# Real-Time Synthetic Primate Vision


Andrew Dankers

Department of Information Sciences and Engineering

Australian National University

A thesis submitted for the degree of

*Doctor of Philosophy*

2007

# Declaration

These doctoral studies were conducted under the supervision of Dr Nick Barnes and Professor Alex Zelinsky. The work submitted in this thesis is a result of original research carried out by myself, except where duly acknowledged, while enrolled as a PhD student in the Department of Information Engineering at the Australian National University. It has not been submitted for any other degree or award.

Andrew Dankers

# Acknowledgements

I wish to express profound gratitude to my supervisors, Dr Nick Barnes and Professor Alex Zelinsky. I thank Professor Zelinsky for the opportunity and inspiration to conduct this research. I thank Dr Barnes for his tenacious guidance and enthusiasm. I wish to thank both Dr Barnes and Professor Zelinsky for the past, present and future opportunities that would not exist were it not for their generosity and support.

I also wish to thank all of my friends and colleagues at the RSISE for making the curricular and extra-curricular environment so rewarding. In particular, I thank Felix Schill for all the frivolous, earnest and always entertaining discussions, that on several occasions elicited side-projects.

Finally, special thanks goes to my family - Colleen, Ivar, Sonia, and Brett - whose unconditional nurture has manifest itself in this thesis.

# Associated Publications

Publications resulting from this research, in print prior to the submission of this thesis:

**A real-world vision system: mechanism, control and vision processing.** Andrew Dankers and Alex Zelinsky, *International Conference on Vision Systems* (ICVS) Graz Germany, 2003.

**CeDAR: mechanism, control and vision processing.** Andrew Dankers and Alex Zelinsky, Machine Vision and Applications (MVA) Journal, 2003.

**Driver assistance: contemporary road safety.** Andrew Dankers, Luke Fletcher, Lars Petersson and Alex Zelinsky, *Australian Conference on Robotics and Automation* (ACRA) Brisbane Australia, 2003.

**Active vision: rectification and depth mapping.** Andrew Dankers, Nick Barnes and Alex Zelinsky, *Australian Conference on Robotics and Automation* (ACRA) Canberra Australia, 2004.

**Bimodal active stereo vision.** Andrew Dankers, Nick Barnes and Alex Zelinsky, *Field and Service Robots* (FSR) Port Douglas Australia, 2005.

**Active vision for road scene awareness.** Andrew Dankers, Nick Barnes and Alex Zelinsky, *International Conference on Intelligent Vehicle Systems* (IVS) Las Vegas USA, 2005.

**MAP ZDF segmentation and tracking using active stereo vision: hand tracking case study.** Andrew Dankers, Nick Barnes and Alex Zelinsky, *Computer Vision and Image Understanding* (CVIU) Journal, 2005.

**Primate structures in synthetic dynamic active visual saliency.** Andrew Dankers, Nick Barnes and Alex Zelinsky, *Epigenetic Robotics* (EPIROB) Paris France, 2006.

**A reactive vision system: active-dynamic saliency.** Andrew Dankers, Nick Barnes and Alex Zelinsky, *International Conference on Vision Systems* (ICVS) Bielefeld Germany, 2007.

# Abstract

We develop a real-time synthetic active vision system based upon observations of the primate vision system. Inspiration was sought from existing knowledge of the biology of the primate visual brain, and from existing models of the primate vision system and its components. A minimal set of biologically plausible processing components of vision was selected, and implemented on a real-time processing network based around a biologically-inspired active vision mechanism. The mechanism's performance capabilities match that of the human vision system, in terms of speed and range of motion.

The processing components include: active rectification for egocentric spatiotemporal perception; a space-variant occupancy grid framework tailored specifically for use with active vision that facilitates estimation of scene structure, motion, and cue-surface correspondence; a foveal MRF ZDF algorithm that permits real-time coordinated stereo fixation upon, and segmentation and tracking of arbitrary, agile, and rapidly deforming visual targets; and, an active-dynamic attention system based on a set of biologically plausible bottom-up saliency cues, that incorporates active-dynamic inhibition of return, an updateable task-dependent spatial bias, moderation of covertly selected saccade destinations before overt attention is deployed, and top-down modulation of all saliency cues, biassing, and moderation parameters.

The system components are combined on a processing network that permits concurrent serial and parallel processing pathways. Processing tasks are distributed over server nodes in the network so as to minimise processing latency and network bandwidth, and so that processing routines are not duplicated. Based on these constraints, the

best functioning solution is found to exhibit a structure broadly similar to that of processing areas of the primate visual brain.

Psycho-physical experiments were conducted with humans to identify inter-individual trends in human gaze behaviours. While viewing a dynamic, repeatable, controlled 3D scene, participants' unconstrained gaze scanpaths were recorded. Parameters useful in characterising human gaze behaviours were selected based upon two pilot trials. Group statistics associated with each selected parameter were then extracted from 20 subsequent human trials.

A comparison was then conducted between the behavioural characteristics elicited by the synthetic vision system and the behavioural benchmarks obtained during the human trials. The synthetic active vision system was subjected to the same trial multiple times, each with different configuration settings. Behavioural parameters were extracted at each iteration. Good coherence was found between all extracted behavioural parameters in the synthetic and biological primate vision systems. Though tunable to some extent by varying system configuration parameters, the synthetic system elicited trial behavioural parameters that fell within the variances of the human benchmarks.

# Contents

# List of Figures

## LIST OF FIGURES

# List of Tables

# Nomenclature

### Acronyms

AI artificial intelligence

C++ enhanced version of the *C* programming language

CAD computer aided design

CamShift continuously adaptive MeanShift

CCD charge coupled device

CCS colour centre surround

CeDAR cable drive active-vision robot

CI confidence interval

CORBA client object request broker

CPD cycles per degree

CPU central processing unit

CRF classical receptive field

DFCS depth flow centre surround

DM dorsomedial area

DOG difference of gaussian

DSP digital signal processor

EW Edinger-Westphal nucleus

FaceLAB a commercial gaze tracking system

fMRI functional magnetic resonance imaging

FPGA field-programmable gate array

FST the floor (or fundus) of the superior temporal sulcus (STS)

HCI human-computer interaction

HMM hidden Markov models

I/O input/output, pertaining to data flow

IOR inhibition of return

IPP intel performance primitives

KL divergence Kullback-Leibler divergence

LED light-emitting diode

LGN lateral geniculate nucleus

MAP maximum *a posterior* probability

MeanShift a nonparametric clustering technique

MMX multimedia extension set

MRF Markov random field

MRF ZDF Markov random field zero disparity filter

MST medial superior temporal area

MT middle temporal

NCC normalised cross-correlation

NDT neighbourhood descriptor transform

NP non-deterministic polynominal time

OCR optical character recognition

OCS orientation centre surround

OpenGL open source hawdware-accelerated graphics library

PC personal computer

PDF probability density function

PID proportional-integral-derivative

PIT posterior inferotemporal

PWM pulse width modulated

QR orthogonal matrix triangularisation

RPC remote procedure calls

SAD sum of absolute differences

SCN suprachiasmatic nucleus

SIFT scale invariant feature transform

SIMD single instruction multiple data

SSE streaming SIMD extensions

STG Servo-To-Go Inc

TPM trapezoidal profile motion

TSB task-dependent spatial bias

TTC time to collision

V4t middle temporal crescent

## NOMENCLATURE

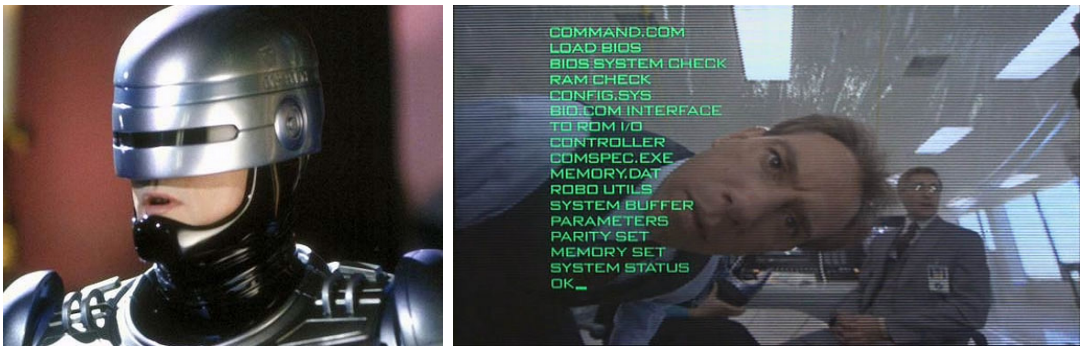
VOR vestibulo-ocular reflex

VP ventral posterior

YUV a colour space: Y represents intensity; U and V are chrominances

ZDF zero disparity filter

# Prelude




Hello world! RoboCop's first bootup - "RoboCop", 1987.

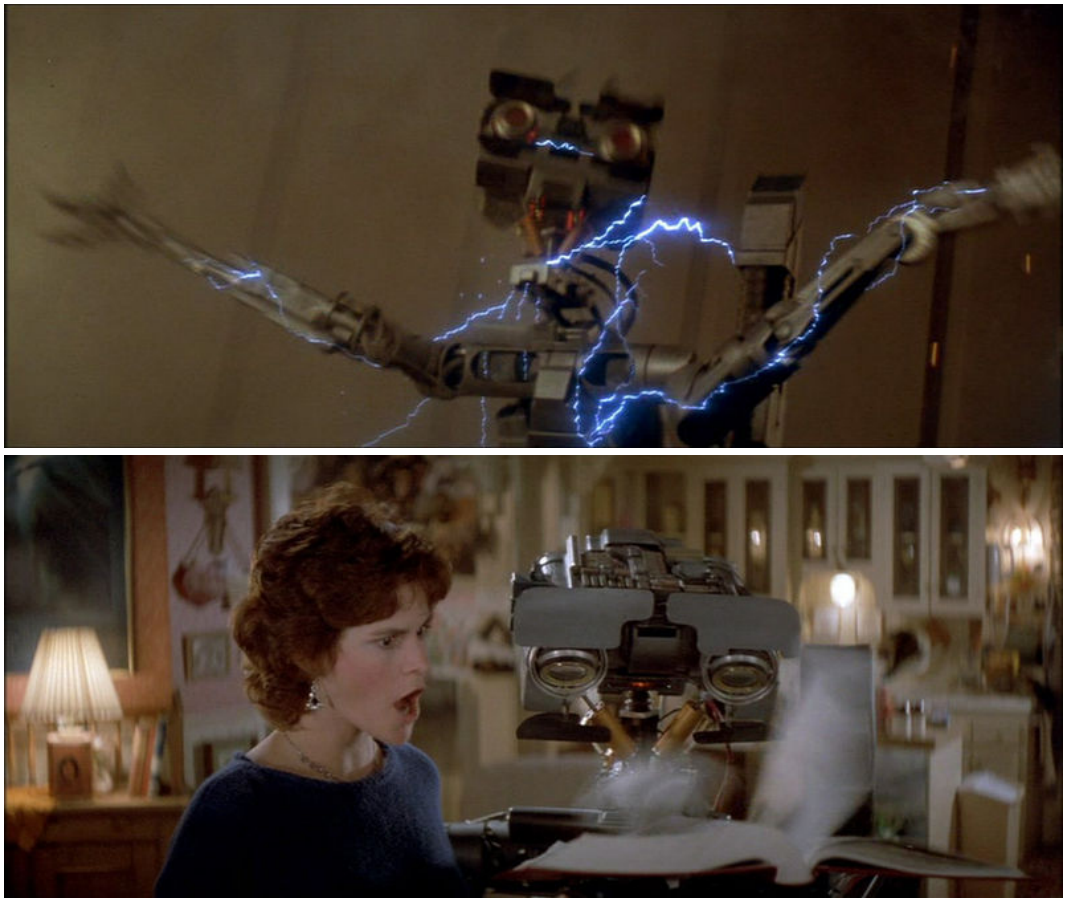
## Hello World!

You have recently announced"Hello World!", and if you are to become an intelligent robotic system, you are probably eager to explore your environment, develop a conscience and experience emotion. But you learnt to walk before you could run, and likewise, your first stepping stone towards an awareness of your surroundings is to develop your basic visual competencies. Whether you are a T-800 looking for a wardrobe, a 'Hive' security system waging biological warfare, a robot cop protecting humanity from its own shortcomings, or a military machine zapped into peace-loving consciousness by a freak bolt of lightning, you have probably found your first glimpse of the world a little daunting. But rest assured, basic scene awareness is indeed within your reach, and if you pay particular attention to those who have gone before you (and possibly whose ponderings and experiments are the very reason for your existence), you will soon be on your

way to deciding where to look next. After all, biology has had millions and millions of years addressing this question, so it is probably best to start with a little biological inspiration before leaping towards the top of the evolutionary chain.


T-800 considers expanding his wardrobe - "Terminator 2", 1991.


The Hive detects a possible intruder - "Resident Evil", 2003.



After being struck conscious, it doesn't take long for Johnny-5 to master basic visual competencies; in no time at all he amazes his friends with speed reading demonstrations - "Short Circuit", 1986.

# Chapter 1

# Introduction

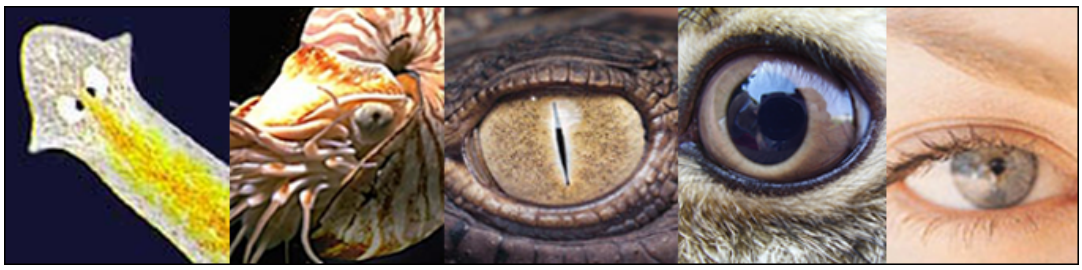


Figure 1.1: Towards human vision in nature: (from left) the proto-eye of a planaria, and the eyes of a nautilus, reptile, mammal and human.

## 1.1 Introduction

Vision is a data-rich sensing modality useful for environmental perception, navigation, search, hazard and novelty detection and communication. In this research, we investigate machine vision for seeing in the real world. In terms of general perception and flexibility, the best performing systems capable of interpreting real-world visual data exist in nature. We aim to reduce the gap between biological vision and synthetic vision, and to investigate primate-like visual perception.

We first outline the domain of this research. We then summarise the evolution of biological vision. We highlight the success of biological vision as motivation for

biological inspiration in the development of synthetic vision systems for sensing the real world. We compare the current capabilities of machine vision with the capabilities of biological vision. Specifically, we justify the investigation of primate vision for improving synthetic systems. We suggest components of primate vision from which machine vision may benefit, and we describe how to combine these elements into a synthetic vision system.

The chapter concludes by summarising the research contributions presented in this thesis. A roadmap of the subsequent contents of the thesis is provided.

## 1.2 Research Domain

In recent years, increased hardware performance versus component cost has brought vision firmly into the realm of practical robot sensors. The domain of computer vision has sufficiently matured to enable researchers to build and experiment with systems that model and interact with what they observe. In particular, we investigate primate-like perception in the real world. We work towards reducing the gap between the visual perception abilities of machines and primates.

We concern ourselves with refining a multi-purpose visual sensor system for real-world, real-time, task-directed perception. The vision system should demonstrate usefulness in performing a diverse range of visual tasks. It must be able to intelligently gather data from its environment in a sufficiently timely fashion to make the decisions for task-oriented behaviour.

A practical system is required to react to the real world in real time. Real environments contain events occurring on all timescales. It is therefore practical to consider real-time as a time period commensurate to the defined task. The real world is an unstructured, possibly cluttered, dynamic environment that extends beyond sensor range. An active vision system operating in the real world must therefore be equipped with mechanisms to fixate its attention upon what is important within the time-frame of its relevance, while simultaneously disregarding background irrelevancies. Appropriate sensory and processing resources must be selected with these considerations in mind.

This research has applications in the development of autonomous agents, investigations into synthetic primate vision, human-computer interactions, auto-

mated novelty detection, human visual assistance such as the development of visual prosthetics, and facilitates exploration of what visual competencies can be implemented in real-time within reasonable processing limitations.

### 1.2.1 Let There Be Light

Where there is light, vision is an especially useful sensing modality. In nature, most animal species need to be able to detect, for example, food and predators. In the presence of light, the eye allows the detection and observation of such objects from a safe distance. The majority of pertinent matter that an overland or marine based organism needs to be spatially aware of (particularly in terms of survival) comprises of the surfaces of solids and liquids - a set of matter that often reflects, refracts, absorbs or emits light. Of course, the naturally occurring mixture of gasses in breathable air is also crucial for the survival of most species of land animals, but it can reasonably be assumed that air does not usually need to be sought or detected, that it is present while an organism is able to respire, and that it does not need to be spatially located. Moreover, it is especially convenient that air rarely interacts with light in a manner that is detectable by the eye, other than where it borders solid or liquid matter (it is typically transparent), allowing seeing animals to detect the far more pertinent set of matter beyond the air immediately in front of their eyes. For marine-based species, water fills the afore-mentioned role of air, and it is again convenient that water is often largely transparent to a submerged marine animal's eyes. Eyes could only evolve because the environmental media in which organisms live submerged is transparent, and illumination is present. In any case, overland and water based animals live in an environmental media where pertinent matter may not always produce audible sounds or detectable smells, but where all solid and liquid surfaces transmit, reflect, refract or absorb light in a manner detectable by the eye.

Given illumination, the visual range is not limited to localised regions, rather, it depends only upon the size of an object and the system's visual resolution. This range property arises because vision does not probe its environment (via touch or by emitting decaying rays – in contrast to radar or laser range-scanning). Unlike touch or smell, vision provides a large search space - the eye is capable of seeing

Figure 1.2: Towards human vision: in science fiction (top); in research (bottom).

things from only millimeters away to the distance of stars. Whilst illumination is present, information from all sources and reflectors of light in sensor view is received simultaneously. The set of visually observable space is only reduced by occlusions. Although light does not transmit detectably through most matter or around corners, the lack of ambiguous reflections, refractions and interference helps to disambiguate spatial localisation, unlike hearing. However, vision is not without sources of error; the eye must cope, for example, with dynamic scene illumination, shadows, aliasing, defocus and saturation.

Vision synergistically compliments our sense of touch when we interact with our environment or manipulate objects. It is certainly very useful to be able to visually locate something before we reach to interact with it, or before it attacks us!

## 1.3 What Machines Have Seen

Fiction would have us believe that the time when robots will look, feel and function like humans is fast approaching. Indeed, research is progressing, and although similarities are apparent (Figure 1.2), science fiction is still a few steps

ahead of science fact (as it ought to be - we cannot build what we cannot imagine). Nevertheless, cameras are becoming increasingly resolute, approaching the acuity of human vision. Modern imaging technology enables us to see that which we once could not.

### 1.3.1 Imaging Capabilities

The functioning of a camera is often compared to that of the single aperture eye because both focus a projection of the observed scene onto a light-sensitive medium. In the case of most common cameras this medium is film or an electronic sensor [Holst (1998)]; in the case of the eye it is an array of photoreceptor cells.

Numerous types of imaging technologies exist that can see what we otherwise could not. Hyper-spectral imaging, for example, allows us to see in infra-red and ultraviolet light, including thermal images. With non-pinhole imaging, such as x-ray, ultra-sound, radiological and magnetic resonance, we can see through matter and inside objects such the human body. Electron microscopes can capture images of minute structures. Telescopic zoom lenses enable us to see long distances. Satellite imagery allows us to see our own planet from orbit, and even peer into space and back in time.

Cameras can act as remote 'eyes'. We can send camera images around the world to remote destinations, enabling common communication technologies such as television and video-conferencing. We can record camera imagery and shift it in time by playing it back when it suits us, or capture a moment with a photograph to help us remember an event. Cameras can provide us with security by watching over our belongings when we would rather be elsewhere. If we are to save time by not manually interpreting camera images, we must give machines the ability to automatically interpret images for us.

Computers may be programmed to automatically acquire and process camera images. *Computer vision* and *machine vision* are the terms commonly associated with the theory and technology of building artificial systems that interpret digital images. Computer vision systems may be used to perform automatic visual or physical tasks.

### 1.3.2 What Success To Date?

The development of artificial vision systems was initially slow, its beginnings emerging in the 1960s. It was not until the late 1970s when computers could process large data sets, such as images, that the field of computer vision took flight.

It was initially thought that, if given reasoning programs and sensors such as cameras, actuators and manipulators, highly efficient autonomous systems would emerge. In fact, it has proved much more of a challenge. Real autonomous agents have to deal with the real world, not just the symbolic representations as assumed by traditional artificial intelligence that human perception seems to extract with ease. In the 1980s scientists instead focussed on reconstructing representations of observed environments, and the shape, location and orientation of objects in this environment, in order to give robots awareness of their surroundings. This approach required a large amount of computational power, having the effect that robotic systems could only operate in constrained environments.

Towards the late 1980s it was realised that it was impossible to deal with all the information in a visual signal in real-time. Rather, it was suggested that the robustness of vision systems could be improved through camera movements and focusing procedures – an active vision approach. A passive system is unable to change the way it views a scene, and must extract all the required information from images captured with the same parameters. An active system is able to acquire more relevant data by adjusting its parameters to recover the visual information required for a task. Ballard published results to prove that moving sensors significantly decreased such computational costs [Ballard (1991)].

Active vision signals a distinctive shift from data-driven to task-driven applications. A task-oriented system that uses an active sensor can select useful information and ignore task-irrelevant parts of the scene. A purposive vision regime changes the requirements of what needs to be perceived.

As well as mimicking ocular motion, researchers have also reproduced primate-like active vision by, for example, using specialised *log-polar* cameras [Rougeaux (1999)]. For computers and the brain alike, foveal vision reduces resolution away

from the fixation point, and accordingly, reduces computational bandwidth. Unlike the human vision system that has a periphery that extends to about 140°, the range of vision provided by a camera with a conventional lens is usually limited to a much smaller angle (and a rectangular frame).

### 1.3.3 Application Domains

Algorithms that perform computer vision tasks often fall into one of the following categories:

- Detection: image data is scanned for a specific condition.
- Recognition: one or several pre-specified or learned objects or object classes can be recognised, usually together with their 2D positions in the image or 3D poses in the scene.
- Identification: an individual instance of an object is recognised, such as a specific person's face.
- Segmentation: pixel-wise extraction or categorisation of regions in 2D images.
- Tracking: following the movements of specific objects.
- Motion estimation: estimating camera egomotion or scene motion.
- Spatial reconstruction: 3D reconstruction of scene structure and surfaces.

These task categories have been applied to numerous application domains, including:

- Medicine: extraction of information from image data for diagnosis. Image data can be in the form of microscopy images, X-ray images, angiography images, ultrasonic images and tomography images. For example, detection of tumors and arteriosclerosis, 3D organ mapping, blood flow, functional structure of brain.

- Geophysics: satellite imagery, hyper-spectral imaging, agriculture, climate/weather monitoring and prediction, mapping, geographical censuses.
- Industry: manufacturing process, quality control (automatic goods inspection), object manipulation.
- Military and surveillance: intrusion detection, missile/vehicle guidance, communication, intelligence,"battlefield awareness", traffic monitoring, crowd monitoring, intrusion detection.
- Autonomous vehicles: either full autonomy or driver support, mapping, navigation, feature and obstacle/incedent detection.
- Embedded devices: face detection on digital still cameras, video-conferencing automatic target selection, vehicular pedestrian/sign detection, optical character recognition (OCR), augmented reality.

### 1.3.4 What's Missing?

Computer vision is presently very useful for performing algorithmic visual tasks that humans would otherwise find dangerous, monotonous or computationally demanding. Today, more and more computer vision applications find their way into commercial products. In terms of commercial devices (as opposed to pure AI research), the trend seems to be towards modular, static, monocular systems. Commercial products in particular, by virtue of the fact that consumers expect a product to perform the task it was purchased for, are often developed for reliability in performing a specific task within defined constraints. This is a valid and very useful approach, especially for modular applications.

Task-specificity of vision systems may originate where computer vision solutions are sought in specific scientific fields, for well-defined problems under controlled conditions. As the complexity and applicability of computer vision systems expands, it is important to reduce the brittleness of integrated system components.

Research into artificial intelligence and humanoid vision is more interested in primate-like scene awareness. The real world is dynamic in many respects,

requiring flexible real-time visual sensing. Real environments contain events occurring at all timescales. Processing resources are always limited to some extent. Primates are able to integrate multiple interpretations of vision into a timely, unified perception. There are certainly gains to be made from observations of primate vision in this regard. We therefore investigate the integration of multiple visual tasks on a primate-like vision platform.

For convenience, computer vision algorithms sometimes assume human-imposed abstract starting conditions. For example:

- Fixed camera systems may negate the need for selection of visual field.
- Images out of their spatiotemporal context may eliminate need for tracking.
- Pre-segmentation may eliminate the need to select a region of interest.
- Clean backgrounds may ameliorate segmentation problems.
- Clean images may ameliorate the requirement to cope with noise.
- Assumptions about relevant features and the ranges of their values reduce their search ranges.
- Assumed knowledge of the task domain may negate the need to search a stored set of all domains.
- Assumed knowledge about which objects appear in scenes may negate the need to search a stored set of all objects.
- Assumed knowledge of which events are of interest may negate the need to search a stored set of all events.

Biological visual perception cannot rely upon such abstract starting assumptions. Such assumptions may help with data reduction and reduce processing times, but they may also adversely affect perception. For example, addressing data and search reduction by imposing a static reference frame may also impose restrictions on perception. Biological reduction techniques, such as attention and foveal vision, are unlikely to affect the successful completion of visual tasks,

and are also less likely to restrict perception. As another example, biology is known to use robust, flexible, low-level consistent scene representations, as a basis upon which higher-level interpretations, tasks and cognitions are computed. The primate visual brain, for example, interprets illumination-independent image representations (similar to contrast images) passed from the retina along the optic nerve, from which a perception is built [Rodieck (1998)]. Rather than addressing computer vision tasks by operating on camera images directly, conversions to robust low-level image representations - that are, for example, illumination independent, camera motion independent, or noise reduced - may be meritous. It is likely that useful low-level, low-bandwidth representations of visual data could be compiled once and then re-used in concurrent (or at least one) high-level computer vision tasks. Like humans, autonomous machines may need to perform multiple visual tasks simultaneously (such as spatial perception, mapping, localisation, search and obstacle avoidance), as well as to be perceptive to novel visual events.

Modular computer vision algorithms may provide insight into how the brain may perform specific visual tasks. However, investigating how biology integrates visual abilities from the lowest sensory level into general scene perception may be more useful in proposing a generalresearch formulation of how more complex computer vision problems can be solved. A key factor in the gap between animal and machine vision could be machine vision's general lack of low-level visual *perception*. For intelligent systems and autonomous agents - regardless of their final application - it may be beneficial to develop effective low-level visual scene perception capabilities instead of directly constructing higher level interpretations and models of the world. Machines exist in and perceive the same world as animals. Exploring flexible synthetic scene perception may benefit considerably from biological inspiration.

## 1.4 A Sense of Vision

We now look at the development of visual perception in biology. We consider the evolution of sight from the conception of vision to primate vision.

### 1.4.1 First Sight

It is widely accepted that all varieties of animal eyes evolved from a proto-eye that first appeared around 540 million years ago [Halder *et al.* (1995); Parker (2003)]. Anatomical and genetic features common to all eyes provide evidence in support of such a common origin. The earliest forms of optical sensing in biology were as simple as detecting the presence or absence of light via photoreceptor cells [Land & Nilsson (2002)]. Patches of such cells evolved that could sense the level of ambient lighting. By depressing the patch to form a pit, it became possible for organisms to sense the direction towards light sources. Modern planaria (Figure 1.1) and other early invertebrates (such as some slugs and snails) that first appeared in the Cambrian period can differentiate the direction and intensity of light sources because of their cup-shaped, heavily-pigmented retina cells, which shield the light-sensitive cells from exposure in all directions except for the single opening for the light. This proto-eye is more useful for detecting the level of ambient light than the direction of its source. As the proto-eye pit deepened and the number of photoreceptive cells grew, visual information could be deciphered with increasing precision [Land & Nilsson (2002)]. Overgrowths of transparent cells prevented contamination and parasitic infestation - the first stage in the evolution of the lens [Land & Nilsson (2002)]. From such common ancestry, multiple types and subtypes of visual sensors developed in parallel.

Trilobites, for example, were amongst the first animals to develop more advanced visual capabilities [Sherwin & Armitage (2003)]. The majority of trilobites possessed a pair of compound eyes. Compound eyes are today found in arthropods such as insects and crustaceans. A compound eye is comprised of several, even thousands, of tiny independent photosensitive units (ommatidia) that are oriented to point in slightly different directions [Sherwin & Armitage (2003)]. Each ommatidia consists of a cornea, lens, and photoreceptor cells. The image perceived by a compound eye combines the input from the numerous ommatidia.

Single aperture eyes evolved from the proto-eye as the pit deepened into a cup, then a chamber [Land & Nilsson (2002)]. By reducing the size of the chamber opening, a 2D projection of the scene could be formed on retinal photoreceptor cells at the rear of the chamber. Similar to the modern pinhole camera model

(see Chapter 5), this type of eye allowed for finer directional sensing and even some edge and shape detection. The nautilus (Figure 1.1) is an example of an existing species that still possesses an early form of single aperture eye. Lacking a cornea or lens, the nautilus eye provides poor resolution and dim imaging but is a significantly more acute sensor than the proto-eye [Land & Nilsson (2002)].

Compared with single aperture eyes, compound eyes usually have lower image resolution. However, they generally possess a larger viewing angle, the ability to detect fast movement, a broader spectral response and, in some cases, the polarisation of light [Land & Nilsson (2002)]. For example, bees can see ultraviolet light [Bellingham *et al.* (1997)]; the mantis shrimp possesses polarisation detection capability, hyperspectral capability[1] [Cronin & King (1989)], and triple redundant depth perception from both their eye constructions and their multiple eyestalk motions (both 2D tracking, and axial rotation). The fact that these capabilities are achieved using a compound eye shows the merit of radically divergent evolution accelerated by an evolutionary visual "arms race" [Parker (2003)].

### 1.4.2 Towards Modern Vision

The majority of the advancements in early eyes are believed to have occurred over only a few million years [Land & Nilsson (2002); Parker (2003)]. With the emergence of the eye, a visual arms race began where prey and predator species alike were forced to rapidly match or exceed any advancing capability of their counterparts. As visual species diversified to find their niche environment, the evolution of each species and how that species visualised their environment became intricately coupled. For example, birds of prey evolved high visual acuity [Land & Nilsson (2002)] - much greater than that of humans - complimenting their evolving ability to hunt and detect prey from altitude, or through undergrowth or camouflage. Animals such as rabbits and chameleons have eyes located so as to reduce sensory overlap [Land & Nilsson (2002)], providing a wider field of view suited to detecting the threat of an advancing predator. Many species, including some mammals, birds, reptiles and fish, have eyes whose fields of vision

[1] The mantis shrimp's hyperspectral capability incorporates three to four times the number of receptors by range as humans (without including interpolation) over a wider spectral range

largely overlap, to allow better depth perception (stereo vision - differences in two simultaneous views of an object provide information about its distance). By virtue of the fact that all animal species are able to move relative to their surroundings, any animal equipped with a vision sensor is able to use that sensor to actively investigate their environment from multiple perspectives. Motion of the eye itself was the next progression in active visual sensing.

### 1.4.3 The Expanding Visual Brain

Links exist between the evolutionary emergence of vision and the expansion of the animal brain [Parker (2003)]. Part of the increase in neural population in vision-equipped species was undoubtedly to cope with the new and increasingly vast channel of information. In fact, the sensory bandwidth of the visual channel is so rich that processes exist to reduce the amount of data to that which is relevant. For example, the contrast-sensitive response of ganglion cells occupies far less bandwidth on the optic nerve ($1/100th$) than the amount of data hitting the retina [Rodieck (1998)]. As we shall see (Chapter 8), saliency and attention in the early brain instantiate an informational bottleneck that serves to detect the most relevant visual regions to prioritise data search so that non-relevant regions are represented with minimal bandwidth. Vision can provide spatial information such as the location of free and occupied space; the ability to localise and navigate within complex environments; novelty and hazard detection; communication; imitation and learning. Colour vision helps to further disambiguate, segment and recognise borders, objects and regions. Colour can also contribute to communication efficiency. The neural population of the visual brain has undoubtedly expanded to allow animals to develop such abilities.

## 1.5 Nurture From Nature

Machines and animals sense the same real world. We have seen that animals have evolved invaluable visual abilities. With observations of biology, it may be possible to accelerate the evolution of machine vision.

### 1.5.1 Why Primates?

In terms of specific properties, the human eye is seemingly outdone by the eyes of what we might otherwise consider to be "lesser species". In many cases, primate eyes have a narrower spectral response than other species. In daylight, human visual acuity is significantly less than that of raptors in terms of spatial resolution, and significantly less than various insects in terms of spectral response range. At night, human vision is again less acute than that of raptors, as well as cats, and even invertebrate molluscs such as squid and octopuses. So why focus on primate vision?

- We *are* primates; self introspection gives us insight into how we interpret light.
- Primates are able to provide high-level experimental feedback. Primates can be more or less cooperative test subjects in psycho-physical experiments.
- Primates perform visual tasks seemingly effortlessly.
- We are interested in developing synthetic systems that can see how primates see, so that we can learn about ourselves.
- Primates are highly intelligent species. They exhibit the visual abilities we wish to synthesise.
- We are interested in investigating machines capable of producing human-like visual abilities.
- We are interested in developing systems capable of assisting humans, for example, via the development of prosthetic vision.

Primates use vision to perceive their environments efficiently and accurately. We now consider how machines can benefit from biology, identifying which primate visual abilities we consider important for synthetic vision.

## 1.6 Components of Perception

We now list and define components of perception considered important to primate vision. We also discuss the relevance of these components to synthetic primate vision. Evidence for the existence of components in the primate brain is presented in the next chapter.

### 1.6.1 Active Vision

In terms of both biology and machine vision, active vision has over the years been defined in various similar ways. In a recent 2007 keynote speech (ICVS'07), Tsotsos summarised these as follows:

- Ullman (1984): "A set of visual routines".
- Bajcsy (1985): "Active sensing is the problem of intelligent control strategies applied to the data acquisition process which will depend on the current state of data interpretation including recognition".
- Aloimonos *et al.* (1988): "Geometric control of sensor".
- Burt (1988): "Dynamic vision - foveation, tracking and high-level interpretation".
- Ballard (1991): "Gaze control in animate vision systems".
- Blake & Yuille (1992): "Active vision emphasises the role of vision as a sense for robots and real-time perception systems, with advantages for structure from controlled motion, tracking, focussed attention and prediction".
- Pahlavan *et al.* (1993): "An active visual system is a system which is able to manipulate its visual parameters in a controlled manner in order to extract useful data about the scene in time and space".
- Aloimonos *et al.* (1993) expands his earlier definition, introducing: "Purposive and qualitative active vision" to the concept of *active perception*.

Further, Tsotsos discusses that machines might use active vision to:

- Attend a selected fixation location.
- Complete a task requiring multiple fixations, tracking or visual feedback such as during object manipulation.
- Compensate for spatial non-uniformity of a processing mechanism such as foveation.
- Track moving objects such that they become pseudo-stationary, reducing motion blur.
- See a portion of the visual field otherwise hidden due to occlusion.
- Expand the visual search space.
- Improve acuity via sensor zoom or observer motion.
- Adjust stereo vergence for spatial perception.
- Disambiguate or eliminate degenerate views.
- Determine induced motion (kinetic depth).
- Address lighting changes (photometric stereo).
- Provide a viewpoint change when viewing a subject.

Of course, active abilities do not come without computational overheads. An active system will also need to:

- Decide that some action is needed.
- Determine which changes are required and decide upon a priority sequence.
- Determine the spatial correspondence between the old and new viewpoints.
- Execute the change.
- Adapt the system to the new viewpoint.

It is well understood that an active vision approach can offer computational benefits for scene analysis in realistic environments [Bajczy (1988)]. In order to incorporate active vision, the benefit must outweigh the cost of such overheads. Tsotsos has described effeciency benefits of active over passive perception [Tsotsos (1992)]. As mentioned, Ballard published results to prove that instead of increasing computation involved with computer vision, moving sensors actually decreased the computational costs [Ballard (1991)]. Where animals use active vision the benefits undoubtedly outweigh the costs. Biology has overwhelmingly adopted active vision, suggesting it is a desirable proficiency in many circumstances, and perhaps a verification of Ballard's findings.

Active vision permits *foveal* vision. Many species exhibit highest visual acuity in a small central region of the retina known as the fovea. The human visual system, for example, exhibits its highest resolution in the fovea, a region approximately the size of a fist at arms length [Wandell (1995)]. It has been estimated that if the human eye exhibited homogeneous resolution distribution, it would weigh approximately 13,500 kg [Aloimonos *et al.* (1988)]. In this manner, foveal vision permits significant data and processing reductions. The rest of the retina constitutes the visual periphery. Despite being less resolute, the periphery is very sensitive to motion [Schwartz (1980)]. By orienting the eye so as to place the fovea over regions of interest in the focussed retinal projection of a scene (a process known as *fixation*), a resolute perception of the scene can be constructed. Humans centre and focus both eyes on surfaces of interest - originally detected in the periphery - within the fovea for subsequent resolute processing. In fact, we find it difficult to defocus our eyes or prevent fixation while our eyes are open. Foveal vision is strongly complementary to both active vision and attention.

It is difficult to identify a vision-equipped animal that does not move to perceive its surroundings from multiple perspectives (*actively*), or that does not have the ability to direct its attention towards different locations in a scene. To investigate primate-inspired synthetic vision we would like to incorporate primate-inspired attention, of which active vision is a required component.

### 1.6.2 Attention

Attention is the cognitive process of selectively concentrating on one aspect of the environment while ignoring less important stimuli. Visual attention is likely to have evolved due to inherent capacity limits in visual processing resources in combination with the evolutionary competitive need for increasingly resolute and intricate perception.

Research has shown that purely feed-forward, unconstrained visual processing seems to have an inherent exponential nature. However, even small amounts of task guidance can turn an *NP-complete* problem into one with linear time complexity [Tsotsos (1989)]. He defines visual attention as: "the set of mechanisms that seek to optimise the search processes inherent in vision" [Tsotsos *et al.* (1995)], and that the deployment of attention broadly incorporates:

- Select viewing parameters.
- Selection of spatial and feature dimensions of interest within the visual field.
- Selection of the general visual field - *active vision.*
- Selection of the visual field for detailed analysis - *active vision.*
- Selection of objects, events and tasks.
- Selection of a world model.

It is noted above that active vision is considered a component of attention.

Biological vision is heavily reliant upon visual attention. Animals are good at selecting an appropriate visual field for performing tasks. They are also excellent visual novelty detectors. Novelty detection and efficient task execution are important to survival. Attention helps reduce the amount of visual data presented to the brain for consideration. It also assists visual search by prioritising the evaluation of salient regions according to a visual task.

Research exists in the area of synthetic machine attention, but relatively few implementations incorporate real-time active vision. The lack of active attention implementations means less practical insight exists in current research. That

is, implementing attentional saliency on an active head platform can produce artifacts that may not have otherwise been predicted or accounted for in existing models. Numerous models of attention have been proposed but few incorporate the use of active cameras and dynamic 3D scenes in real-time. As we shall see, active attention reveals issues that must be addressed.

#### 1.6.2.1 Covert Vs Overt

Overt attention involves explicitly directing sensors towards a source of stimulus such that information about that stimulus is maximised. Covert attention involves the consideration of one stimuli in a non-overt manner. Covert attention is thought to be a neural process that enhances the signal from a particular region of the sensory panorama without overtly directing sensors towards that region. The concept of covert attention was documented as early as 1890 when H. von Helmholtz noted that he found himself "able to choose in advance which part of the dark field off to the side of the constantly fixated pinhole [he] wanted to perceive, by indirect vision" [Nakayama & Mackeben (1989)].

Humans and primates can overtly gaze in one direction but may covertly attend in another. For example, if individuals attend to the right hand corner field of view, movement of the eyes in that direction may have to be actively suppressed while visual tasks are executed. It is likely that covert attention is a mechanism for rapidly scanning the field of view for interesting locations and is involved in the assessment of the next fixation point. In neural recording studies with monkeys, scientists found they could predict the occurrence of saccades by monitoring the activity of certain neurons [Sugrue *et al.* (2005)].

### 1.6.3 Dealing with Dynamics

The real world is dynamic. Incorporating active vision sensing introduces further visual dynamics. In order to cope with dynamic lighting conditions, ganglion cells in the retina convert the retinal projection of the scene into contrast response output. The output on the optic nerve is similar to the output a *difference-of-Gaussian* (DOG) convolution over the original image. Such a representation is considerably more illumination-independent than the original image. Primates

also cope with deliberate eye motions via an egocentric reference frame where spatial locations of scene contents are related across eye motions. Monkeys retain a short term memory of attended locations across saccades by transferring activity among spatially-tuned neurons within the intra-parietal sulcus [Merriam *et al.* (2003)], thus retaining accurate global representations of visual space across eye movements.

Machine vision could benefit from illumination-independent image representations. Also, projecting images into a static egocentric reference frame would relate active camera images over time and space, and from multiple cameras, into a common representation.

### 1.6.4 Coordinated Fixation

Coordinated fixation involves enforcing stereo camera fixation upon a scene point. Ideally, it is the specific propensity to enforce fixation of multiple cameras upon precisely the same scene point, rather than try to point each camera at each location independently via some form of search. Humans find it impossible to fixate both eyes on different scene points. If machines are to experience primate-like vision, they should exhibit the same behaviour.

Implementations of real-time machine attention are not common, so there has been no significant cause to investigate real-time coordinated stereo foveal fixation suited to such real-time attention. Coordinated fixation would compliment attention, foveation and active vision, and help integrate multiple views of a scene into the aforementioned common representation.

### 1.6.5 Spatial Perception

Vision provides animals with a 3D perception of free space and occupied space in their vicinity. Animals are good at collision avoidance, even at high speeds. Humans experience an egocentric spatial reference frame. Egocentric real-time local spatial awareness would be especially useful for autonomous machines and object manipulation.

### 1.6.6 Efficient Representations

The bandwidth of visual data projected onto the primate retina is so rich that processes exist to reduce the amount of data that reaches the visual brain. Accordingly, a synthetic visual system would benefit, in terms of data transfer and bandwidth limitations, from efficient image representations. By efficient, we infer that the bandwidth-reduced representation of the sensory image does not significantly impede perception. Subsequent (higher level) scene interpretations, such as a spatial representation of the scene, should also be constructed in such a way that they do not adversely affect performance.

### 1.6.7 Task Flexibility

Animals efficiently interpret retinal projections of scenes. They are able to rapidly perform multiple visual tasks in parallel. Often their survival depends on this. Primates perform multiple tasks simultaneously. Biology may therefore offer some insight into how to structure such a processing framework. Similarly, a synthetic vision system should be capable of rapidly performing multiple simultaneous visual tasks. The framework should permit concurrent serial and parallel processing. Indeed, numerous scalable vision processing networks exist in current research [Ude *et al.* (2005)]. Such frameworks are useful for investigating primate-like capabilities for synthetic vision systems. The vision system is developed on a processing network of high-end, yet common available, computers.

## 1.7 Building a Model

Over the last century, there have been extensive studies of eyes, neurons and the brain structures devoted to processing visual stimuli in both humans and various animals. This has led to a coarse, yet complicated, description of how biological vision systems operate in order to solve certain vision-related tasks. These results have led to a subfield within computer vision where artificial systems are designed to mimic components, processing and behaviour of biological systems, at different levels of complexity.

We devise a model capable of primate-like perception that incorporates abilities defined in the previous section, based upon literature and experimentation. We consider existing neural and behavioral studies, and psycho-physical evidence. We conduct psycho-physical trials to evaluate aspects of unconstrained 3D human attention. We implement such desirable behaviours on a processing network by minimising network bandwidth and latency. We consider the processing structure in light of neurobiological studies. The implementation incorporates both biologically inspired, and computational algorithms from literature where possible, and additional functionality is engineered as necessary. We allow the cameras to automatically adjust parameters such as contrast, brightness and saturation. The algorithms used should subsequently be robust enough to cope with image variations this may introduce.

The biological hardware of the brain has impressive capabilities. A computer is capable of millions of exacting floating-point calculations per second. Of course, different processing hardware exhibits different strengths and capabilities that may not be transposable. We are, however, interested in what can be achieved on computers in terms of primate-inspired vision. In using computer vision, it is only possible to develop algorithms whose function is similar to human capabilities, but it is not possible, or necessary, to use the exact implementation employed by the brain. For example, the centre-surround contrast response of retinal ganglion cells to the optic nerve can be synthesised by using a DOG convolution applied to a memory buffer containing pixel intensities. The difference in hardware means that in some instances it is necessary to use non-biological methods to generate outputs that synthesise biological functionality. For example, we may produce a depth map using area-based correlations, despite the fact that it is known the brain does not determine scene depths the same way.

Finally, once we have a system capable of synthesising primate scene perception, we want to evaluate it. We want to assess its capabilities in terms of both tangible metrics and primate-like behaviours. Little psycho-physical data exists that evaluates *unconstrained* primate attention. When observing unbounded 3D scenes humans are unlikely to exhibit behaviours identical to those exhibited when they observe static pictures or 2D videos (as utilised in most attentional

experimentation to date). Depth and covert (peripheral) object tracking, for example, are likely to affect attention in a manner that the use of static images cannot demonstrate.

## 1.8 In This Thesis

### 1.8.1 Contributions

We look to biology for inspiration in addressing the challenges of developing a synthetic primate vision system. We develop a model of primate vision that does not contradict behavioural observations of biology. We synthesise components of the primate vision system. With consideration of biology, we develop a flexible framework upon which components can be integrated. We integrate synthesised components into a 'minimal' but expandable system suitable for many tasks. We define methods to quantitatively evaluate unconstraind 3D primate attention. We conduct pychophysical trials to compile behavioural human attention data. Based on this data, we evaluate the implemented system according to behavioral similarity to primate attention.

We investigate what can be implemented in real-time within reasonable processing limitations. We address hardware limitations by implementing components with outputs similar to biological functionality, but whose actual implementation may not resemble that of the primate brain. We produce a system capable of simultaneous spatial awareness, novelty detection and performing visual tasks. In terms of deploying attention, the system behaves similarly to humans. The system allows the investigation of primate-inspired synthetic vision. Its manifestation contributes practical insight into the development of machines that are capable of producing human-like visual behaviours.

Specific contributions (subsequently defined in this thesis) are:

- Active online epipolar rectification.
- Active mosaicing, including spatio-temporal binding of active vision images into a globally epipolar rectified static reference frame.

- The ability to perform any static stereo algorithms on an active stereo platform.
- A Bayesian occupancy grid framework for spatial awareness, incorporating an egocentric reference frame.
- A *Markov random field zero disparity filter* (MRF ZDF) foveal segmentation, coordinated fixation and tracking.
- A real-time implementation of MRF optimisation using graph cuts.
- Primate-inspired synthetic active attention incorporating real-time bottom-up visual saliency and *task-dependent spatial biassing* (TSB).
- Primate-inspired synthetic active-dynamic attention incorporating dynamic *inhibition of return* (IOR).
- A psycho-physical evaluation of unconstrained 3D human attentional behaviours.
- Extraction of parameters from gaze data for evaluation of unconstrained 3D synthetic primate attentional behaviours.
- An expandable real-time vision processing network framework.
- A flexible real-time primate-inspired vision system.

### 1.8.2 Roadmap

Background information upon which this research builds is presented as follows:

- Chapter 2: We describe basic components of the primate vision system.
- Chapter 3: We introduce relevant existing synthetic models of human vision and function of relevant components of human vision. We then propose our system model.
- Chapter 4: We present the biologically-inspired research platform, including its control and *input and output* (I/O) for system integration.

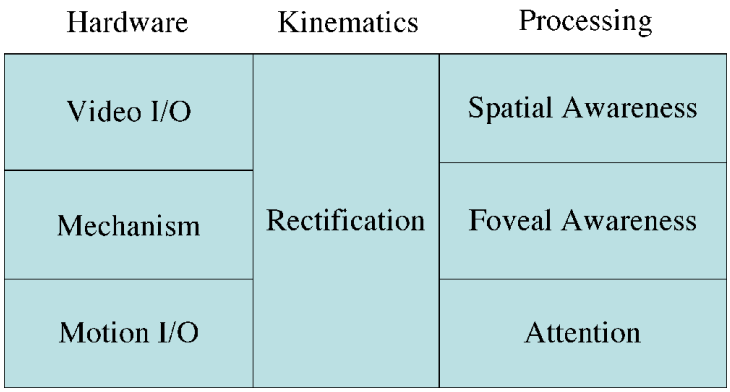



Figure 1.3: Components of the synthetic vision system presented in this thesis.

After presenting background information, the system is described in terms of its components. It incorporates: 1) hardware that produces data; which must 2) be put into spatiotemporal context, for 3) processing and interpretation. Figure 1.3 shows a broad system map divided into three sections: hardware, kinematics and processing. The diagram does not constrain the system in any way. Its boundaries are not necessarily rigid, in particular, the divisions separating rows in the processing column.

The next four chapters contain the major technical contributions in this thesis:

- Chapter 5: We develop methods necessary to cope with active vision including active rectification and mosaicing.
- Chapter 6: We develop a real-time perception of spatial awareness.
- Chapter 7: We develop a robust real-time foveal algorithm that ensures coordinated stereo fixation upon scene surfaces. The algorithm also enables subject segmentation and tracking.
- Chapter 8: We consider where to look. We develop primate-inspired attention suitable for concurrent visual tasks in dynamic scenes with active vision.

The research system is based on properties of primate vision, so it should reflect primate-like behaviours. The system is also designed to operate in the

same environment as primates - the real world. We therefore conduct psychophysical human trials to extract some behavioural characteristics of unconstrained human attention in a 3D scene. Having conducted the trials, we return to the synthetic system to conduct identical synthetic trials for comparison with human results. The corresponding chapters are:

- Chapter 9: Human trials; investigating behavioural characteristics of human visual attention while freely observing a 3D scene, given a simple search task.
- Chapter 10: Synthetic trials; investigating behavioural characteristics of the developed model of synthetic visual attention. A quantitative comparison of the behaviour of the synthetic vision system and human vision.

Finally, the body of research is summarised, its implications discussed, and the thesis is brought to conclusion in Chapter 11.

### 1.8.3 Summary

We investigate machine vision for seeing the real world. We explore what biology finds relevant for visual perception in the real world. We conduct this research using readily available equipment. We apply knowledge from literature, biology and human experience.

# Chapter 2

# Primate Vision System

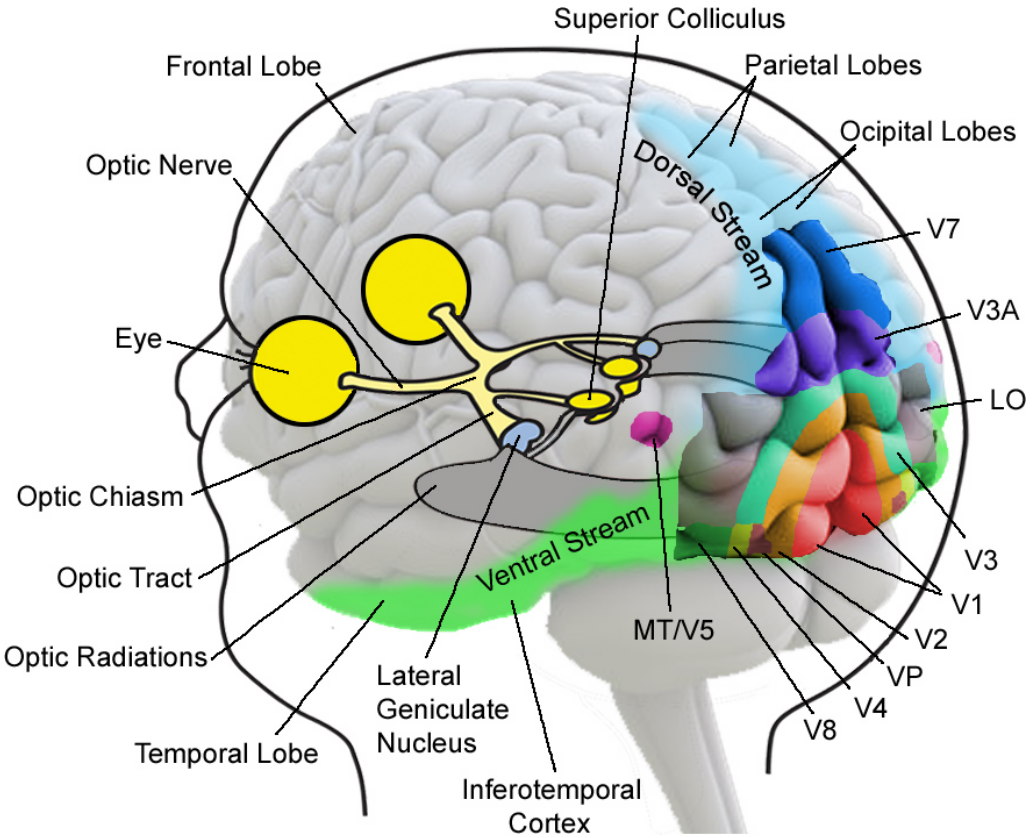



Figure 2.1: Components of the human vision system. The approximate surface profiles of areas in the primate visual cortex are shown.

In this chapter we begin by reviewing components of primate vision. We then discuss the locations in the brain where perceptions such as spatial awareness and attention take place.

## 2.1 Introduction

We have motivated primate vision as inspiration for synthetic visual perception. The visual system is the part of the nervous system that senses and interprets the information available from visible light. Visual information is used to build environmental perception. The visual system has the complex task of interpreting a 3D world from 2D projections of that world.

We now investigate how the primate brain achieves scene perception and what components contribute to perception. Before we can identify where visual perceptions occur in the primate brain, we must first consider the structure of the primate visual brain. We then look for evidence of perception within this structure.

We first consider the physical properties and mechanical extra-ocular structure of the eye as inspiration for a synthetic platform. We also consider behavioural eye movements so that we may incorporate such behaviours into a synthetic model, and compare the resultant synthetic behaviours to that of primates. We look at the acquisition of visual data in the retina and investigate the structure of the visual brain by following the propagation of the retinal response to visual stimulus through the brain. In doing so, we may provide insight into structuring a synthetic vision processing architecture. For example, we observe the propagation of responses along somewhat separate channels in the brain, note the basic neuronal responses, and observe where serial and parallel processing occurs. As we shall see, following the propagation of response to retinal stimulus through the brain yields evidence of characteristics such as parallel processing, and the separation of stimulus into cues and other channels. We also consider how the brain constructs efficient representations of visual data, and the interaction between processing areas, including the forward flow and feedback between brain areas. When considering the physical structure of the visual brain, we do not consider what is perceived on more than a local functional level.

Having summarised the physical structure of the visual brain, we then consider the components of primate visual perception as inspiration for the development of the components of synthetic visual perception. We concentrate on those regions involved in dealing with kinematics, spatial awareness and attention. We present evidence of the components of perception, and look at existing literature for information about where components of perception such as spatial awareness and attention manifest within the visual cortex. We consider how the existence of such components of perception have been confirmed. By localising where perceptions occur within the visual brain, we can understand what input and output responses are likely to be involved in each component of perception.

Figure 2.1 shows the main constituents of the primate visual system, including:

- The eye, including intra-ocular components such as the retina, and extra-ocular components such as muscles.
- The optic nerve, chiasm and tract.
- The *lateral geniculate nucleus* (LGN) and optic radiations.
- The visual cortex.
- The dorsal and ventral streams.

## 2.2 The Primate Eye

For the purpose of gaining insight into the development of a synthetic vision mechanism, we review the eye, including its structure, function and ocular performance. Where not explicitly referenced otherwise, this section (Section 2.2) makes reference to biology texts such as [Rodieck (1998)].

The structure of the primate eye (Figure 2.2) can be divided into two main geographical segments: the anterior segment and the posterior segment. The anterior segment is the front third of the eye that includes the structures in front of the vitreous humour: the cornea, iris, ciliary body, and lens. The posterior segment is the back two-thirds of the eye that includes the anterior hyaloid membrane, vitreous humor, retina, choroid and optic nerve. The structure can also

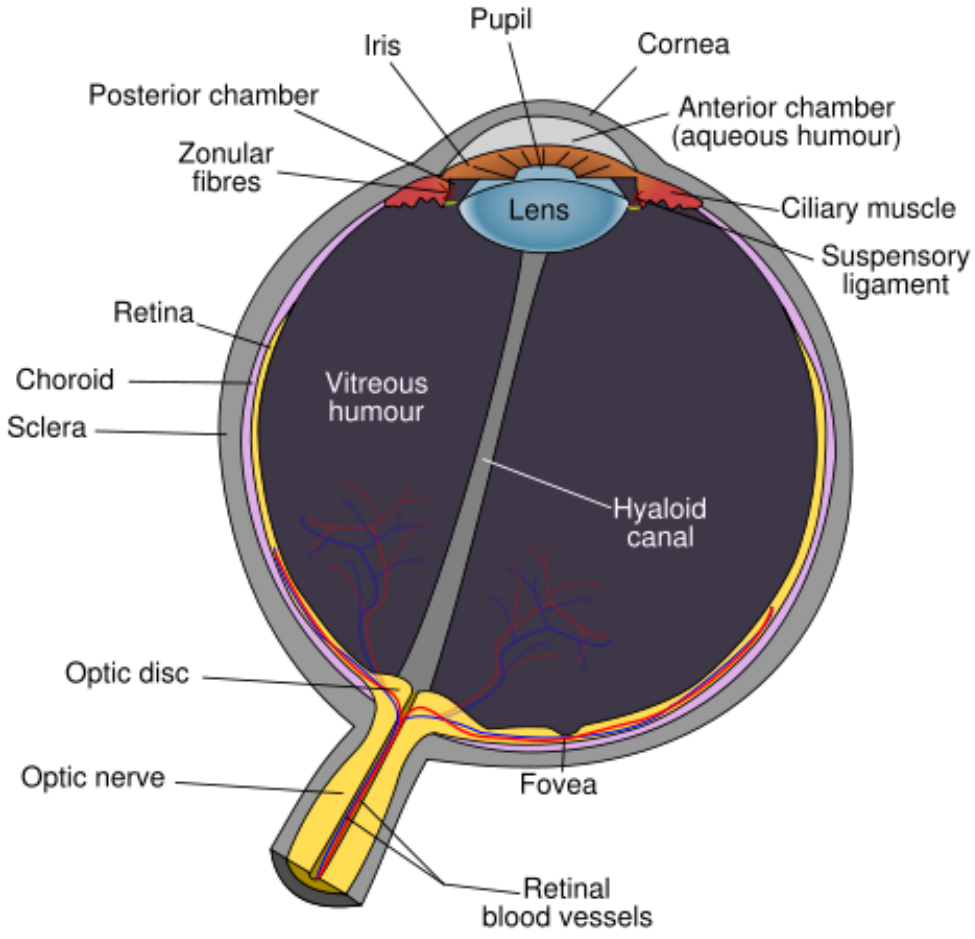



Figure 2.2: Structure of the primate eye.

be considered in terms of the appearance of its three main functional layers: the fibrous tunic contains the cornea and sclera; the vascular tunic which includes the iris, ciliary body and choroid; and the nervous tunic. The nervous tunic is the inner sensory region which includes the retina.

The retina contains photosensitive rod and cone cells and associated neurons. The retina is a relatively smooth layer with two distinct features - the fovea and optic disc. The fovea is a dip in the retina directly opposite the lens. It is largely responsible for sensing colour, and enables high acuity. The optic disc is a point on the retina where the optic nerve enters the retina to connect to the nerve cells on its inside. No photosensitive cells exist at the optic disc, which is why it is sometimes referred to as the anatomical blind spot. The pupil, lens and retina form an optical structure similar to that of the aperture, lens and focal plane of modern cameras. Both systems may be approximated by the pinhole camera model.

### 2.2.1 Interpreting Light - Retinal Structure

We now consider how the retina interprets projections of a scene. We can obtain insight into how it encodes images into efficient representations and channels for subsequent processing by the primate visual brain. An understanding of the function of the human eye serves to distinguish functions occurring a sensory level (retinal functions) as opposed to processing level (visual cortex functions). It also serves to benchmark human vision performance for comparison with synthetic vision systems.

Light enters the eye through the pupil. The lens focusses light on the retina causing chemical reactions in photosensitive cells, the products of which trigger nerve impulses that travel to the brain. The retina contains two forms of photosensitive cells that are structurally and metabolically similar - rods and cones. Rod cells are highly sensitive to light, but they cannot discriminate colour. Cone cells enable high visual acuity and need more intense light to elicit a response. Different cone cells respond to different wavelengths of light, which allow primates to perceive colour.

Rod cells contain the protein rhodopsin which is highly sensitive across the spectrum of visible light. Cone cells contain different proteins sensitive to each of the three primary colours: red, green and blue. When subjected to electromagnetic radiation, the proteins break down into two constituent products, creating ion channels on the cell membranes that hyperpolarise the cell leading to a release of transmitter molecules at the synapse.

The fovea, directly behind the lens, consists of mostly densely packed cone cells. Each cone cell is connected to a single bipolar cell, increasing detail and resulting in detailed visual acuity of the fovea, but also reducing low light sensitivity. The density of rod cells increases towards the periphery. Several rod cells are connected to a single bipolar cell, which then connects to a single ganglion cell that relays input to the visual cortex. This allows rods to accumulate input over an area for transmission at a single synapse. Figure 2.3 shows the interaction of the layers of cells in the retina.

There are two functional modes of ganglion cells that produce different outputs on the optic fibres: the on-centre and off-centre responses. For both, the strongest ganglion response is elicited when the spatially central region of the receptive field of the ganglion cell experiences opposite stimulus to its surroundings. This occurs, for example, when the centre is illuminated but the surroundings are not, or vice versa. Figure 2.4 shows the modes of stimulation, and the respective outputs, of on-centre and off-centre ganglion cells.

Each ganglion cell produces either an on-centre or off-centre opponency response. The response comes from either light-dark illumination opponency (from rod cells), or red-green or blue-yellow colour opponency (from cone cells). A single ganglion cell can take input from either rods, cones or both. In the fovea, ganglion cell inputs are mostly from cones (colour opponency). Towards the periphery, the ganglion inputs are predominantly from rods (light-dark opponency). Where ganglion cells take input from both rods and cones (in particular, within the fovea), their output is either colour opponency *or* light-dark opponency. The type of output elicited depends on the level of ambient lighting. In low lighting conditions, the outputs are predominantly from light-dark responses from rod cells. In good lighting, the outputs are predominantly colour opponency from cone cells.

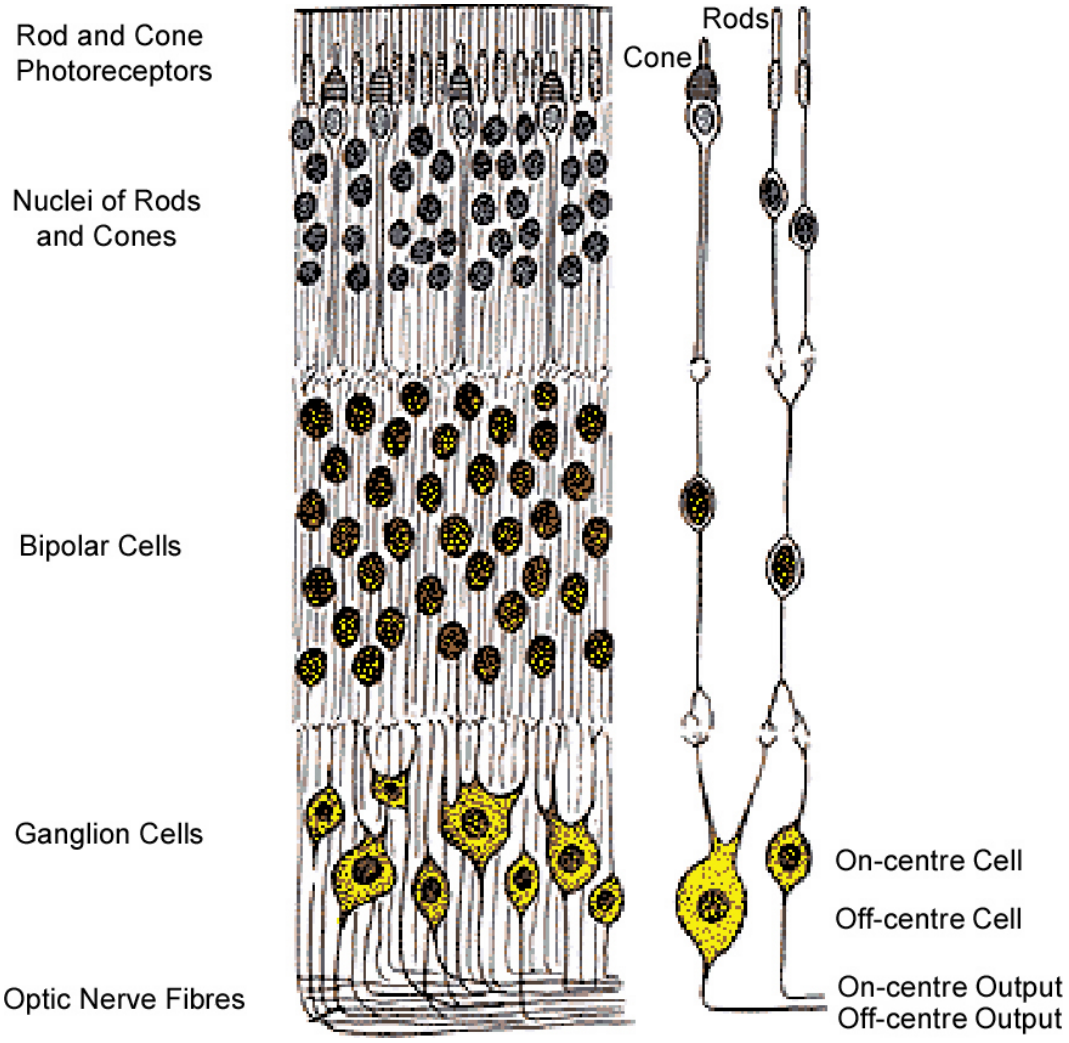



Figure 2.3: Layers in the retina. Light enters from bottom.

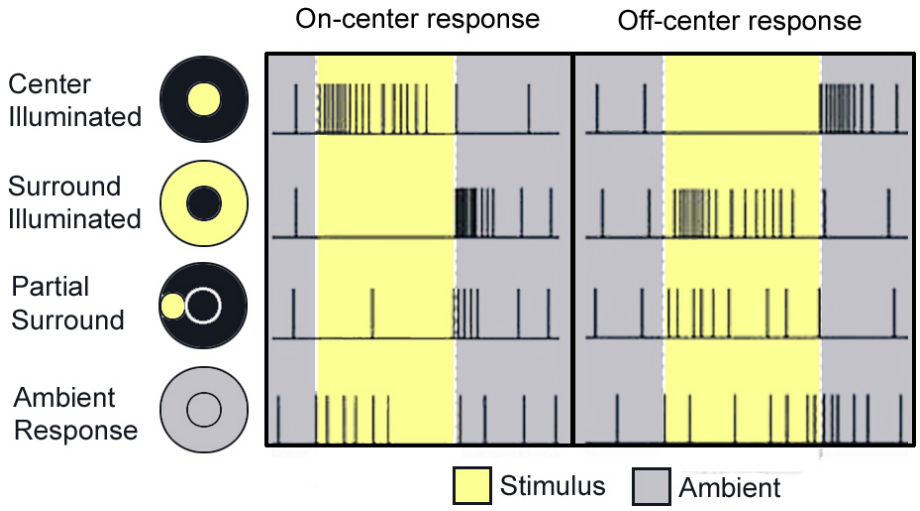



Figure 2.4: Output firing of ganglion cells during different modes of stimulation of its receptive field, as described by [Rodieck (1998)].

There are several classes of ganglion cells, named according to the the type of photoreceptors they connect to and the brain locations to which the optic fibres project the output. Namely:

- Midget (Parvocellular, or P pathway) ganglions receive inputs from relatively few rods and cones over a small centre-surround receptive field. They primarily respond to changes in colour but respond weakly to changes in contrast unless the change is great [Kandel *et al.* (2000)]. They project to the parvocellular layers of the LGN. About 80% of retinal ganglion cells are midget cells.

- Parasol (Magnocellular, or M pathway) ganglions receive input from numerous rod and cone cells. They respond well to contrast (even low contrast) stimuli, but are not very sensitive to changes in colour [Kandel *et al.* (2000)]. They have much larger centre-surround receptive fields. They project to the magnocellular layers of the LGN. About 10% of retinal ganglion cells are parasol cells.

- Bistratified (Koniocellular, or K pathway) ganglions project to the koniocellular layers of the LGN. About 10% of retinal ganglion cells are bistrat-

ified. They receive inputs from intermediate numbers of rods and cones. They have moderate spatial resolution and can respond to moderate contrast stimuli. They may be involved in colour vision. They have very large receptive fields that only have centres (no surrounds).

Other smaller populations of ganglion cell types exist. These project to the *suprachiasmatic* nucleus (SCN) for moderating circadian rhythms; to the superior colliculus for controlling eye movements; and to the *Edinger-Westphal* nucleus (EW) for control of the pupillary light reflex [Kandel *et al.* (2000)].

### 2.2.2 Retinal Performance

Retinal performance parameters may be used to quantify the visual abilities that humans have evolved. It is desirable that a synthetic primate vision sensor exhibits similar performance. The acuity and dynamic response of the human eye have been determined using various metrics.

**Visual Acuity**: *Cycles per degree* (CPD), is the most common method used by optometrists to measure human angular resolution. The test involves discriminating equal width black and white lines at a distance of 1m. Various estimates have been documented, reporting 'average' human performances ranging between 60CPD [Curcio *et al.* (1990)] and 150CPD [Campbell & Green (1965)], the former corresponding to line widths of 0.93mm at a distance of 1m for the fovea centralis. The angle of sharp foveal vision is just a few degrees (typically $\sim 2^o$) over which 60CPD corresponds to a resolution of approximately one megapixel.

**Equivalent Resolution**: The perception of wide and sharp human vision is based on actively turning the eyes towards multiple points of interest in the field of view. The brain augments resolute foveal imagery over time and spatial saccades into the unified perception. Various estimations of equivalent resolution have been conducted by extrapolating the resolution of the fovea centralis over the entire visual field. For example, for a 'conservative' square field of view of $120^o$, it has been estimated as from 81 megapixels [Fischer & Tadic (2000)] to 576 megapixels [Clark (2005)]. The attentional scanpaths chosen by the observer

in building such a resolute perception, as we shall discuss later, is highly dependent on scene content.

**Dynamic range**: It is generally accepted that the human retina has a static contrast ratio of around 100:1, and a total dynamic contrast ratio of about 1,000,000:1, depending on illumination. The eye re-adjusts exposure sensitivity both chemically, and by adjusting the iris. Peak adaptation typically occurs within 30 minutes. The adjustment rate is non-linear and often interrupted by illumination variations or saccades. Significant adaptation takes place within seconds of perturbation.

**Spectral response**: The visible spectrum of light ranges approximately from 400 to 700nm.

Modern video cameras commonly exhibit a broader spectral response than humans. The human periphery extends over a field of view of approximately $140^o$. Pinhole model cameras can achieve fields of view up to $180^o$, depending upon lens choice. Camera acuity depends largely upon zoom lenses, but the resolution of existing CCD/CMOS sensors per unit area is significantly lower than that of the human retina. An active vision system's equivalent resolution would depend upon the image sampling resolution as well as methods used to integrate images across pan and tilt motions.

### 2.2.3 Extraocular Structure - Eye Motion

An understanding of the extraocular structure of the eye potentially provides insight into factors important in the design of a synthetic active vision mechanism. It reveals the degrees of freedom and the ranges of motion that evolution considers important for vision in the real world.

Each eye has six muscles that control its movements: the lateral rectus, the medial rectus, the inferior rectus, the superior rectus, the inferior oblique and the superior oblique. The actuating muscles of the human eye are shown in Figure 2.5. When the muscles exert different tensions, a torque is exerted on the

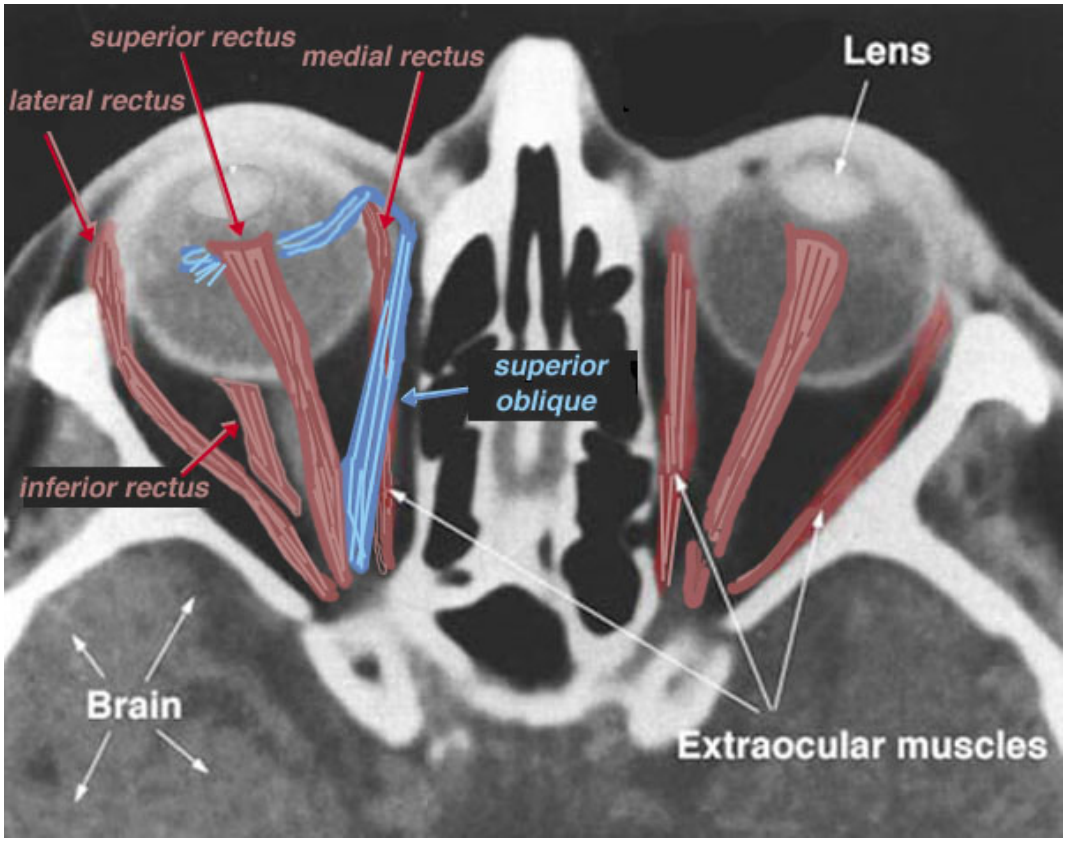



Figure 2.5: Actuating muscles of the human eye.

eye that causes it to move. This is an almost pure rotation, with only about 1mm of translation. The eye is commonly approximated as undergoing rotations about a single point in the centre of the eye.

#### 2.2.3.1 Agility

Human eyes are capable of tracking objects moving across the field of view at up to around $100^o/s$. Attention shifts can achieve velocities around $400^o/s$. Humans are typically capable of performing a maximum of three point-to-point-return fixations per second. They can typically focus on surfaces from around 10cm distance up to infinity. An approximate control resolution for eye motions is around $0.2^o$, however visual perception may provide the sensation of greater accuracy (see Section 2.2.3.2).

#### 2.2.3.2 Behavioural Eye Movements

Primates exhibit various instinctual and reflexive eye motions that help to stabilise the projection of a scene onto the retina. Knowledge of these behaviours provides useful information for comparison with those elicited by a synthetic system. It also highlights the usefulness of active eye motion and provides insight into how primates control gaze. Behavioural eye movements include:

**Vestibulo-ocular reflex**: This is the ability to physically counteract shifts in the projection of a scene onto the retina during head movement by instinctively producing an eye movement in the direction opposite to such head movements, thus preserving the image on the centre of the visual field. For example, when the head turns to the right, the eyes turn to the left, and vice versa.

**Smooth pursuit**: The is the ability to track moving objects. Tracking updates occur with less accuracy and at a slower rate than the vestibulo-ocular reflex, as pursuit requires cognitive processing of incoming visual information and the supply of feedback. During smooth pursuit the eyes will often 'jump' to keep up with a moving target and to correct tracking errors.

**Saccades**: Rapid simultaneous movements of both eyes towards the same visual target.

**Microsaccades**: When concentrating on a single scene point, though apparently stable, gaze occasionally exhibits tiny saccades. This motion is thought to stimulate individual retinal photoreceptor cells that would otherwise stop generating output. Microsaccades typically move the eye less $0.2^o$.

**Vergence**: Binocular attention towards an object involves rotating the eyes around a vertical axis so that the projection of the image is aligned in the centre of the retina of both eyes. In observing an object closer than the previously attended location, the eyes converge; for an object farther away, they diverge. Vergence movements are closely connected to focal adjustments (*accommodation*). A shift

in fixation to an object at a different depth will involve both vergence and accommodation.

**Optokinetic reflex**: When in motion, the optokinetic reflex cyclically saccades and smoothly pursues stationary objects as they are passed.

## 2.3 Structure of the Visual Brain

Figure 2.1 shows the main areas involved in acquiring, transferring and processing visual information in the primate brain. We now look at physical connections in the visual brain and the associated flow of information. The processing response to retinal visual stimulus first propagates from the retina to the visual cortex. From the visual cortex it propagates into the dorsal and ventral streams. In this section we summarise the flow of information along these streams and look briefly at the specific function of regions in the visual cortex. Where not explicitly referenced otherwise, this section (Section 2.3) makes reference to biology texts such as [Kandel *et al.* (2000)].

### 2.3.1 From the Retina to the Visual Cortex

Neural responses to visual stimuli propagate to the visual cortex according to where they originated in the retina. This provides further evidence for the preservation of separate channels during processing in the visual brain, according to how channels were generated in the retina. For example, the response generated by K cells (colour) in the left retina are coherently transferred to a different processing area than responses generated by right retinal M cells (contrast/texture).

The main components involved in propagating retinal responses from the eye to the visual cortex are:

**Optic nerve:** About 90% of retinal ganglion cells transfer information to the brain via axons along the optic nerve.

**Optic chiasm:** The optic nerves from both eyes meet and cross at the optic

chiasm, at the base of the hypothalamus. At this point the information coming from both eyes is combined and then splits according to the visual field. The corresponding halves of the field of view (right and left) are sent to the left and right halves of the brain, respectively, to be processed. That is, the right side of primary visual cortex deals with the left half of the field of view from both eyes, and similarly for the left brain [Nolte (2002)]. A small region in the centre of the field of view is processed redundantly by both halves of the brain.

**Optic tract**: Information from the right visual fields of each eye travels along the left optic tract. Similarly, information from the left visual fields travels along the right optic tract. Each optic tract terminates in the LGN.

**Lateral geniculate nucleus LGN**: The LGN is a sensory nucleus in the thalamus of the brain. It consists of six layers in humans and most other primates [Nolte (2002)]. Layers one, four, and six correspond to information from one eye; layers two, three, and five correspond to information from the other eye. Layer one connects to the M cells (depth and/or motion) of the optic nerve of the opposite eye. Layers four and six also connect to the opposite eye, but to the P cells (colour and edges) of the optic nerve. In contrast, layers two, three and five connect to the M cells and P cells of the optic nerve for the same side of the brain as its respective LGN. In between the six layers are smaller cells that receive information from the K cells (colour) in the retina. The output of the LGN propagates to the primary visual cortex (V1) via the optic radiations. Recent research suggests that some modulation responses are elicited in the LGN, and that it may not merely function as a relay nucleus [Sherman (2006)].

**Optic radiations**: Information is transferred from the LGN to the visual cortex via the optic radiatoins. The P layer neurons of the LGN relay to the V1 layer known as $4C\,\beta$. The M layer neurons relay to the V1 layer known as $4C\,\alpha$. The K layer neurons in the LGN relay to large neurons called *blobs* in layers *2 and 3* of V1.

There is a direct correspondence from an angular position in the field of view of the eye, all the way through the optic tract to a nerve position in V1. From there, more cross-connections exist within the visual cortex.

### 2.3.2 Visual Cortex

In this section, we follow the response to visual stimulus as it propagates through the visual cortex. This enables us to understand the extent of the preservation of the initial separation of visual responses according to the retinal origin such as colour or intensity channels, or left and right visual field channels. We may also extract information about modulation and feedback, which provides insight into how some visual functions are affected by higher brain areas and cognition, and the sequencing of processing. The main processing areas in the visual cortex include the primary visual cortex (V1), V2, V3, V4 and MT (V5).

#### 2.3.2.1 Primary Visual Cortex (V1)

The primary visual cortex (V1) is the earliest cortical visual area. Visual information relayed to V1 is more or less coded in terms of local contrast levels. There is a well-defined spatial mapping of the visual image from retina to V1 - even the blind spots are mapped into V1. V1 is divided into six functionally distinct layers. Layer four receives most of the visual input from the LGN.

V1 processes information about static and moving objects, and pattern recognition. Neurons in V1 have the smallest receptive field size of any visual cortex region, perhaps for the purpose of accurate spatial encoding. Individual V1 neurons tune to one of the two eyes (ocular dominance). Early neuron responses (up to 40ms in propagation time into V1) are thought to consist of tiled sets of selective spatiotemporal filters that can discriminate small changes in spatial frequency, orientation, motion, speed, colour, and other spatiotemporal features. In the visual cortex in general, neurons with similar tuning properties tend to cluster together as cortical columns. The exact organisation of such cortical columns within V1 is not known.

Later in propagational time (beyond 100ms into V1), neurons are progressively more sensitive to the global organisation of the scene [Lamme *et al.* (2000)]. This

property may stem from recurrent processing (the proximity influence of higher-tier cortical areas) and lateral connections. As information relayed beyond V1, it is increasingly non-local in character.

#### 2.3.2.2 V2

Visual area V2 is the second major area in the visual cortex. It receives strong feed-forward connections from V1 and sends strong connections to V3, V4 and V5. It also sends strong feedback connections to V1. V2 is split into four quadrants: a dorsal and ventral quadrant in each of the left and the right hemispheres. Together, these four regions provide a complete spatial map of the visual world.

Functionally, V2 has many properties in common with V1. Cells are tuned to simple properties such as orientation, spatial frequency and colour. The responses of many V2 neurons are also modulated by more complex properties, such as the orientation of illusory contours, and whether the stimulus is part of the foreground or the background [Qiu & von der Heydt (2005)].

#### 2.3.2.3 V3

Visual area V3 is the cortical area located immediately in front of V2. A distinction is often made between "dorsal V3" and "ventral V3" (or ventral posterior area, VP) according to their upper and lower cerebral hemisphere locations respectively. They also have distinctly separate connections with other parts of the brain, appear physically different, and contain neurons that respond to different combinations of visual stimulus (for example, colour-selective neurons are more common in ventral V3).

Dorsal V3 is normally considered to be part of the dorsal stream, mainly receiving inputs from V2, and projecting to the posterior parietal cortex. Other studies prefer to consider dorsal V3 as part of a larger area, named the *dorsomedial* area (DM), which is thought to contain a representation of the entire visual field. Work with *functional magnetic resonance imaging* (fMRI) has suggested that area V3 may play a role in the processing of global motion [Braddick & O'Brian (2001)]. Neurons in area DM respond to coherent motion of large patterns covering extensive portions of the visual field.

VP has much weaker connections from the primary visual area, and stronger connections with the inferotemporal cortex. It was originally thought that VP contained a representation of the upper part of the visual field but it is now thought that it contains a complete visual representation.

#### 2.3.2.4 V4

V4 is the third cortical area in the ventral stream, receiving strong input from V2 and projecting strong connections to the *posterior inferotemporal* (PIT) cortex. It also receives inputs from V1, especially for central space. It has weaker connections to V5 and some other areas.

V4 is the earliest area in the ventral stream shown to be modulated by attention [Moran & Desimone (1985)]. Most studies indicate that selective attention can change firing rates in V4 by about 20%. Like V1, V4 is tuned to orientation, spatial frequency and colour. It is also tuned for object features of intermediate complexity, like simple geometric shapes. The full extent of tuning in V4 is not known. Unlike other areas in the inferotemporal cortex, V4 is not tuned for complex objects such as faces.

#### 2.3.2.5 V5/MT

There is uncertainty as to the exact function of area visual area V5, also known as visual area MT (middle temporal). It is thought to play a major role in the perception of motion, the integration of local motion signals into global precepts, and the guidance of some eye movements [Born & Bradley (2005)].

MT is connected to numerous cortical and subcortical brain areas. V1 provides the strongest feed-forward connection to MT. Input is also received from V2, dorsal V3 and the koniocellular regions of the LGN. MT sends its major outputs to areas located in the cortex immediately surrounding it, including the floor of the superior temporal sulcus (FST), the superior temporal area (MST) and V4*t* (the middle temporal crescent). It also projects to the eye movement-related areas of the frontal and parietal lobes.

## 2.4 Perception in the Primate Visual Brain

We have described the major physical connections between areas in the visual brain, and the general increase in functional complexity that occurs as responses propagate through these areas. We now consider the development of perception in the visual cortex, and along the dorsal and ventral streams. We concentrate on cue extraction and modulation, and the components of spatial perception and attention. We do not consider higher cognition such as recognition. We summarise neurobiological and psycho-physical evidence for the existence and interaction/ordering of such components. This process highlights the importance of such components in human perception. Moreover, it provides insight into the integration of relevant components into a primate-inspired system.

### 2.4.1 Early Visual Cues

Pre-attentive computation of visual cues occurs across the entire visual field [Itti & Koch (2001)]. Cue processing takes around 25-50ms. Cue feature maps are computed in parallel but separate cortical streams (hypercolumns) [Dacey (1996)], and computation is not solely feed-forward. Top-down modulation of cue pre-attentive cue processing is known to occur, affecting things like cue priority and cue sensitivity. Pre-attentive computation has been shown to occur persistently; neurons involved in pre-attentive cue processing fire vigorously even if the subject is attending away from the receptive field or is anesthetised [Treue & Maunsell (1996)]. Neuronal tuning becomes increasingly specialised with progression from low to mid-level visual areas. Mid-levels include those that respond to corners or junctions [Pasupathy & Connor (1999)], shape-from-shading [Braun (1993)], and basic object recognition.

As described, neurons at the earliest stages are known to respond to simple features such as intensity contrast, colour, orientation, motion and stereo disparity. Spatial feature contrast is important, not local absolute feature strength [Nothdurf (1990)]. Early visual neurons are tuned to spatial contrast in cues, and neuronal responses are strongly modulated by context, in a manner that extends far beyond the range of the classical receptive field (CRF) [Allman *et al.*

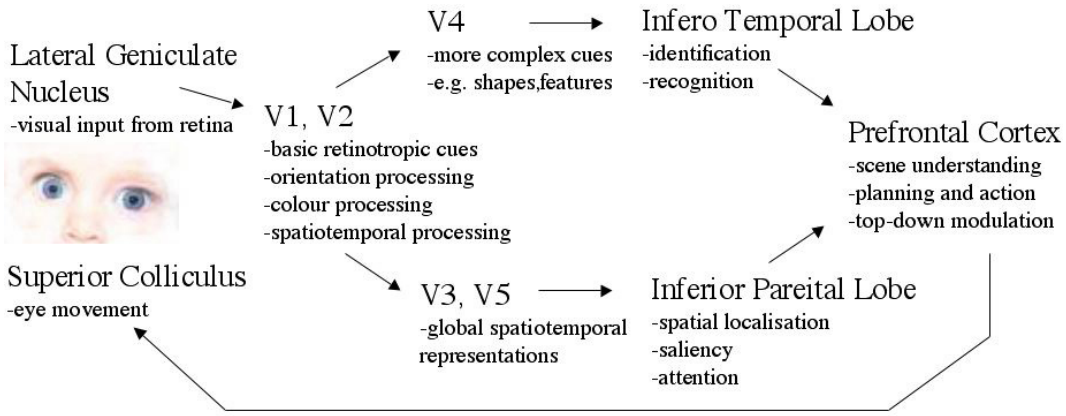



Figure 2.6: Main forward propagation of responses to visual stimulus through the human brain. The ventral stream passes through the inferotemporal lobe, the dorsal stream through the parietal lobe. Not all pathways shown.

(1985)]. A broad inhibitory effect occurs when a neuron is excited with its preferred stimulus but that stimulus extends beyond the neuron's CRF. Conversely, little inhibition occurs when the stimulus is restricted to the CRF, and surrounds contain non-preferred stimulus [Sillito *et al.* (1995)]. Additionally, long-range excitatory connections in V1 appear to enhance responses of orientation-selective neurons when stimuli extend to form a contour [Gilbert *et al.* (2000)]. A result is that monkeys exhibit sparse activity when viewing complex natural scenes, compared to the vigorous response elicited by small laboratory stimuli in isolation. These observations point towards non-classical surround modulation.

### 2.4.2 Perception in the Dorsal and Ventral Streams

After reaching the visual cortex, responses propagate towards 'higher' levels along two general pathways - the dorsal and ventral streams (Figure 2.6). The dorsal stream connects the visual cortex to the posterior parietal lobe. The ventral stream connects the visual cortex to the inferotemporal lobe.

The dorsal stream is primarily involved in spatial localisation and directing gaze towards objects of interest in a scene. The control of attention is believed to take place in the dorsal stream. The dorsal stream is often generalised as the

"where" stream [Rodieck (1998)].

The ventral stream is mainly concerned with recognition and identification. It is also involved in the representation of the attended objects that pass through the attentional bottleneck. Although probably not directly involved with the control of attention, the ventral stream areas have been shown to receive attentional feedback modulation [Moran & Desimone (1985)]. The ventral stream is often generalised as the "what" stream.

The dorsal and ventral streams interact because scene understanding involves both recognition and spatial deployment of attention. Some such interactions occur in the prefrontal cortex, which is bi-directionally connected to both the inferotemporal cortex and the posterior parietal cortex [Kandel *et al.* (2000)]. The prefrontal cortex is responsible for planning and action and as such has a role in modulating, via feedback, the dorsal and ventral processing streams.

### 2.4.3 Spatial Perception

Humans experience a rich egocentric perception of scene structure and motion. Mechanisms of spatial updating maintain accurate representations of visual space across eye movements. We now consider how and where this perception develops within the brain.

#### 2.4.3.1 Motion Perception

First order motion perception refers to the perception of the motion of an object that differs in luminance from its background, such as a black bug crawling across a white page. Second order motion occurs when a moving contour is defined by contrast, texture, flicker or some other quality that does not result in an increase in luminance or motion energy of the stimulus. There is evidence to suggest that early processing of first and second order motion is carried out by separate pathways [Nishida *et al.* (2001)]. As described earlier, individual neurons early in the visual system (LGN, V1 and even V3) respond to motion that occurs locally within their receptive field. Each local motion-detecting neuron may suffer from

the aperture problem [1], so that the estimates from many neurons need to be integrated into a global motion estimate. This appears to occur in area MT/V5 in the human visual cortex where first and second order signals appear to be fully combined [Kandel *et al.* (2000)].

While the eye is stationary, primates can perceive relative velocities of scene surfaces with high accuracy. However, during eye movements accuracy is reduced. When a non-fixated object moves towards or away from an observer without being attended, the ability to discern absolute and relative velocities is still present, although not as accurate, via disparity. Velocity estimations also improve with lighting intensity.

#### 2.4.3.2 Depth Perception

Depth perception is the visual ability to perceive the world as three-dimensional. *Stereopsis* is depth perception from binocular vision that exploits parallax disparities. Animals that have their eyes placed frontally can also use information derived from the different projection of objects onto each retina to judge depth. By using two images of the same scene obtained from slightly different angles, it is possible to triangulate the distance to an object with a high degree of accuracy. If an object is far away, the disparity of that image falling on both retinas will be small. If the object is close or near, the disparity will be large.

In the 1980s, neurons were found in V2 of the monkey brain that responded to the depth of random-dot stereograms [Poggio *et al.* (1988)]. It is now known that numerous visual brain areas contain neurons involved in depth perception. Recent experimental results [Neri *et al.* (2004)] determined that dorsal areas (V3A, MT/V5, V7) show more adaptation to absolute than to relative disparity; ventral areas (hV4, V8/V4) show an equal adaptation to both; and early visual areas (V1,

[1] Each neuron in the visual system is sensitive to visual input in a small part of the visual field, as if each neuron is looking at the visual field through a small window or aperture. The motion direction of a contour is ambiguous, because the motion component parallel to the line cannot be inferred based on the visual input. This means that a variety of contours of different orientations moving at different speeds can cause identical responses in a motion sensitive neuron in the visual system.

V2, V3) show a small effect in both experiments. These results indicate that processing in dorsal areas may rely mostly on information about absolute disparities, while ventral areas split neural resources between the two types of stereoscopic information so as to maintain an important representation of relative disparity.

Primate depth perception benefits from binocular vision, but it also uses numerous other monocular cues to form the final integrated perception. Monocular cues that contribute to the perception of depth include:

**Motion parallax**: When an observer moves, the apparent relative motion of several stationary objects against a background gives hints about their relative distance. This effect can be seen clearly when driving in a car - nearby things pass quickly, while far-off objects appear almost stationary.

**Kinetic depth perception**: As objects in motion recede into the distance they appear to become smaller. Conversely, uniformly expanding objects appear to be coming closer. Kinetic depth perception enables the brain to calculate time to collision (TTC) assuming a particular velocity.

**Perspective**: Parallel lines converge at infinity, allowing us to reconstruct the relative distance of two parts of an object, or of landscape features.

**Relative size**: Objects that are close to us look larger than similar objects far away; our visual system exploits the relative size of similar (or familiar) objects to judge distance.

**Focus**: The lens of the eye can change its shape to bring objects at different distances into focus. Knowing at what distance the lens is focused when viewing an object means knowing the approximate distance to that object.

**Accommodation**: This is an oculomotor cue. When focussing on distant objects, ciliary muscles stretch the eye lens, making it thinner. The kinesthetic sensations of contracting and relaxing ciliary muscles (intraocular muscles) are sent to the visual cortex where they contribute to interpreting depth [Zajac (1960)].

**Convergence**: This is also an oculomotor cue. By virtue of stereopsis the two eyes can converge on the same object. The angle of convergence is larger when the eye is fixating on far away objects. The convergence will stretch the extraocular muscles. Kinesthetic sensations from these extraocular muscles also help in depth/distance perception.

**Shading**: The intensity of the surface of 3D objects changes according to the location of light sources. The location of light sources and the shape of a surface can be inferred from such intensity gradients.

**Occlusion**: Depth ordering of objects can be inferred by occlusions.

**Texture gradient**: Textures (for example, that of an area of grass at one's feet) transition to a more homogonous appearance (that cannot be clearly discerned as textured) with increased distance.

Other monocular phenomenon can also contribute to the perception of depth. The colour of distant objects may, for example, be shifted towards the blue end of the spectrum (for example, distant mountains). Due to light scattering by the atmosphere, objects that are a great distance away may also look hazy.

Of all the above cues, only convergence, focus and object familiarity provide direct absolute distance information. When combined with gaze geometry information from kinesthetic feedback, disparity estimation provides absolute depth information. It is likely that primates use depth estimates from absolute disparity to *ground* the relative estimations provided by other depth relative depth cues.

### 2.4.4 Attention

Primates use foveal attention with rapid eye movements to analyze complex visual inputs in real time [Vidyasagar (1999)]. Attention breaks down scene understanding into a rapid series of computationally less demanding, localised visual analysis

problems. It selectively reduces the quantity of visual input that reaches short-term memory and visual awareness [Desimone & Duncan (1995); Crick (1998)]. In this section, we present observations of the components of primate visual attention based on the thorough review of the field provided by Itti & Koch (2001).

Psycho-physical experimentation confirms the propensity for primates to respond to unique/salient features (bottom-up attention). Conversely, during the execution of a task, attention can be likened to a "stagelight" that strategically illuminates different scene regions for specific analysis as they are reasoned to be interesting with respect to the current task (top-down attention). In the latter case, attention is task-dependent, and can vigorously modulate early visual processing in both spatial and feature-specific manners [Reynolds *et al.* (2000); Weichselgartner & Sperling (1987)]. In both instances, attention implements a bottleneck that reduces the quantity of visual information to be processed. Top-down attention involves cognition and is slower than bottom-up attention. Psycho-physical experiments can also provide insight into how such feature maps are computed. For example, Zetzsche used eye trackers to observe that humans preferentially fixate on regions with multiple orientations, such as corners [Zetzsche (1998)]. Zetzsche then computed feature maps that highlighted these preferential features using Gabor wavelets. Similarly, spatial contrast neuronal responses can be synthesised using a DOG pyramid approximation.

Attention can be involved in triggering visual and/or physical behaviours, and is intimately related to recognition, planning and motor control [Miller (2000)]. Detection of visual stimulus may initiate instinctual reactions. Further, it has been shown that primates can recognise objects by explicitly replaying a sequence of eye movements and matching expected features with those observed [Rybak *et al.* (1998)] . As such, purposeful gaze direction is selected based upon what features are *expected* at particular spatial locations (based on past experience), not merely on the cue responses elicited by the visual stimulus that actually *is* at that location. This observation supports the notion of top-down modulation of early cues, but such modulation is used to verify or oppose the existence of *expected* features, rather than for detecting *present* features.

Gaze-directed attention is not necessarily mandatory for early vision: humans can make simple judgements about objects they are not attending [DeSchepper

& Treisman (1996)], but these judgements may be less accurate than attended objects. This type of judgement is often referred to as *covert attention*.

#### 2.4.4.1 Evidence of Attentional Maps

Most of the early visual processing areas participate in attention. A possible explanation for widespread attentional activity throughout most visual areas could be that some neurons in all those areas are concerned with the explicit computation of saliency, but are found at different stages along the sensory-motor processing stream [Itti (2005)]

Single unit recordings in the visual system of the macaque monkey indicate the existence of a number of maps of the visual environment that appear to encode salience, and/or the behavioural significance of targets [Murthy *et al.* (2001)]. Such maps were concluded to exist in the superior colliculus, infero and lateral subdivisions of the pulvinar, the frontal eye fields, and areas within the intraparietal sulcus. Numerous neural correlates throughout the human brain, including areas in the lateral intraperietal sulcus of the posterior parietal cortex, the frontal eye fields, the inferior and lateral subdivisions of the pulvinar, the superior colliculus, the retina and the LGN suggest it is not likely that a single or centralised saliency map exists [Kustov & Robinson (1996); Gottlieb *et al.* (1998); Suder & Worgotter (2003)].

It is generally accepted that early visual features or cues are computed in topographic feature maps in V1. Salience may then be expressed as a modulation onto such feature responses [Zhaoping (2005)]. Further, Desimone and Duncan suggest that salience is not explicitly represented by specific neurons, instead it is implicitly encoded in a distributed modulatory manner across the various feature maps [Desimone & Duncan (1995)].

These neurons are found in different parts of the brain that specialise in different functions, so they may encode different types of salience. [Navalpakkam *et al.* (2005)] propose that the posterior parietal cortex encodes a visual salience map; the pre-frontal cortex encodes a top-down task relevance map; and the final eye movements are subsequently generated by integrating information from both regions to form an attention guidance map possibly stored in the superior colliculus.

#### 2.4.4.2 Integration of cues for attention

Attention is thought to activate a *winner-take-all* competition amongst neurons tuned to different orientations and spatial frequencies within one cortical hypercolumn [Lee *et al.* (1999); Carrasco *et al.* (2000)]. Different cue features contribute with different strengths to perceptual saliency [Braun & Julesz (1998)]. This relative feature weighting can be influenced by top-down modulation and training. Within a broad feature dimension, strong local interactions between filters (for example, various orientations within the general orientation feature) have been characterised via neuronal correlates [Carandini & Heeger (1994)]. Less evidence exists for within-feature competition across different spatial scales [Itti & Koch (2000)].

#### 2.4.4.3 Inhibition of Return and Attentional Memory

It is necessary for efficiency to have some coarse short-term knowledge of past fixation locations so as to reduce the likelihood of unnecessarily returning to the same scene location. Inhibition of return (IOR) encapsulates the notion that the gaze is temporarily prevented from unnecessarily re-attending saliency maxima. Experimental support exists for transiently inhibiting neurons in the saliency map at the currently attended location [Klein (2000)]. In the intraparietal sulcus of monkeys, the activity of spatially-tuned neurons at salient locations was shown to be transferred to other neurons according to eye motion [Merriam *et al.* (2003)]. A short-term inhibitory effect then prevents previously attended stimuli from being immediately re-attended.

Horowitz and Wolfe proposed that visual search is memoryless - when elements of a search array randomly reorganised while subjects searched for a specific target, search efficiency was not degraded [Horowitz & Wolfe (1998)]. Performance gains for searches on a stable array would indicate memory use. However, this may just preclude perfect memorisation and does not necessarily preclude the possibility that the last few attended locations are remembered, in accordance with the limited lifespan of IOR.

Kahneman and Treisman proposed that a short-term memory maintains information about visual features and their locations ("object files") across saccades

[D Kahneman (1984)]. Psycho-physical experimentation suggests that up to three or four object files may be retained [Irwin & Zelinsky (2002)]. Wilson provides evidence suggesting two concurrent, dissociated types of memory: one stores object features, the other stores spatial location information [Wilson *et al.* (1993)].

#### 2.4.4.4 Task-Dependency and Top-down Modulation

In the first few hundred milliseconds after viewing novel stimulus, bottom-up uniqueness detection across cue features may well describe how attention is deployed. However, a more complete primate model must include top-down biasing. Top-down modulation is thought to be controlled from higher areas including the frontal lobes, which are known to connect directly to the visual cortex and earlier visual areas. Responses along these more cognitive pathways take 200ms or more, comparable to the time required to effect eye motion. Task-dependency modulates neural activity by enhancing the response of early stage visual neurons tuned to the location and features of a stimulus [Desimone & Duncan (1995)].

As discussed previously, attention can be seen as "stagelight" successively illuminating different regions as they are reasoned to be interesting [Weichselgartner & Sperling (1987)]. Such feedback is believed to be essential for binding the different visual attributes of an object, such as colour or form, into a unitary precept [Trieisman & Gelade (1980)]. Knowledge of an object enhances its extraction from visual clutter - the unitary precept suggesting also that cues and features are bound to the representation of an object. The known features of an object help us identify it, and look for other expected identifying features to verify initial perception. Attention is involved in selecting a location of interest for evaluation, and enhances the cortical representation of the object at that location.

#### 2.4.4.5 Top-down Search

During search, knowledge of a target amplifies its salience. The prefrontal cortex implements attentional control by amplifying task-relevant information relative to distracting stimuli [Nieuwenhuis & Yeung (2005). For example, vertical lines are more salient if we are looking for them [Blaser *et al.* (1999)]. A better knowledge of a target also leads to faster search. For example, an exact picture of a target will

facilitate a faster search than a semantic description [Kenner & Wolfe (2003)]. Similarly, Wolfe proposed that top-down knowledge emphasises features which may distinguish a target from surrounding clutter [Wolfe (1996)] . For example, if searching for a red object, the contribution of colour cue would be emphasised while orientation cue may be reduced.

Triesman showed that there is a performance distinction between "pop-out" and "conjunctive" search tasks. Specifically, that a conjuctive search task (for example, colour and orientation: find red vertical oriented target amongst red horizontal oriented targets) is slower than pop-out search task (for example, a red target amongst green surroundings) [Trieisman & Gelade (1980)]. This eliminates the possibility of primates generating new composite features on-the-fly, and imposes constraints on possible biasing mechanisms.

Humans achieve nearly optimal search performance even though they integrate information poorly across fixations [Najemnik & Geisler (2004)]. This suggests that there is little benefit from perfect integration across fixations, rather, that efficient processing of information during each fixation is more important. Visibility peaks at the optical centre where more identification processing resources are allocated, so foveal saliency is better trusted for verification of the presence/absence of a feature/object. As such, a visual surface is attended for confirmation that it conforms to search criteria. In terms of search efficiency, it may be necessary to have some coarse record of such past fixation locations so as to reduce the likelihood of unnecessarily returning gaze to the same scene region [Itti & Koch (1998)].

#### 2.4.4.6 Contextual Search

Contextual information is known to guide eye movements in primate attention [Oliva (2005)]. The "gist" of a scene (for example, "road scene", "beach", etc) is thought to be computed rapidly (within 150ms of scene onset), but supporting neural correlates of this computation are yet to be revealed [Rullen (2003)]. Scene gist is believed to be used as a contextual guide in search and attention schema to compliment target saliency. For example, if searching for a car in a road scene, one would expect to find it on the road, and would search that area preferentially.

## 2.5 Summary

We have reviewed components of primate vision useful for developing synthetic primate vision systems. Primates benefit from active vision in several ways. It enables continual alignment of the fovea with objects in the scene. It permits correction of retinal shifts induced by head perturbations within reflexive, rather than cognitive, timespans. It permits coordinated fixation and smooth pursuit of targets such that target motion blur is reduced. Active foveal perception and attention allows data reduction and high equivalent resolutions in observing a scene. An egocentric spatial perception provides primates with an awareness of the location of visual surfaces in a scene, and their motion.

# Chapter 3

# Synthesising Primate Vision

In this chapter we examine hypothesised and implemented models of vision that are built upon observations of the components of the primate vision system described in the previous chapter. In consideration of such models we motivate and propose components of a synthetic primate vision system.

## 3.1 Introduction

We concern ourselves with refining a multi-purpose visual sensor system for real-world, real-time, task-directed applications, capable of supporting investigations into synthetic primate vision and perception. It is desirable that the vision system is capable of performing a diverse range of tasks. Processing resources are always limited to some extent. The system must be able to intelligently gather data from its environment rapidly enough for it to make the decisions for task-oriented behaviour and to react to novel events. Real environments contain events occurring at many timescales. It is therefore practical to consider real-time as a time period commensurate to the defined task. The real world is an unstructured, possibly cluttered, dynamic environment that extends beyond sensor range.

The success of biological vision justifies the use of primate inspiration in developing a real-world vision system. Having broadly reviewed the functional components of early visual perception in the primate brain (Chapter 2), we highlight important aspects of primate vision from neurobiological and psycho-physical research observations. We also consider existing models of the components of

primate vision. We revisit the main biological relevance of each component and briefly propose the minimal engineering requirements of each. Components must be necessary for basic scene awareness, and be suitable for real-time implementation. The model is tailored such that it does not significantly (preferably not at all) violate observations of biology. Finally, we consider the integration of components into a unified system.

## 3.2 Components of a Synthetic Primate Vision System

We now consider some important basic components of a synthetic vision system, based on the observations of primate vision. Where possible, we consider existing theoretical and synthetic models of such components. Components can be broadly separated into classes summarised by the processing loop shown in Figure 3.1. These classes may not represent the grouping of functions in the brain, they merely group conceptually similar functions for the purpose of presenting a synthetic system. Beyond image acquisition, component classes discussed include: an egocentric reference frame, attention, spatial awareness and foveal fixation.

### 3.2.1 Egocentric Reference Frame

Binocular primate vision combines visual stimulus from two eyes into a unified representation that accounts for convergence in interpreting retinotropic stimulus. When perceiving scene motion, estimates from many neurons need to be integrated into a global motion estimate. First and second-order flow perceptions appear to be fully combined at the level of area MT/V5. A similar global integration exists for depth perception: local disparities and cues are converted to a global, egocentric perception.

Recent experimental results [Neri *et al.* (2004)] indicate that processing in dorsal areas may rely mostly on information about absolute (egocentric) disparities, while ventral areas split neural resources between the two types of stereoscopic information so as to maintain an important representation of relative (retinal) disparity. Dorsal areas (V3A, MT/V5, V7) showed more adaptation to absolute

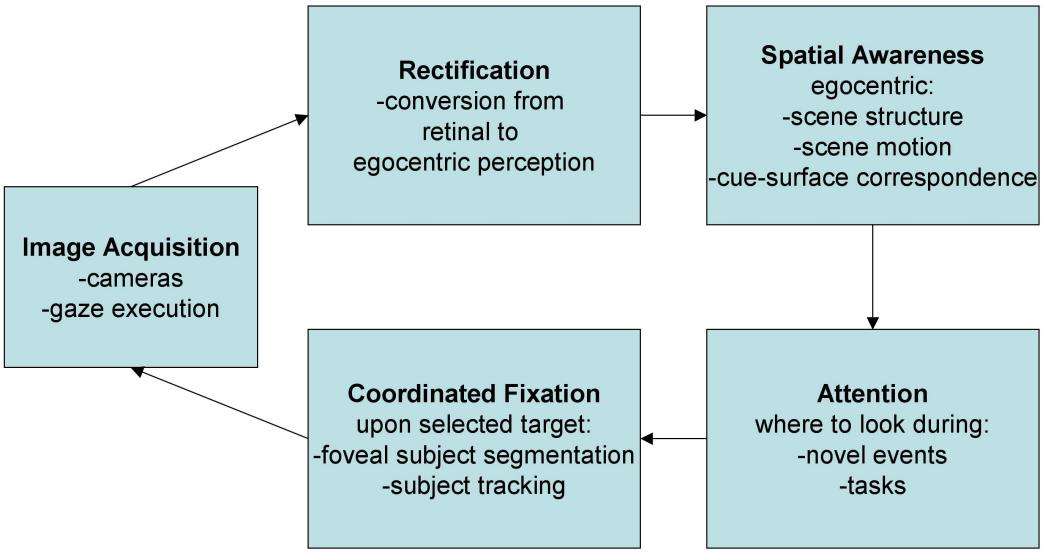



Figure 3.1: Classes of components of synthetic primate vision.

than to relative disparity; ventral areas (hV4, V8/V4) showed an equal adaptation to both; and early visual areas (V1, V2, V3) showed a small response to both absolute and relative disparities. These observations may suggest that the dorsal stream is involved in egocentric perception.

As discussed, monkeys transfer the response of spatially-tuned neurons across eye movements, thus retaining accurate global representations of visual space. The high equivalent ocular resolution of primates is due to this incorporation of visual information over time and gaze shifts. Accordingly, a synthetic system may benefit by transferring imagery from a retinotropic coordinate system to a global and static reference frame so that the relations between left and right, and between successive images, are known despite camera motions such as smooth pursuit and saccade.

Few synthetic vision models deal with active vision convergence by projecting visual stimulus into a static, egocentric (absolute) reference frame. They may instead operate in a retinal (relative) reference frame. To achieve absolute egocentric perception, a synthetic system should be capable of accounting for convergence. We propose a method to achieve a global reference frame based

on online image rectification that integrates images and accounts for perspective changes across time and camera motions. Rectification involves determining and accounting for camera geometry changes so that binocular image pairs can be projected into the static reference frame. As we shall see (Chapter 5), this characterisation also enables the operation of static stereo algorithms with an active stereo platform. Rectification also accounts for lens effects such as barrel distortion.

### 3.2.2 Spatial Awareness

Transferring to an egocentric reference frame enables head-centred egocentric spatial perception. Components of spatial awareness include perception of scene structure and scene motion.

#### 3.2.2.1 Scene Structure

Although numerous cues provide spatial information, primates mainly interpret scene depth from estimates of binocular disparity. Absolute disparities are likely to be interpreted in dorsal areas (V3A, MT/V5, V7). Early visual areas (V1, V2, V3) and ventral areas (hV4, V8, V4) are likely to be involved in processing both absolute and relative disparities. Gaze convergence, focal length and prior familiarity with an object's size can provide information for conversion from relative to absolute depth distances. Gaze convergence stretches extraocular muscles. Kinesthetic sensations from these extraocular muscles have long been known to contribute to absolute depth perception in primates [Zajac (1960)].

As we describe in Chapter 5, there are various ways to calculate disparity from multiple camera views of a scene. These include correlation-based and frequency/phase-based techniques, as well as geometric analysis methods. Many such methods assume fronto-parallel camera geometries to determine relative disparities. Other methods assume static non-parallel geometries and are able to account for non-parallel epipolar geometry to determine absolute disparities. However, the latter methods do not account for camera motions (pan/tilt) and the induced variable epipolar geometry. Moreover, few methods account for variations in epipolar geometries due to dynamic camera convergence at frame rate,

or provide an egocentric 3D perception of scene structure across time and gaze changes.

#### 3.2.2.2 Scene Motion

As with depth perception, there are various ways to calculate optic flow within 2D camera projections of a scene. For our purposes, the main criterion for the selection of a suitable synthetic method is real-time performance. Again, few methods calculate absolute scene flow, that is, account for the image frame effects of deliberate camera motions. Additionally, most methods deal with retinal flow, not absolute scene flow. The method of Kagami [Kagami *et al.* (2000)], though not tailored for active vision or an egocentric perception, seems most promising for real-time 3D scene motion estimation. Kagami uses a static stereo rig. Relative retinal optic flow operations are used to estimate scene horizontal and vertical components of flow while analysis of consecutive depth maps provides the third component of 3D scene flow.

### 3.2.3 Attention

Over the decades, various definitions of attention have emerged to utilise the associated benefits of reduction in complexity. Various general models of attention have been developed for specific purposes:

- Early Selection (Broadbent 1958).
- Attenuator Theory (Treisman 1960).
- Late Selection (Norman 1968, Deutsch & Deutsch 1963).
- Neural Synchrony (Milner 1974).
- Spotlight (Posner 1978).
- Feature Integration Theory (Treisman & Gelade, 1980).
- Object-based (Duncan 1984).
- Zoom Lens (Eriksen & St. James 1986).

- Premotor Theory of Attention (Rizzollati et al. 1987).
- Biased Competition (Duncan & Desimone 1995).
- Feature Similarity Gain (Treue & Martinez-Trujillo 1999).

Noton and Stark developed a "scanpath theory" [Noton & Stark (1971)], proposing that what we see is only remotely related to the patterns of activation in our retinas. This suggestion was based on our permanent illusion of crisp perception over the entire visual field, although only the central two degrees are actually crisp. He suggested a cognitive model where what we expect to see is a basis of perception; the sequence of eye movements is then controlled top-down by our cognitive model of the scene. This theory has been used to restrict analysis of video to a small number of circumscribed regions important for a given task. Rensink proposed a triadic architecture incorporating: 1) pre-attentive processing where low-level visual features are computed in parallel over the entire visual field, up to levels of complexity termed proto-objects; 2) identification of scene gist and structure/layout; 3) attentional vision with detailed object recognition within the fovea (proto-objects are combined for object identification) [Rensink (2000)] .

Such models of attention are plausible and useful, however they have not generally been tested on real-time active vision systems with novel stimulus. They also do not propose how a synthetic egocentric 3D perception of attention is obtained. Rather than assessing high-level models such as those described above, we look at the basic components of vision common to most models of primate attention. We first look at bottom-up attention, then top-down modulation of attention.

#### 3.2.3.1 Bottom-up Attention

Neurons at early stages in the primate visual brain are tuned to simple features like intensity contrast, colour opponency, gradient orientation, motion and stereo disparity. Koch and Ullman proposed that several such feature maps are computed in parallel. They are then combined into a single saliency map from which a selection process sequentially deploys attention to locations in decreasing order of saliency [Koch & Ullman (1985)]. Itti's widely accepted model proposes that

spatial competition for saliency is directly modelled upon non-clasical surround modulation effects [Itti & Koch (1998)] . He uses an iterative spatial competition scheme where, at each iteration, a feature map is convoluted with a 2D difference-of-Gaussian. This is currently accepted as the basis of many models of synthetic attention. After competition, all feature maps are weighted and summed to yield a scalar saliency map. Many models [Tsotsos *et al.* (1995); Cave (1999); Itti (2005)] adopt the winner-take-all method in finally selecting attentional locations. More complex models have been proposed. Milanese used a relaxation process to optimise an energy measure consisting of four contributing factors: 1) inter-feature incoherence favours regions that excite several feature maps; 2) minimising intra-feature incoherence favours grouping of initially spread activity into small numbers of clusters; 3) minimising total activity in each map enforces intra-map spatial competition for saliency; and, 4) maximising the dynamic range of each map ensures process does not converge towards uniform maps at some average value [Milanese *et al.* (1994)] .

Synthetic implementations of attention do not all deal with active cameras or dynamic scenes. It is often assumed that cameras are already pointing appropriately, and that saliency only needs to be determined within a static image frame. One of the functions of attention in primates is to guide gaze for foveal vision and data/search reduction. Non-active attentional implementations may help low-level investigations of attention, but do not synthesise this main function. In incorporating eye motion, a representation of visual space may be required to integrate information across such movements, and to enable a history of past locations to be retained. In monkeys, salient locations are retained across saccades by transferring activity among spatially-tuned neurons within the intraparietal sulcus [Merriam *et al.* (2003)]. A short-term inhibitory effect prevents previously attended stimuli from being immediately re-attended. Few synthetic attention systems address this type dynamic prioritising of attention with moving cameras or in dynamic scenes where objects move, in real time. Real-time functionality is important for active attention - if attention is to be deployed towards a moving target, low latency functionality is important.

#### 3.2.3.2 Inhibition of Return

Koch [Koch & Ullman (1985)] implemented synthetic IOR where, after inhibition, a winner-take-all network then shifts attention towards the next most salient location. This process repeats, generating attentional scanpaths over a static image.

#### 3.2.3.3 Top-down Modulation of Attention

The prefrontal cortex implements attentional control by amplifying task-relevant information relative to distracting stimuli [Nieuwenhuis & Yeung (2005); Wolfe (1996)]. Various manifestations of this type of top-down search have been implemented and shown to assist search in static images: Wolfe [Wolfe (1996)] used the Koch and Ullman model [Koch & Ullman (1985)] with feature-based biasing by weighting feature maps in a top-down manner (for example, bias red features when searching for a "red book"). Navalpakkam proposed a method to optimally set relative feature weights for this type of search [Navalpakkam & Itti (2004)].

A "feature gate" model has also been proposed where a neural network implementation is adopted to determine which cue weightings are relevant for a search task, and modulates bottom-up cue extraction mechanisms accordingly [Cave (1999)]. Similarly, Rao proposed that saliency can be computed from the Euclidean distance between target feature vector and feature vectors extracted at all locations in visual input [Rao *et al.* (1997)]. Torralba extended upon this, developing a Bayesian framework with coarse global analysis where gist gives guidance cues (for example, when in a road-scene, attention may be preferentially deployed to the road and cars) [Torralba (2005)]. Torralba a holistic representation of a scene based on spatial envelope properties (for example, openness, naturalness, etc) that represents the scene as a single identity, bypassing analysis of component objects [Torralba (2005)]. The scene gist is formalised as a vector of the contributing features. Torralba then used learning to find the associations between scene context and categories of objects such as their typical locations, sizes, scales, etc.

Knowledge of the gist of a scene can be used as a contextual guide in search/attention schema to compliment target detection. Extending on Oliva's analysis of the gist

of a scene [Oliva (2005)], Siagian and Itti propose that by sampling a full image, a vector that contains a summary of cue responses for a particular scene can be created [Siagian & Itti (2007)]. Such vectors can be used to identify the type of scene that is being viewed, and recording such vectors can be used in re-localisation.

Top-down bias may also be preempted for regions of the scene not currently in view, but whose position relative to the current fixation point is known. It may accordingly predict the spatial location of cue responses. For example, scanpath theory proposes that attention is guided in a top-down manner based on an internal model of the scene [Noton & Stark (1971)]. Rybak's model proposed is related - scanpaths are learned and then executed for each object to be recognised [Rybak *et al.* (1998)]. This method reduces the emphasis on bottom-up attention and may be difficult to apply to dynamic environments or flexible or moving subjects.

These top-down methods may not be specifically required for scene awareness, but the general ability to modulate bottom-up attention (for whatever reason or by whatever top-down or task-specific process) is seen as a useful feature for a synthetic primate vision system.

#### 3.2.3.4 Covert and Overt Attention

There are three generally accepted models of the interaction between covert and overt attention:

- Independence Model: covert and overt attention are independent and co-occur because they are driven by the same visual input [Klein (1980)].
- Sequential Attention Model: eye movements are necessarily preceded by covert attentional fixations [Henderson (1992)].
- Pre-motor Theory of Attention: covert attention is the result of activity of the motor system that prepares eye saccades - attention is a by-product of the motor system [Rizzolatti *et al.* (1987)].

In each of these models, covert attention involves consideration of factors not directly associated with the current target at fixation. It involves consideration of regions towards or beyond the periphery, whether real, expected or hypothetical.

### 3.2.4 Foveal Fixation

As discussed previously, attention can be seen as "stagelight" that successively illuminates different scene regions as they are considered interesting [Weichselgartner & Sperling (1987)]. Such feedback is believed to be essential for binding the different visual attributes of an object, such as colour or form, into a unitary precept [Trieisman & Gelade (1980); Reynolds *et al.* (2000)]. Feature Integration Theory proposes that only simple visual features are computed in a massively parallel manner over the entire visual field [Trieisman & Gelade (1980)]. Attention binds such early features into a unified object representation. Triesman then suggests that the selected bound object representation is the only part of the visual world that passes through the attentional bottleneck. We do not necessarily need to bind such precepts for the purposes of target identification. We would primarily like to perform foveal figure-background subject segmentation and coordinated fixation, leaving processes such as identification/representation to higher-level processes.

Monkeys exhibit vigorous responses elicited by small laboratory stimuli in isolation, compared to sparse neuronal activity when viewing broad scenes [Vinje & Gallant (2000)]. In humans, long range excitatory connections in V1 appear to enhance responses of orientation selective neurons when stimuli extend to form a contour [Gilbert *et al.* (2000)]. One cue useful in rapidly extracting the boundary of an attended object is *zero disparity*: an attended object appears at near identical positions in left and right retinas, whereas the rest of the scene usually does not. The attended object appears once and crisp in our fused cyclopean view. During stereo fixation, the foveas are aligned over the target in a truly coordinated manner, requiring accurate vergence control. Accordingly, a synthetic system may benefit from a response to the contours of the object upon which fixation occurs. In this manner, foreground objects can be extracted for higher-level consideration such as identification. Of course, accurate segmentation also

assists target tracking.

## 3.3 The Proposed Model

We propose a model for reactive visual analysis of dynamic scenes in terms of a system specification. We specify a minimal set of system features for basic scene perception, based upon biological observations. It is desirable that system operation does not contradict known properties of the primate vision system. Where possible, we use observations of biology to specify methods to deal with active cameras, to define features of fixation control, and to choose a relevant set of early visual cues.

### 3.3.1 System Components

We now list important basic components of a primate-inspired synthetic vision system.

#### 3.3.1.1 Image Aquisition

- cameras
- active platform
- camera parameter control
- simultaneous, accurate head axis motion control

#### 3.3.1.2 Rectification

- camera barrel rectification
- epipolar rectification for use of static stereo algorithms on active platform
- accounting for convergence
- converting from retinal to head-centred egocentric reference frame

- distribution of rectified camera data (colour opponents, intensity) and rectification parameters

#### 3.3.1.3 Attention

Where to look:

- during novel visual stimulus
- during tasks

Permit online top- down modulation from higher processes:

- for search
- for tasks

Cues:

- colour opponency
- intensity centre-surround
- depth
- optic flow
- orientations
- TTC

Coping with the real world and active cameras:

- egocentric IOR for retaining suppression over gaze shifts
- dynamic IOR for propagating suppression as objects move
- TSB

#### 3.3.1.4 Spatial Awareness

An egocentric 3D perception for determining:

- scene structure
- scene motion
- cue-surface correspondences

#### 3.3.1.5 Coordinated Foveal Fixation

Once attention directs foveas to interesting location:

- extract target
- track target

#### 3.3.1.6 Flexibility

Incorporating rapid I/O and real-time performance for:

- environmental data distribution
- task specific tuning (search/track/mapping/etc)
- top-down modulation of attention

### 3.3.2 Discussion

In any implementation, processing resources are limited. For this reason component implementations need to balance the trade-off between processing time and accuracy. Similarly, it may not always be possible to implement algorithms exactly as hypothesised in primate vision. We have defined components based on the minimum requirements required for primate-like awareness.

We leave camera parameter control on automatic settings, handled by the cameras themselves, where possible. For example, camera hue, saturation, contrast, brightness, etc are controlled automatically by the cameras. We therefore

adopt algorithms, where possible, that can cope with these variations. For example, we employ the use of difference-of-Gaussian representations where images are intensity-normalised.

We can use existing research for many components of the system, such as the implementation of cues like disparity and optic flow. The parts of the system for which further research is required are:

- Egocentric/absolute perceptions developed from active cameras.
- Coordinated fixation and target segmentation.
- Active-dynamic attention.

These areas are discussed further in subsequent chapters.

Of course, having proposed such a system based on primate vision, we want to evaluate it. We want to assess its capabilities in terms of both tangible metrics and primate-like behaviours. There exists little psycho-physical data for evaluating unconstrained human attention for such a comparision. It is unlikely that when observing unbounded 3D scenes humans exhibit behaviours identical to when they observe static pictures or 2D videos, as utilised in most attentional experimentation to date. Depth and covert (peripheral) object saliency suppression, for example, are likely to affect attention in a manner that the use of static images and image frames cannot demonstrate.

## 3.4 Aspects of Implementation

We now consider the underlying requirements necessary to support such processing capabilities. We consider the nature of the processing structure and engineering components, including hardware and software.

### 3.4.1 An Expandable Processing Network

Real-time vision processing requires significant processing power, such as that available via a network or cluster. Such a processing network would require:

- Support for multiple processors with vector processing capabilities.
- Efficient data distribution.
- Simultaneous serial and parallel processing.
- Expandability.

### 3.4.2 Framework Components

The synthetic vision architecture comprises of hardware and software components.

#### 3.4.2.1 Hardware

We use readily available off-the-shelf components. Hardware includes:

- Video cameras - 2 x Sony FCB EXT-37 series analog.
- Video capture cards - 2 x Brooktree type-29 framegrabbers.
- An active vision head capable of moving cameras at high speed - CeDAR (see next chapter).
- Motion axis amplifiers.
- Motion control card for head axis control - Servo-To-Go inc. (STG).
- Processing computers - dual CPU 3.2GHz.
- Network components such as 10/100/1000Mbit ethernet cards and hubs.

#### 3.4.2.2 Software

The system is capable of supporting processing on multiple computer nodes simultaneously with cross-communication of data. We use CORBA (client object request broker) to distribute data over the network and to initiate remote procedure calls (RPC). For example, motion control and video capture drivers are embedded within CORBA wrappers that allow control of head motion and

camera parameters by RPCs. CORBA enables serial and parallel processing on computers in the network. It facilitates expansion of the processing network by permitting the addition of extra processing nodes without affecting existing network functionality.

Processing algorithms onboard each computer are multi-threaded, allowing further simultaneous serial and parallel processing within a node. Built on a Linux environment, programmes are written in C++ and make use of OpenGL, MMX and SSE hardware accelerations. We use Intel's *Intel Performance Primitives* (IPP) library to take advantage of the MMX/SSE vector processing instruction set, and OpenGL for graphics card display acceleration. Beyond these libraries, functions are written in-house for system performance optimisation and fine-grained control.

### 3.4.3 Network Structure

We adopt a CORBA client-server architecture to allow concurrent serial and parallel functional network processing. At the lowest level, a video server controls image capture, handles remote requests for images, and distributes images to other computers (nodes) for subsequent processing. Similarly, a motion control server handles head motion requests and distributes head status parameters to nodes. To minimise network bandwidth, to cope with the processing load of each frame, and to prevent repetition of computations, nodes in the structure are configured simultaneously as clients of processes preceding them in the functional serial pathway, and as servers to nodes following (dependent upon) them. Multiple such serial processing pathways can also exist in parallel. Each node corresponds to a physically separate PC and all are dual CPU hyper-threaded machines, with two physical CPUs amounting to four virtual processors. Trade-offs exist between splitting tasks into subtasks, passing subtasks to additional nodes and minimising network traffic. The best performing solution involves grouping of serialised tasks on each server, and that as many operations are done on the image data on the same server as possible, so there is minimal CPU idle time and minimal network traffic between servers. The serial nature of cue computations means

there is often no gain possible in distributing the task – in fact further network transfer of data between servers would slow performance significantly.

### 3.4.4 Data Transfer

Data transfer between nodes can be either *push* or *pull*. That is, a node can *expect* or *request* data from preceding nodes (respectively). The type of data transfer selected depends upon the function of the node. For example, if node processing is heavy, it may not operate at the same rate as the preceding node can distribute images. In this case, depending on the function, we could either stop processing on the current frame and move into the next available frame as it arrives (push), or complete processing on the current frame, perhaps dropping a few frames available from the previous node, and then request the next image when ready (pull). Alternatively, if it is possible to estimate the processing time remaining on a frame, and the network data transfer latency is known, the node may request (pull) the next frame before it has finished processing the current frame, such that it arrives in a timely fashion (as if pushed) when processing on the current frame is finished. In this manner, CPU idle time can be minimised, and no CPU time is expended managing a buffer.

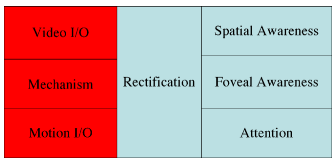
Ideally, images are time-stamped upon capture so that if a later processing node requests images from two preceding nodes that take different processing times, they can be matched. Alternatively, algorithms may be designed such that precise synchrony is not essential; instead discrepancies due to a few frames of delay can be absorbed.

## 3.5 Summary

We have looked at existing models of primate vision and justified components using biological inspiration. We have presented a list of desirable basic components of a synthetic primate vision system for basic scene awareness. We have considered aspects of implementation of components, and of a framework capable of supporting distributed image processing.

# Chapter 4

# Active Vision Platform



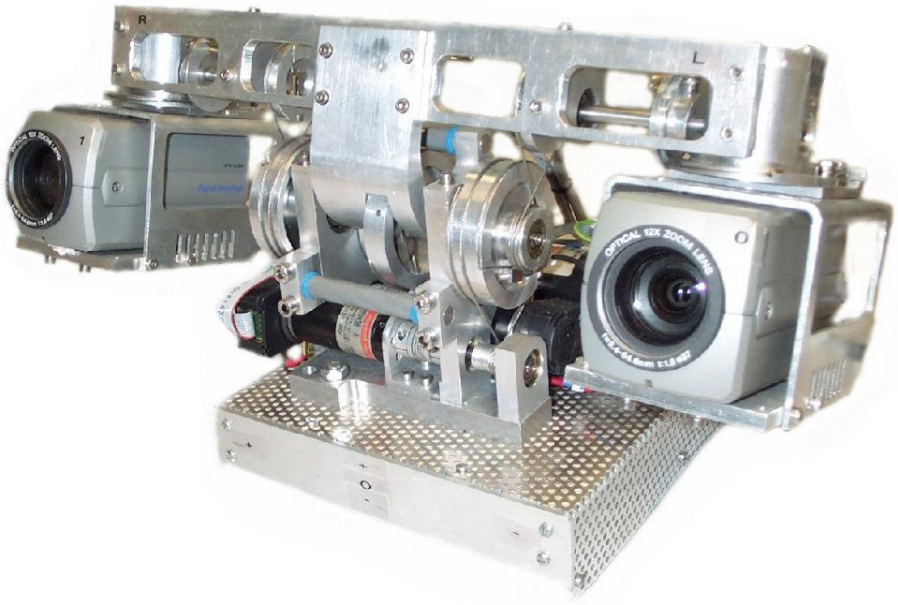



Figure 4.1: CeDAR (Cable-Drive Active-vision Robot).

In this chapter we describe the biologically-inspired active vision mechanism with which this research is conducted. We motivate and present aspects of its design and control that make it particularly suited to investigating synthetic primate vision. We demonstrate its mechanical performance.

## 4.1 Introduction

The concept of controlled camera movements to facilitate vision undoubtedly originated from observations of the biological world. As discussed in Chapter 2, primates benefit from active vision in several ways. Active foveal perception allows data reduction and high equivalent resolutions in observing a scene. It enables continual foveal alignment of objects in the scene. It permits correction of retinal shifts induced by head perturbations within reflexive, rather than cognitive, timespans. It permits smooth pursuit of targets such that target motion blur is reduced. The same benefits are possible for a synthetic vision system.

An active mechanism has several benefits over the use of high resolution wide-angle static cameras that may or may not incorporate moveable *virtual*, pseudo-foveal processing areas (for data reduction). Narrow angle video cameras are common, provide images of resolutions suitable for sustainable processing at frame rate, exhibit increasingly small form factors, and are financially economical. A high resolution camera with or without variable virtual fovea processing would need a high frame rate to remove motion blur of moving objects. Objects tracked in a virtual fovea will exhibit motion blur that can be largely removed by *physical* foveal tracking (for example, a moving car photographed at the same shutter speed will exhibit less motion blur if the camera is panned with the car than if it is held stationary). Physical active vision can therefore provide more resolute information about selected targets in the deployment of attention if the subject (or the active vision platform itself) is moving. High resolution, high frame rate cameras produce large amounts of data requiring accordingly large computational resources for real-time processing, and may require high volumes of processor-intensive memory writes, even if a *virtual* cropped fovea is selected. Lack of physical gaze direction would require a wide-angle lens to achieve a large range of vision. Wide-angle lenses introduce lens distortions that must be accounted for. This can be an expensive operation for high resolution cameras, and errors may compromise image resolution, especially towards the image edges where most distortion exists. A virtual fovea in a static camera would exhibit an asymmetric periphery which may impede primate-like perception. High frame rate mechanical image stabilisation, such as the vestibulo-ocular reflex, is also not

possible with static mechanisms. Biology achieves high equivalent resolutions, large and redirectable peripheries, high tracking acuity and retinal stability using *physical* foveal active vision.

These benefits point strongly towards the use of an active mechanism. Scientists have endeavored to mimic the abilities of biology when designing synthetic active vision systems. Synthetic binocular active vision systems are usually composed of a pair of cameras that, mimicking the coordination of successful biological systems, share a common tilt axis and rotate symmetrically about separate vertical axes for the control of vergence, or independently for both vergence and version. Other systems go further, mimicking neck rotation and/or the small torsional rotations of the human eye around its optical axis enabled by the superior oblique muscle.

We proceed by presenting relevant existing active vision platforms. We then present background information about the mechanism with which this research is conducted. The mechanism was modelled and manufactured in house by Harley Troung *et al.* prior to commencement of this research [Troung (1998)]. We summarise the previous work, including the platform's mechanical design, kinematics and mechanical performance evaluation. Previous work also involved the implementation of biologically-inspired low-level motion control [Sutherland *et al.* (2000)]. Standard methods to stabilise images obtained from an active vision platform have been implemented and are presented as background knowledge. Because of the system's reliance upon the active vision mechanism, it is important that the capabilities of the active vision platform are understood. It is also important that its mechanical capabilities reflect those of the primate vision system from which the synthetic vision system is inspired.

After presenting this background work we discuss integration of the platform into a network-based flexible vision processing structure. We describe the regime used to allow distributed network access to images captured by the platform's cameras, and to the platform's mechanical status and motion control.

### 4.1.1 Related Work

The human eye achieves extraordinary performance through its low weight and low inertia muscle actuation. Accordingly, existing synthetic vision systems have been built that endeavour to mimic the properties of biological vision systems.

A brief overview of recent active vision devices reveals a trend towards smaller, more agile systems. In the past the goals were to experiment with different configurations using large systems with many degrees of freedom like the KTH active head [Pahlavan & Eklundh (1992)] with its 13 degrees of freedom and Yorick 11-14 [Sharkey *et al.* (1997)] with a 55cm baseline[1] and reconfigurable joints. Although useful for experimentation, these systems were cumbersome, affecting agility and rendering them difficult to configure for mobility. Smaller active heads such as the palm-sized Yorick 5-5C [Sharkey *et al.* (1997)] and ESCHeR [Kuniyoshi *et al.* (1995)] (with an 18cm baseline), were developed as light-weight systems suitable for mobile robot and teleprescence applications. All three versions of Yorick as well as ESCHeR use harmonic or gear drive technology. A limitation of the technology is an unavoidably large speed-reduction ratio that limits the output speed to less than 100rpm. In many instances the size of the motors and cameras limits the compactness of the active head and the motors themselves add to the inertia of the moving components. An exception is the Agile Eye [Gosselin *et al.* (1996)] where no motor carries the mass of any other.

Active vision heads used in humanoid research also tend to incorporate a smaller baseline, in line with the human vision system. A narrow baseline has the disadvantage of eliciting less binocular disparity for a given camera resolution. Sometimes, due to power and payload requirements, humanoid heads exhibit less agility than vision-specific robots. Notable exceptions include the DB vision head at ATR Japan [Ude *et al.* (2005)], and the SARCOS[2] head that is able to achieve angular accelerations and velocities beyond that of a human.

[1] The *baseline* is the distance between camera optical centres.
[2] Sarcos Research Corporation - http://www.sarcos.com.

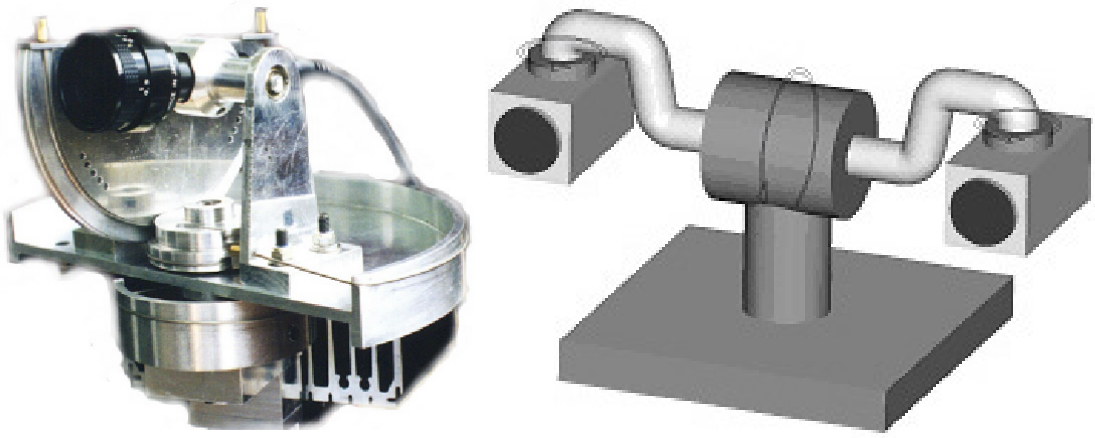


Figure 4.2: CeDAR's development. An early prototype (left) and the Helmholtz configuration of CeDAR (right).

## 4.2 Mechanical Design

The mechanism was designed to incorporate aspects of design from other laboratories and the mechanics of the human visual system [Truong *et al.* (2000)]. Muscles are lightweight, exhibit high accelerations and minimal backlash. Muscle-skeletal structures do not suffer significantly from the common problem encountered in serial active head designs whereby each degree of freedom requires sufficiently powerful actuators to move all previous degrees of freedom, including their actuators. By adopting a parallel architecture and relocating the motors to a fixed base, thereby reducing the inertia of the active components to little more than the mass of the cameras, problems inherent in serial design were alleviated [Brooks *et al.* (1997)]. However, the parallel architecture does have the added complexity of coupling, where the motion of one axis causes the movement of another. This effect is counteracted through a software decoupling function [Truong *et al.* (2000)]. To map between the joint and actuator spaces, the coupling ratio (which defines the amount of vergence compensation required to decouple the verge joint from the tilt joint) is determined.

Backlash-free speed reduction is essential for high-speed performance, so the choice of transmission system for the parallel architecture is important. During high-speed movements such as saccades, where motors are driven at maximum

acceleration, velocity saturation for harmonic-drive gearboxes is of concern. Cable drive, a novel alternative for use with repeated bounded motion, does not induce speed limitations, operates lubricant-free with low friction, exhibits high torque transmission, and is low-cost.

An earlier prototype [Brooks *et al.* (1997)] (Figure 4.2) proved the usefulness of cable drive transmissions and parallel mechanical architectures in a two-degree-of-freedom active 'eye' system. The prototype was fast (able to achieve an angular velocity of $600^o s^{-1}$ for each axis), responsive (angular accelerations of up to $72000^o s^{-2}$) and accurate (to a resolution within $0.01^o$). In 2000, the prototype's architecture was transferred to a stereo Helmholtz configuration [Murray *et al.* (1992)] (Figure 4.2), resulting in the present mechanical design of CeDAR (Figure 4.3). The platform has three mechanical degrees of freedom. The cameras share a common tilt axis, while the independent left and right verge axes enable asymmetric vergence. In this manner, any 3D scene location in front of the platform may be attended by both cameras simultaneously (within the range limits of each axis). This configuration is similar to the human vision system, as human eyes are also able to move independently for asymmetric vergence. Even though the eyes have separate muscles to tilt each eye up and down, they are normally constrained to move together within the same horizontal plane.

An important kinematic property of the design is that the axes intersect at the optical centre of each camera, minimising kinematic translational effects. This property of the stereo camera configuration reduces complexity in stereo algorithms such as depth reconstruction through image disparity calculations, a competency that biological vision systems exhibit [Wilson & Cowan (1972)].

Actuation has been transferred through cable drive circuits that integrate with the parallel architecture. Power ratings for the actuators (70W tilt axis, 20W each verge axis) are such that the unit can operate in mobile robotic situations, where low power consumption is desirable.

A multi-modal systems approach was adopted where the mechanism and its control were developed in parallel, with integration in mind [Brooks *et al.* (1998)]. The system was designed to achieve the mechanical capabilities of existing synthetic vision systems and the abilities of the human vision system, while incorporating reasonably sized payloads (for example, two 700g cameras). The

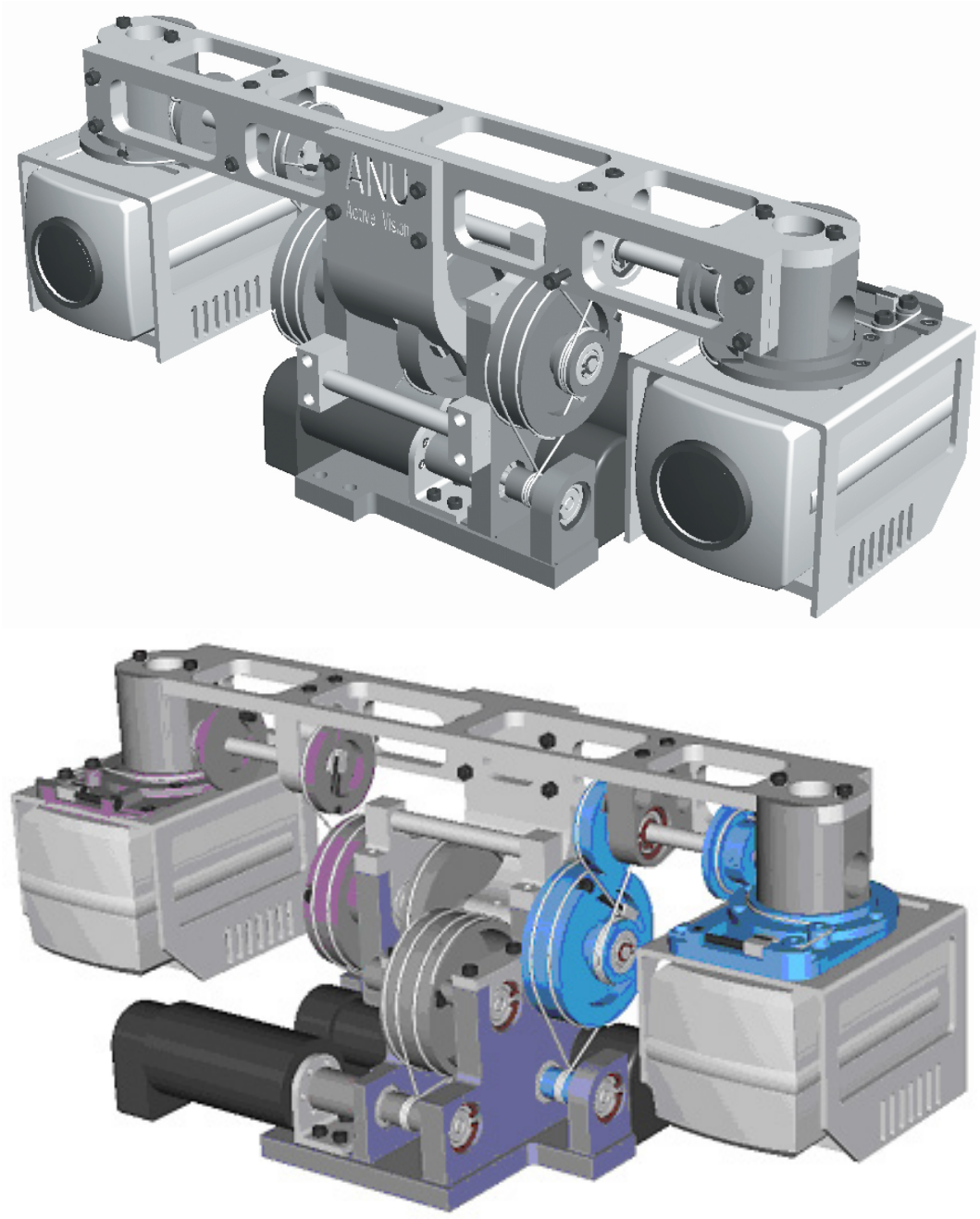



Figure 4.3: CAD model of CeDAR [Troung (1998)]. The rear view (bottom) shows CeDAR's parallel architecture and cable drive (bottom).

mechanism and control modules have been conceived with the purpose of having a high-level of mechanical performance to allow rapid motion and short reaction times, as well as being reconfigurable for application to many situations with minimal modification. The control and vision processing modules have been developed for use with standard digital visual hardware operating at a 30Hz frame rate. Low cost, ease of reproduction and configurability for mobility were also significant factors in the evolution of the design.

## 4.3 Kinematics

We can relate the position of the joints, as measured by the encoders, to the position of the active head in a real world coordinate system. The cameras share a common tilt plane, and have independent verge axes to enable both vergence and version. The coordinate frame for each joint is shown in Figure 4.4. Each joint has been placed in its respective home position, where the tilt axis frame coincides with the defined world coordinate system $[x, y, z]$. The $[X_{VL}, Y_{VL}, Z_{VL}]$ and $[X_{VR}, Y_{VR}, Z_{VR}]$ coordinate frames are fixed to the left and right cameras, with the Z-direction pointing along their optical axes and the X-Y planes parallel to the camera image planes. From the definitions in Figure 4.5, we can relate axis angles to real-world Cartesian coordinates as follows:

$$x = \frac{\tan\theta_L}{y} \tag{4.1}$$

$$y = \frac{l\tan\theta_R}{\tan\theta_L + \tan\theta_R} \tag{4.2}$$

$$z = x\tan\theta_T \tag{4.3}$$

where $l$ is the baseline length separating the cameras (30cm).

## 4.4 Mechanical Performance

The maximum velocity of each joint actually exceeds the maximum velocity achievable by the human eye. The CeDAR can exceed $600°s^{-1}$, whereas the

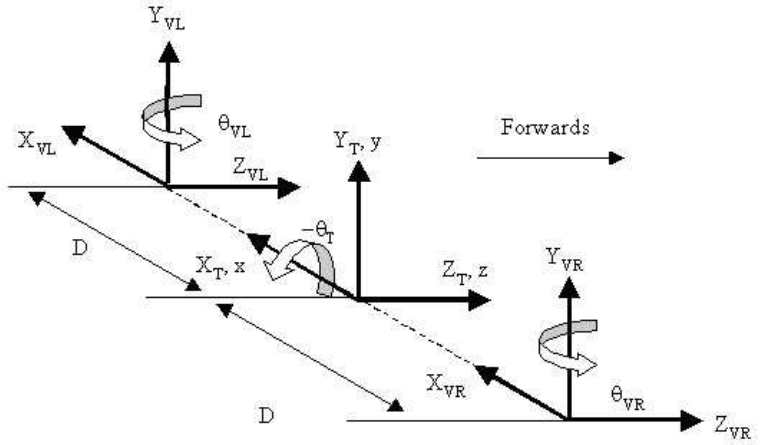



Figure 4.4: Joint kinematics.

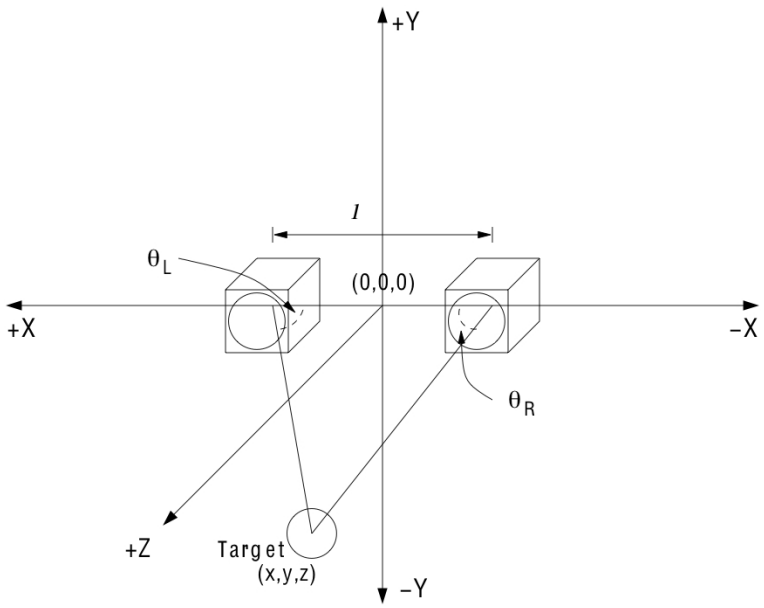



Figure 4.5: Conversion from Cartesian coordinates to axis angles.

human eye is limited to $400°s^{-1}$ [Rodieck (1998),Truong *et al.* (2000)]. The maximum acceleration, which is greater than 18 $000°s^{-2}$ is also comparable to the human eye.

The angular resolution and repeatability of each joint was measured to be 0.01°, which enables highly accurate control of both position and speed during any necessary motions including saccade, smooth pursuit and fixation.

The performance figures traditionally reported for active vision mechanisms consist of maximum angular velocity, maximum angular acceleration, angular resolution and axis range. While the latter two are highly relevant, we consider the others to be not especially useful in that they do not detail any form of specific task competency. Additional specifications for an active vision system which not only involve the speed and acceleration of the axes, but also express the usage intention of the system in the form of a functional requirement, are given in Table 4.1. The tabled values need to be achieved in order to satisfy constraints related to desired motion abilities. CeDAR's maximum allowable full-speed saccade time was required to be 0.18s to enable three $90^o$ gaze shift saccades, with an allowance for each to be preceded by four target location video frames and succeeded by one stabilisation video frame per second. Just over five frames are captured during the saccade itself.

The minimum allowable full speed stop-to-stop angular change within one video frame was required to be $15^o$, which equates to the ability to track an object moving past the cameras at up to $4ms^{-1}$ at a distance of 1m. The angular resolution has been selected with the aim of allowing the platform to perform meaningfully small camera movements and to allow single pixel selection.

The required maximum range, payload and baseline specifications were based on the desire to use commonly available motorised zoom cameras. However the unit is reconfigurable to incorporate many off-the-shelf cameras. With the potential to incorporate smaller cameras, minor improvements in mechanical performance are likely.

The saccade rate and pointing accuracy were chosen to reflect biological system performance. It is desirable that the system can rapidly attend novel or salient visual events.

Table 4.1: Performance specifications and test results.

| | Test | | Specification | |
|---|---|---|---|---|
| Specification | Tilt | Vergence | Tilt | Vergence |
| Max velocity | $600^o s^{-1}$ | $800^o s^{-1}$ | $600^o s^{-1}$ | $600^o s^{-1}$ |
| Max acceleration | $18,000^o s^{-2}$ | $20,000^o s^{-2}$ | $10,000^o s^{-2}$ | $10,000^o s^{-2}$ |
| Saccade rate | $5 s^{-1}$ | $6 s^{-1}$ | $5 s^{-1}$ | $5 s^{-1}$ |
| Ang repeatability | $0.01^o$ | $0.01^o$ | $0.01^o$ | $0.01^o$ |
| Ang resolution | $0.01^o$ | $0.01^o$ | $0.01^o$ | $0.01^o$ |
| Max range | $90^o$ | $90^o$ | $90^o$ | $90^o$ |
| Payload | Two 700g cameras | | | |
| Baseline | 30cm | | | |

Speed performance was determined by driving the joints to their maximum range, speed and acceleration in a cyclical fashion (repeated saccades) [Troung (1998)]. The command positions and actual positions of the joints were logged at millisecond intervals. The position data was then differentiated using a three-point rule and filtered using a seven-point moving average to obtain velocity and acceleration profiles.

A series of accuracy tests were also conducted using laser pointers mounted on the robot head [Troung (1998)]. Repeatability, the ability to return to an absolute position after a series of complex movements, was demonstrated by moving the joints to an arbitrary position, relocating to another location and then returning to the original point. In systems that suffer from backlash, friction or poor compliance, the return point differs from the original. Angular resolution, the smallest angle that can be actuated was measured by moving the joints a minimal increment. Coordinated motion, the joints' ability to move in unison, was demonstrated by verging both laser pointers to the same location on a wall then commanding the system to follow a predetermined trajectory. Coordination was evaluated according to how closely the lasers were converged throughout the motion. Table 4.1 lists results of the accuracy tests along with results of the speed tests and the design specifications.

## 4.5 Motion Control

### 4.5.1 I/O

Motion commands need to be transformed into voltage and current signals as inputs to the motors in order to move the joints. A *Servo-To-Go Inc* (STG) motion controller card is used to transform the motion commands into an analogue signal output to the *pulse-width modulated* (PWM) amplifiers using a *proportional-integral-derivative* (PID) control algorithm. The low-level PID controller compares the actual position of each joint to the command position at each time-step, and the resulting error is used for adjustment.

The PWM amplifiers amplify the analogue signals from the STG card to drive the three motors at the base of the active head. The actual positions of the joints are determined by the optical encoders, which are attached to each motor. The particular encoders used on CeDAR have a high-level of precision, guaranteed to $0.01^o$. The position measured by the optical encoders is fed back to the motion control card to be used by the PID controller. The STG card in turn feeds the positions back to the computer to be used by the high-level software controller.

The range of motion of each joint is discretised into a defined number of positions in accordance with encoder resolution. The (tiny) detectable constant distance between adjacent discrete encoder positions is known as a *click*. The STG card reports all information back to the software control level in terms of clicks and the current click position. Joint positional and velocity information can be directly converted to degrees and degrees per second for intuitively easier understanding.

### 4.5.2 Trapezoidal Profile Motion

As observed in nature, gaze control can be broken down into two basic tasks: saccade and smooth pursuit. CeDAR's control routines are an extension of work undertaken by [Murray *et al.* (1992)] on *trapezoidal profile motion* (TPM). In particular, the approach allows for the implementation of a single algorithm for both saccade and smooth pursuit, enhancing the simplicity and compactness of

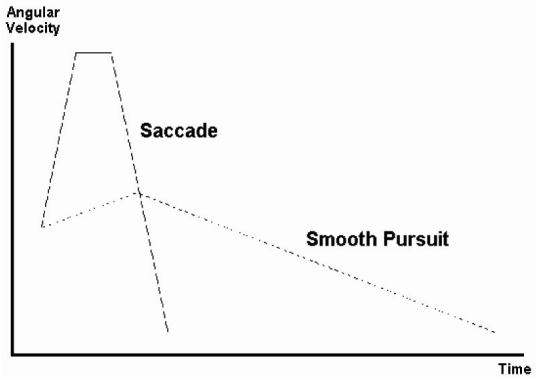



Figure 4.6: Trapezoidal profile motion (TPM) velocity control profiles.

the controller design. We describe the TPM control regime as implemented by Sutherland *et al.* [Sutherland *et al.* (2000)].

The essence of the TPM problem is to detect and transfer gaze to the desired target point a distance from the image centre either in the shortest time possible (saccade) or as smoothly as possible (smooth pursuit). Both the joints' and target's starting velocity are potentially non-zero and disparate. As the name of the controller suggests, the trajectory calculated is completely characterised by a trapezoidal-shaped velocity profile. The profile consists of three main stages, and the starting point is taken to be the current velocity of the joint. Specifically, we cause each visual axis to accelerate constantly to a calculated ceiling velocity[1], coast at this velocity for a given period, then decelerate at the same constant rate as the acceleration until the target velocity is reached (Figure 4.6). Mathematically, it is a four-dimensional problem per axis where the acceleration $a$, ceiling velocity $v$, move time $T$ and total distance travelled $x$ are unknown. The initial joint velocity $v_1$, target velocity $v_2$ and the target's initial distance from the image centre $x_0$ are the givens.

If the acceleration $a$ is assumed to be constant, the time taken by the head to accelerate from its initial velocity to the ceiling velocity is

$$T_a = \frac{sv - v_1}{sa}, \tag{4.4}$$

[1]The maximum absolute velocity of the TPM trajectory

Where $s$ is positive for $v_1 < v$ and negative for $v_1 > v$. Similarly, the time to decelerate to target velocity is

$$T_d = \frac{sv - v_2}{sa}. \tag{4.5}$$

Note that acceleration and deceleration rates are equal. If $T_c$ is the time spent coasting at the ceiling velocity, the total time for TPM is

$$T = T_a + T_c + T_d. \tag{4.6}$$

The distance travelled by the head in time T is

$$x = \frac{sv + v_1}{2} T_a + svT_c + \frac{sv + v_2}{2} T_d, \tag{4.7}$$

but can also be considered as the sum of the initial distance of the target from the foveal centre $x_0$ and the distance travelled by the target during the move

$$x = x_0 + Tv_2. \tag{4.8}$$

These general equations can be used to develop the case for saccade and smooth pursuit.

#### 4.5.2.1 Saccade

Saccade involves changing the head's current position and velocity state to that of the target, as inferred by its previous states, in the shortest time possible. Motion smoothness is not a concern and hence acceleration is set to its maximum possible magnitude. Two cases can arise:

1. The ceiling velocity required for the action is less than the maximum allowed velocity and hence no time is spent coasting.

2. The theoretical ceiling velocity required for the action is greater than the maximum allowed velocity and hence some time must be spent coasting.

It is useful to assume $T_c$ is initially zero so that $T$ can be deduced from Equation 4.4 to 4.8 with $sv$ calculated as

$$sv = v_2 \pm \frac{1}{2}\sqrt{4sx_0a - 2(v_1^2 + v_2^2 - 4v_1v_2}, \tag{4.9}$$

where the smaller value is taken for $v_2 > v_1$ and vice-versa. If $v$ exceeds the maximum allowed velocity, $a$ and $v$ are replaced by their maxima. Then

$$T_c = \frac{kT_a - x_0}{v_2 - sv}, \tag{4.10}$$

where

$$k = sv - \frac{v_1 + v_2}{2} \tag{4.11}$$

is calculated to deduce $T$. Equation 4.9 also defines the value of $s$ so that the operand of the radical is greater than or equal to zero

$$s = \begin{cases} 1 \text{ for } (v_1 - v_2)^2 + 4x_0a \geq 0 \ , \\ -1 \text{ otherwise.} \end{cases} \tag{4.12}$$

An example of a saccade trapezoidal velocity profile that reaches maximum achievable velocity is depicted in Figure 4.6 (marked "saccade").

#### 4.5.2.2 Smooth Pursuit

Smooth pursuit involves moving from one position and velocity state to the next in a given amount of time with optimal smoothness. To achieve this, the acceleration in moving to and from the ceiling velocity must be as small as possible. Again both the coasting and non-coasting cases are relevant. With the assumption that the coasting velocity is initially zero, Equations 4.4 – 4.8 yield

$$v = \frac{x}{T} \pm \frac{1}{2T}\sqrt{4x^2 - 4Tx(v_1 + v_2) + 2T^2(v_1^2 + v_2^2)} \tag{4.13}$$

If these values are in excess of the maximum allowable velocity of the head, the time constraint is unrealisable. In this instance, a saccade is initiated.

The coast at constant velocity is not always necessary if the ceiling velocity is less than or equal to the maximum velocity achievable. An example is the velocity profile of a smooth pursuit motion to acquire a target moving at a non-zero velocity is displayed in Figure 4.6 (marked "smooth pursuit").

As discussed, the parallel mechanism architecture prevalent in the design of the CeDAR results in a coupling effect between the tilt axis and the vergence axes. When motion occurs in the tilt joint, the left and right vergence axes also move even if their respective motors are stationary. Therefore a function was written to

compensate each parameter calculated by the TPM algorithm, before any motion commands were sent to the PID controller. This was achieved by implementing the translation from joint to motor space. Coupling ratios have been determined and their complimenting decoupling ratios are used to counteract this coupling effect.

### 4.5.3 Gaze Stabilisation

Gaze stabilisation is an important aspect of platform control. The primate vision system incorporates both image-based and gyro-based vision stabilisation. We look briefly at how stabilisation is achieved in the primate vision system, and how these features are commonly implemented to help stabilise synthetic vision platforms. Research in subsequent chapters is performed under stable laboratory conditions where stabilisation is not required. However, stability is an important component of primate vision, and would be highly relevant for gaze stabilisation in mobile applications for a synthetic vision system. Both gyro-based and image-based gaze stabilisation have been previously implemented on various active vision systems [Panerai *et al.* (2000)]. To demonstrate the primate-like capabilities of the CeDAR platform we now discuss biological evidence for these reflexes, and present a simple implementation largely similar to previous implementations.

#### 4.5.3.1 Gyro-based Stabilisation

The vestibulo-ocular reflex (VOR) is a reflex eye movement that stabilises images on the retina during head movement by producing an eye movement in the direction opposite to head movement that preserver the location of the image on the retina. For example, when the head moves to the right, the eyes move to the left, and vice versa. Since slight head movements are present all the time, the VOR is important for stabilising vision. The primate vision system struggles to capture visual information if the projected image slips across the retina at more than a few degrees per second [Westheimer and McKee, 1954]. For humans to be able to see with acuity while the head is moving relative to the world or a visual target, the vision system must compensate for the motion of the head by turning the eyes to stabilise the image in the retina. Patients whose VOR is impaired find

it difficult to read print because they cannot stabilise the eyes during small head tremors [Baloh *et al.* (1981)]. The VOR reflex does not depend on visual input and works even in total darkness or when the eyes are closed [Rodieck (1998)].

The "gain" of the VOR is defined as the change in the eye angle divided by the change in the head angle during the head turn. If the gain of the VOR is wrong, for example, if eye muscles are weak, or if a person puts on a new pair of eyeglasses - then head movements result in image motion on the retina, resulting in blurred vision. Under such conditions, motor learning adjusts the gain of the VOR to produce more accurate eye motion. This is referred to as VOR adaptation [Gluck *et al.* (1990)].

The main neural circuit for the VOR is simple [Kandel *et al.* (2000)]: vestibular nuclei in the brainstem receive signals related to head movement from the scarpa ganglions located in CN VIII and the vestibular nerve. From the vestibular nuclei, excitatory fibres cross to the contralateral CN VI nerve nucleus where they split into two additional pathways. One projects directly to the lateral rectus of the eye. The other projects to the oculomotor nuclei, which contains motorneurons that drive eye muscle activity, specifically activating the medial rectus muscles of the eye. The cerebellum is essential for motor learning to correct the VOR in order to ensure accurate eye movements [Gluck *et al.* (1990)].

A seven degree-of-freedom SARCOS stereo head was equipped with high-speed, high-precision four-axis independent pan and independent tilt gaze control. A six-axis gyro/accelerometer unit was rigidly attached to the head at the centre of its rotation, fixed with respect to the baseline of the stereo camera pair. This meant that external perturbations to the robot head could be detected at a much higher frequency than the camera frame rate, such that a compensation signal could be injected into the gaze control loop that minimised the effect of such perturbations on gaze direction, significantly compensating for such perturbations so that their effect on visual tasks is reduced. Translational perturbations that induce forwards/backwards motions of the head require a corrective motion that depends on the distance to the attended object, unless the object is directly in front of the head and located near infinity. Where objects are located in the near foreground, forwards/backwards translations induce scale variations that cannot be corrected using only pan/tilt motions. Horizontal and vertical translational

perturbations require corrective measurements that depend on the distance to the attended object. However, forwards/backwards, lateral, and vertical translational perturbations elicit only minor image frame shifts in comparison to rotational perturbations. This, and the added complexity associated with depth-dependent corrections, means that we concentrate on detecting and correcting for rotational perturbations.

The head is not capable of rotating the cameras about their optical axes. Primate exhibit minimal rotational abilities in this regard. Therefore only perturbation rotations that can be corrected by pan and tilt motions were considered. Gyro data obtained at 1000Hz was integrated to determine the approximate rotational perturbation angles $R_x, R_y$ (angles about reference frame coordinate system axes as defined earlier). The gaze was shifted by $-R_x, -R_y$ to approximately correct for the detected perturbation. The displacement of the camera centres from the gyro unit was small enough ($\sim$3cm in horizontal direction only) such that the kinematic effect on the correctional rotation angle for each eye was considered insignificant. Gyros are prone to drift, but lower rate image acquisition frequency signals, such as a frame-rate target tracking signal, was used to correct any drift and ensure gaze remained on target. Errors present in the frame rate target tracking signal also render both the minor effect of the non-zero displacement of the camera centres from the gyro and the slight image frame effects of translational perturbations even more insignificant.

An experiment was conducted to demonstrate the task-specific performance improvements associated with the use of synthetic VOR. It was shown that VOR dramatically improved visual acuity and tracking quality during the performance of a basic visual task:

- The visual task was to track a coloured object using a previously implemented chrominance-based colour tracker.

- Randomly varying sinusoidal nodding and shaking motions (perturbations) of the head were induced in the three neck axes. The perturbation control loop was separated from the four-axis, frame rate, camera colour tracking control loop.

- The gyros were used to calculate the gaze corrections to counteract the effect of nodding/shaking on gaze angles by simply applying a hand-tuned gain to the gyro signals.

- Motion commands were given to counteract the image-frame effect of the induced head shaking and nodding.

Tracking quality and the image-frame effect of perturbations applied to the head were assessed with and without injection of the VOR into gaze control. Recorded video logs were used to judge image stability and tracking quality for both cases.

The same experiment was again conducted while randomly perturbing the active head (bumping/shaking it by hand). Figure 4.7 shows a snapshot of footage obtained during the experiment. The video logs are provided in Appendix C for assessment.The evaluation confirmed that tracking quality was greatly improved using gyro stabilisation. Image blur was also significantly reduced, producing crisper images of the tracked target. The experiment demonstrates that VOR injection stabilised gaze sufficiently fast (latency was sufficiently lower than camera frame rate) such that tracking quality was significantly improved and track was rarely lost.

#### 4.5.3.2 Image-based Stabilisation

VOR is a reflexive, non-cognitive process. Without image-based verification (albeit at a lower control rate), the VOR may not fully stabilise the view. Cognitive verification and VOR gain refinement both compliment the reflex. Here, we implement image-based methods to further minimise image retinal shifts.

A grid of sample points is placed over the image $I_n$ (left, Figure 4.8). Small templates around these sample points are copied to memory (right, Figure 4.8). The grid is placed over images during forward motion, such that the left and right side magnitude of flow are approximately balanced. The templates are then searched for in image $I_{n+1}$, using normalised cross-correlation. At each grid location, a best estimate of the template's horizontal and vertical translations is recorded (in whole pixels). Then, the mode of vertical whole pixel translations


Figure 4.7: Gyro-based gaze stabilisation demonstration (*snapshot - see Appendix C for full video*).

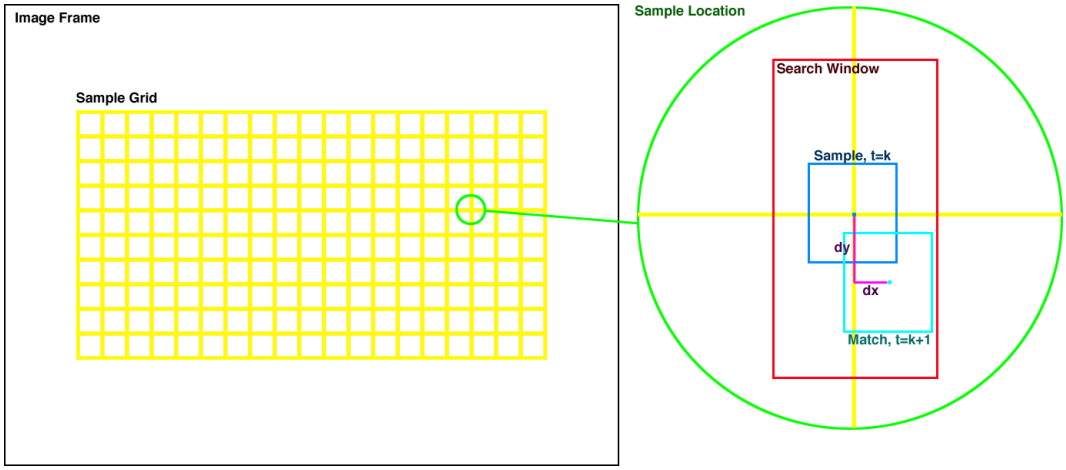



Figure 4.8: Image-based gaze stabilisation (translational).

and the mode of horizontal whole pixel translations is calculated. An output image is created that is equal to the original image shifted horizontally by $-M_h$ and $-M_v$ pixels, removing the effect of angular changes in the direction of the camera's optical axis.

Torsional perturbations can also be removed by estimating the vertical mode shift in each column of the grid. A best fit linear vertical shift gradient is then estimated across the image (Figure 4.9). The original image can then be counter-rotated to remove the perturbing torsion.

To eliminate drift and to allow auto recentreing of images after large shift corrections, the cumulative shift corrections are steadily reduced to zero at a controlled rate. This effectively high-pass filters the correction shifts and ensures that, in the case that the camera becomes steady, the shift-corrected image over time returns to being identical to the input image. For example, the correction rate may be selected such that a low-speed rotation due to smooth pursuit does not induce a correction shift in the images, whereas a fast jolt to the left would be corrected. Figure 4.10 shows a snapshot of image-based gaze stabilisation footage from the CeDAR head mounted behind a car windscreen.

This breaks down if, for example, a truck drives in front of the viewing apparatus and induces a uniform translational flow that does not correspond to apparatus motion. Image-based stabilisation does not remove motion blur due

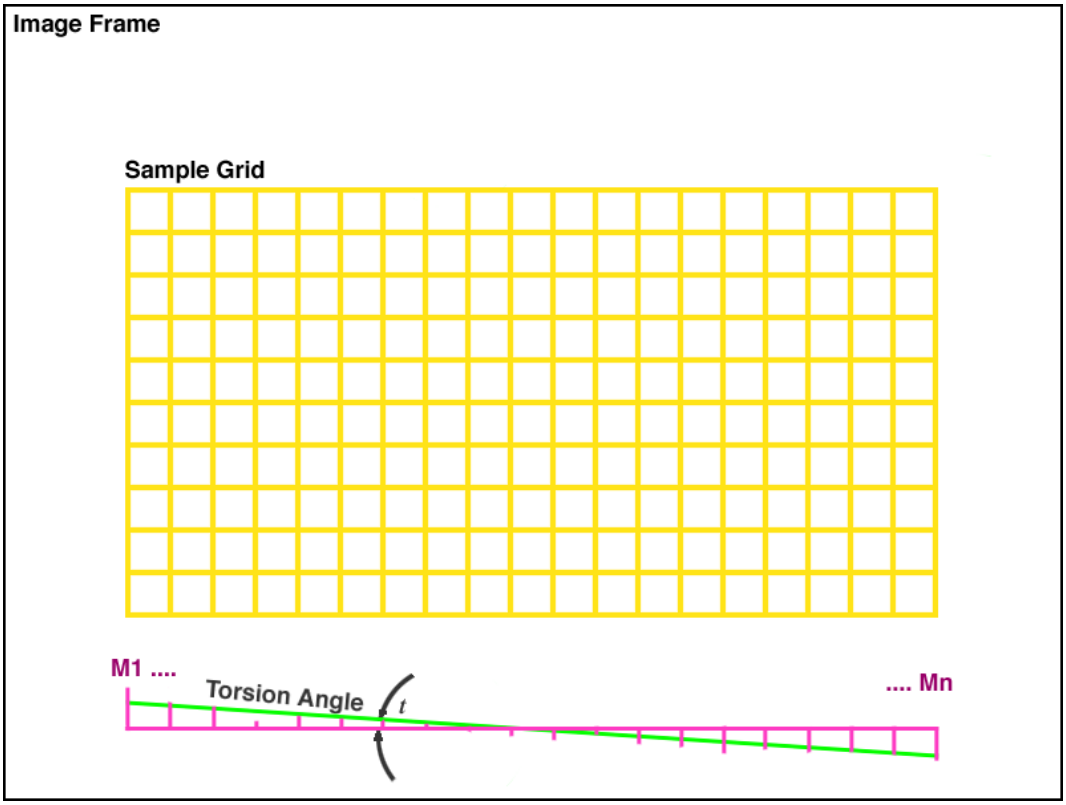



Figure 4.9: Image-based gaze stabilisation (torsional).


Figure 4.10: Online image-based gaze stabilisation demonstration (*snapshot - see Appendix C for full video*).

to perturbations (unlike the VOR), but it does help to stabilise perception of the ground plane, especially when objects moving relative to the ground plane are being tracked. In this manner, it is useful in mobile robot applications.

## 4.6 I/O Dataflow

Synchronised images with a field of view of $45^o$ are obtained from each analog camera at 30Hz at a resolution of 640 by 480 pixels. A dedicated video server reads the images from analog capture cards, resizes and converts images as required, and makes them available to other processing computers, via either push or pull requests, over a gigabit network interface. The video server uses CORBA to distribute video data over the network according to remote client requests.

Similarly, the mechanical status of the viewing apparatus and acceptance of motion control commands are handled by a separate and dedicated motion control server. The server handles PID control of all axes simultaneously, sending motion commands to encoder-controlled servo axes via the STG motion control card. The motion server also uses CORBA to handle remote client requests for axis status

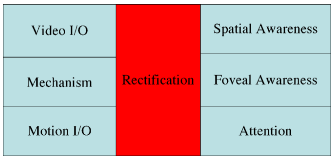
and motion commands.

## 4.7 Summary

We have presented a biologically-inspired active vision mechanism. We have also presented its control and I/O capabilities. The platform's mechanism is capable of performing the behavioural eye movements of primates. In subsequent chapters we turn our attention to developing a more flexible primate-like system capable of egocentric scene awareness.

# Chapter 5

# Active Rectification



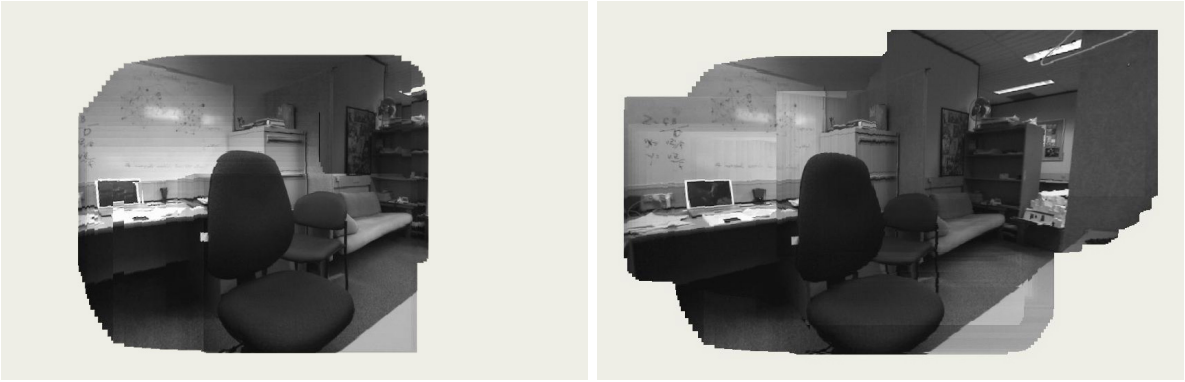



Figure 5.1: Online output of the active rectification process. Mosaics of rectified frames from right CeDAR camera at time 10sec, (left) time 20sec (right).

In this chapter we present a method to cope with the perspective distortions induced by active camera motion in real time. We transfer visual information from camera reference frames into a continuous, egocentric reference frame.

## 5.1 Introduction

Primates combine retinotropic imagery from two eyes into a unified egocentric representation that accounts for convergence. For example, we perceive long straight lines as straight and continuous in our cyclopean perception, even when

they cross the view of both eyes. For an active stereo head, online evaluation of epipolar geometry and/or direct image rectification is required to account for the image frame effect of gaze convergence (see Figure 5.7). Camera lenses can also introduce barrel distortions that need to be accounted for. Any curvature in the projection of straight lines can be removed so that the lines appear straight (and continuous where they cross binocular views), as they exist in the real scene.

Few synthetic vision models deal with active vision convergence by projecting visual stimulus into a static, egocentric (absolute) reference frame. They may instead operate in a retinal (relative) reference frame. We combine the advantages of active stereo vision and static stereo vision by rectifying and projecting active camera images into a static egocentric reference frame. We begin by briefly reviewing biological evidence for the transformation from retinal imagery to an absolute perception in primates. We then outline our approach to synthesising such a transformation. An algorithm is presented and implemented accordingly. Finally, we provide results demonstrating the output of the approach.

### 5.1.1 Evidence in Biology

As discussed, recent experimental results [Neri *et al.* (2004)] indicate that disparity processing in ventral areas involve retinal (relative) disparity, while dorsal areas are more involved in processing absolute disparities. When perceiving scene motion, estimates from many neurons are integrated into a global motion estimate. First and second-order flow perception appear to be fully combined at the level of area MT/V5 [Kandel *et al.* (2000)].

Monkeys retain a short term memory of attended locations across saccades by transferring activity among spatially-tuned neurons within the intraparietal sulcus [Merriam *et al.* (2003)], thus retaining accurate retinotropic representations of visual space across eye movements, a concept known as *efference copy*. This transfer of activity across eye movements may be used to maintain some form of static reference frame for egocentric perception. Similarly, the apparent high ocular equivalent resolution humans experience is likely to be due to the incorporation of visual information over time and viewing angles into a unified egocentric representation.

### 5.1.2 A Synthetic Approach

To achieve an absolute, egocentric perception, perspective changes due to convergence distortions need to be accounted for. This process involves online characterisation of epipolar geometry and/or image rectification. In estimating disparities along epipolar lines (or along horizontal scanlines in the case of parallel epipolar geometry rectified images) it is not necessary to account for perspective distortions due to tilt if both left and right cameras tilt simultaneously (for example, if they share a common tilt axis as exhibited by CeDAR). However, for a binocular active head with independent left and right camera tilt axes, perspective changes due to independent tilt motions need to be accounted for (although independent tilt is not a primate-inspired ability, the algorithm we present can indeed project images from independent tilt axes into a common, static, egocentric reference frame). Similarly, to integrate images into a continuous absolute perception, long straight lines should appear continuous and at the same orientation across eye motions. Perspective distortions due to tilt, whether binocularly common or independent, or even monocular, must then be accounted for.

We propose a simple method that enables active multi-camera image rectification. As we shall see, our approach transforms images into mosaics that are globally fronto-parallel. Projection of images into a reference frame exhibiting parallel epipolar geometry enables existing static multiple-camera (stereo) algorithms that benefit from pre-computed or parallel epipolar geometry (such as depth mapping) to operate on active multi-camera platforms. The algorithm therefore enables the operation of any static stereo algorithms on active stereo platforms. We analyse the general case where any number of cameras in any geometric configuration can be used, for example, any relative translations and rotations between multiple cameras.

## 5.2 Background

The rectification algorithm is based upon the common pinhole camera model. We review the pinhole camera model and associated multiple camera epipolar geometry. In so doing we define the nomenclature adopted to formulate the algorithm.

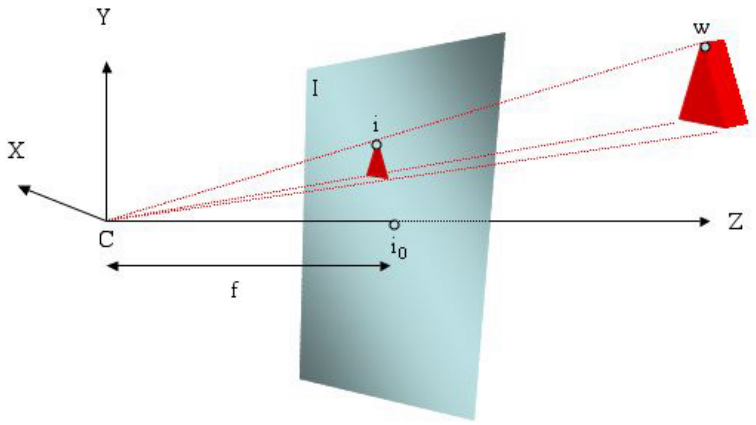



Figure 5.2: Pinhole camera model. Definition of parameters.

Unless specifically referenced otherwise, theory in this section (Section 5.2) is referenced from texts such as [Hartley & Zisserman (2004)].

### 5.2.1 Camera Model

The pinhole camera model represents the camera by its optical centre $C$ and image plane $I$. The image plane is reflected about the optical centre to be located in front of the camera. A line passing through a point $w$ in the real world $W$ at coordinates $w \in W$ and the camera optical centre at $c \in W$ intersects the image plane $I$ at image coordinates $i$. The distance along the optical axis from the optical centre $c \in W$ to the image plane centre $i_0 \in W$ is equivalent to the camera focal length $f$. Figure 5.2 shows the pinhole camera model.

The linear transformation from three-dimensional homogeneous world coordinates $\tilde{w} = [x, y, z, 1]^\top$ to two-dimensional homogeneous image coordinates $\tilde{i} = [u, v, 1]^\top$ is the perspective projection $\tilde{P}$ [Hartley & Zisserman (2004)]:

$$\tilde{i} \cong \tilde{P}\tilde{w} \tag{5.1}$$

The perspective projection matrix can be decomposed by QR factorisation into the product:

$$\tilde{P} = A[R|t] \tag{5.2}$$

where rotation matrix $R$ and translation vector $t$ denote the extrinsic camera parameters that align the camera reference frame with the world reference frame, and $A$ depends only on the intrinsic camera parameters.

Rotation matrix $R$ is the standard 3 by 3 rotation matrix constructed from rotations about the $x$, $y$ and $z$ axes:

$$\begin{vmatrix} c(\theta_y)c(\theta_z) & s(\theta_x)s(\theta_y)c(\theta_z) - c(\theta_x)s(\theta_z) & c(\theta_x)s(\theta_y)c(\theta_z) + s(\theta_x)s(\theta_z) \\ c(\theta_y)s(\theta_z) & s(\theta_x)s(\theta_x)s(\theta_z) + c(\theta_x)c(\theta_z) & c(\theta_x)s(\theta_x)s(\theta_z) - s(\theta_x)c(\theta_z) \\ -s(\theta_y) & s(\theta_x)c(\theta_y) & c(\theta_x)c(\theta_y) \end{vmatrix} \tag{5.3}$$

where *s()* denotes sin() and *c()* denotes cos().

A is of the form:

$$A = \begin{vmatrix} \alpha_u & \gamma & u_0 \\ 0 & \alpha_v & v_0 \\ 0 & 0 & 1 \end{vmatrix} \tag{5.4}$$

where $\alpha_u$ and $\alpha_v$ are the focal length, expressed in units of pixels, along the horizontal and vertical image plane axes respectively; $(u_0, v_0)$ is the image plane coordinate of principal point $i_0$; $\gamma$ is the skew factor that models any deviation from orthogonal $u - v$ axes. Traditionally, the origin of the $u - v$ axis is in the top left corner of image plane $I$.

### 5.2.2 Epipolar Geometry

A point $i$ in the image plane $I$ corresponds to a ray in three-dimensional space $W$. Given two stationary pinhole cameras, $C_a$ and $C_b$, pointed towards the same three-dimensional world point $w$, points in the image plane $I_a$ of camera $C_a$ will map to lines in the image plane $I_b$ of camera $C_b$, and vice versa. Such lines are called epipolar lines. All epipolar lines in image plane $I_b$ will be seen to radiate from a single point called the epipole, which lies in the plane of $I_b$, but depending on camera geometry, may or may not lie within the viewable bounds of $I_b$. The epipole is the mapping of the world coordinates of the optical centre of camera $C_a$ to the extended image plane $I_b$ of camera $C_b$. The baseline connects optical centres of $C_a$ and $C_b$, and intersects the image planes at the epipoles. Figure 5.3 shows the described epipolar geometry.

Stereo algorithms may require locating the same real-world point $w$ in two camera image planes $I_a$ and $I_b$. This involves a two-dimensional search to match

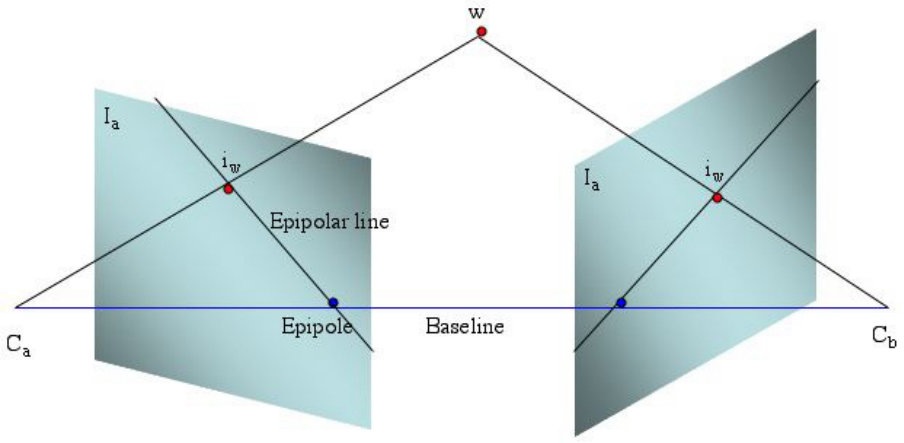



Figure 5.3: Epipolar geometry. Definition of parameters.

point $i_{w_a} \in I_a$ with the corresponding point $i_{w_b} \in I_b$. Once the epipolar geometry is known, this two-dimensional search is reduced to a one-dimensional search for $i_{w_b} \in I_b$ along the epipole in $I_b$ that corresponds to $i_{w_a} \in I_a$. In the special case that image planes are coplanar, both epioles are at infinity and epipolar lines will appear horizontal in each image frame. In this case, the correspondence problem is further simplified to a one-dimensional search along an image row (Figure 5.4). Any set of images acquired from cameras with overlapping fields of view can be transformed such that this special case is enforced - a process called parallel epipolar rectification.

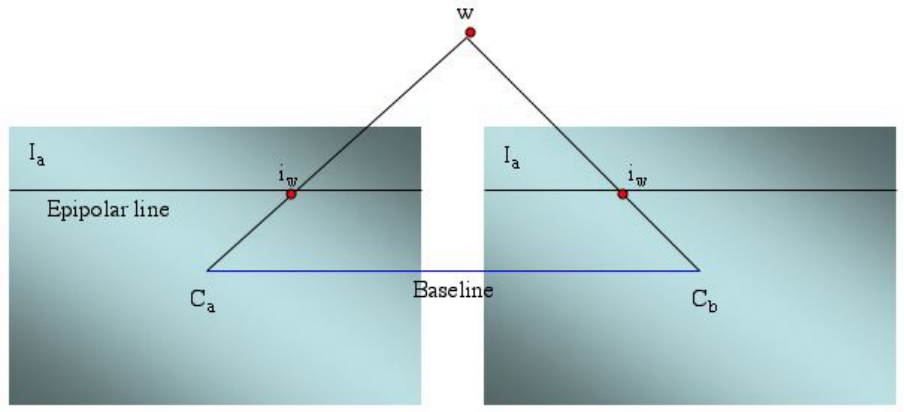



Figure 5.4: Rectified epipolar geometry enforced by fronto-parallel cameras. The image planes exhibit horizontal epipolar lines lines.

## 5.3 Active Rectification

We project images of the scene through the camera optical centres onto a virtual plane parallel to the baseline, whose orthogonal vector points in the head's z-axis direction, as per Figure 5.5. The projection (described below mathematically) transforms the camera images into the large fronto-parallel aligned imaging plane from a virtual camera pointing in a direction aligned with the z-axis of the head (the starting/home direction of the cameras). Where the cameras deviate from the home/start position (fronto-parallel alignment) via pan and/or tilt, the transform distorts the images; an example is shown in Figure 5.7. Such camera motion also alters the projected location of the camera image in the virtual plane. The *mosaic* image is the accumulation of multiple such projections onto the virtual plane from different camera geometries. The construction of two such fronto-parallel mosaics (by projecting both left and right camera images onto the same plane) ensures that parallel epipolar geometry is maintained throughout the contents of the mosaics. The relative relations between the observed parts of the scene are preserved across camera axis motions in this statically-imposed reference frame. The mosaic, or regions of it, can then be fed into standard multi-camera functions that rely on parallel epipolar geometry. As an example, this mosaicing active rectification approach will be shown to function with a standard depth mapping algorithm (Chapter 6), and thereby actively build an occupancy grid representation of the scene. Before applying the rectifying transformation, lens barrel distortion and camera geometry must be determined.

First, the intrinsic camera parameters must be determined for each camera. Lens distortion has two forms, barrel and pincushion. Distortion tends to be most significant in wide angle, telephoto and zoom lenses. It can be highly visible on tangential lines near the boundaries of the image, but it is not visible on radial lines. In a well-centred lens, distortion is symmetrical about the centre of the image but lenses can be decentred due to poor manufacturing quality or shock damage. We use the standard Matlab camera calibration toolbox[1] to characterise each camera individually. A lookup table is created from the parameters output from the Matlab calibration toolbox that maps camera image pixels to distortion

[1]The Matlab camera calibration toolbox is available at *http : //www.vision.caltech.edu/bouguetj/calib_doc/*.

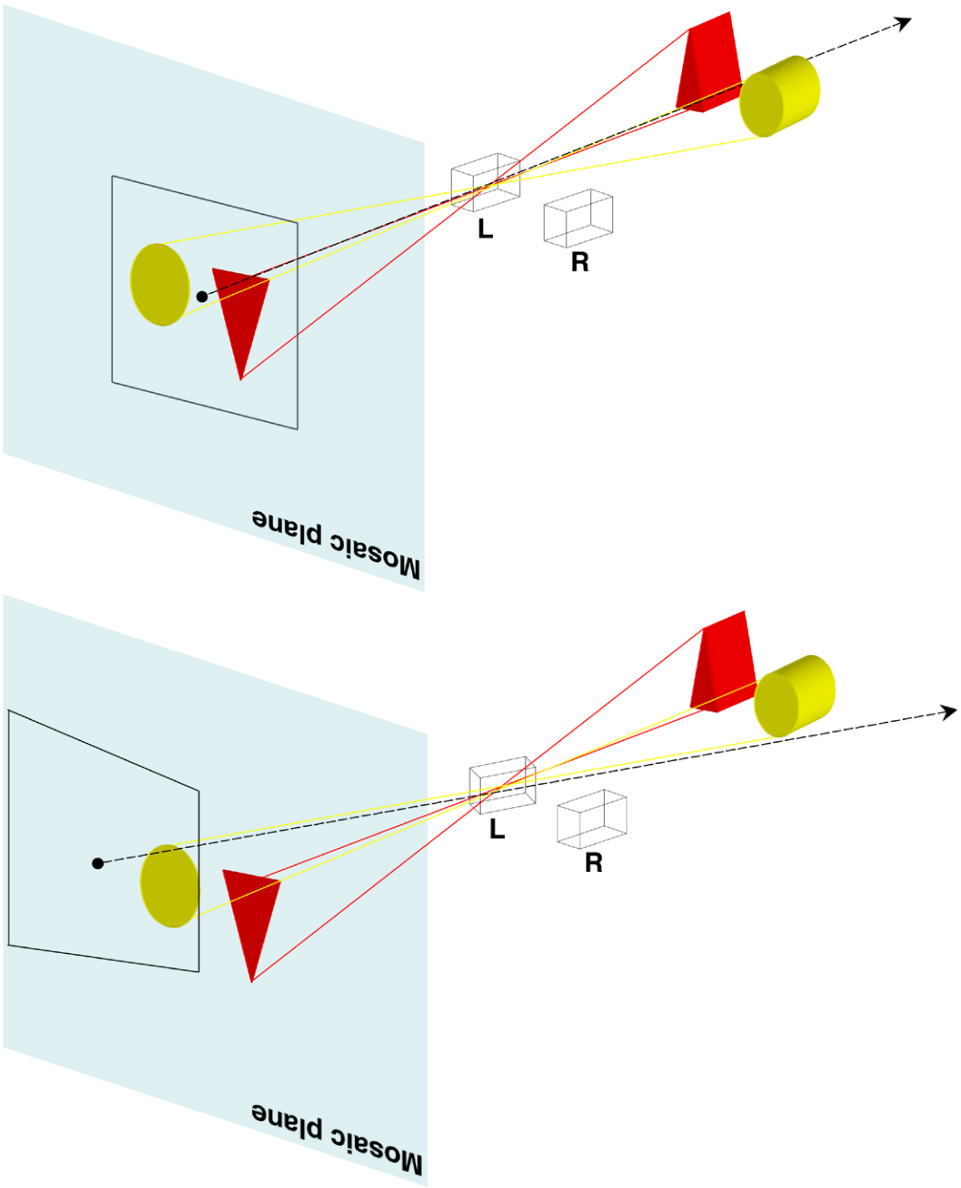



Figure 5.5: Camera images projected into a static reference frame. Top: both cameras at the home position; the left camera view of the scene is projected through the left camera optical centre onto the mosaic plane that is oriented parallel to baseline. Bottom: later, the left camera verges right about its optical centre, projecting a new view of the scene onto a different coverage area of the mosaic plane; the projection of the image left frame is no longer rectangular in mosaic space; however, the projection of objects still in the camera view (for example, the yellow cylinder) remains in the same mosaic location. The projection of the right camera images onto the right mosaic is not shown but differs only by a horizontal translation.

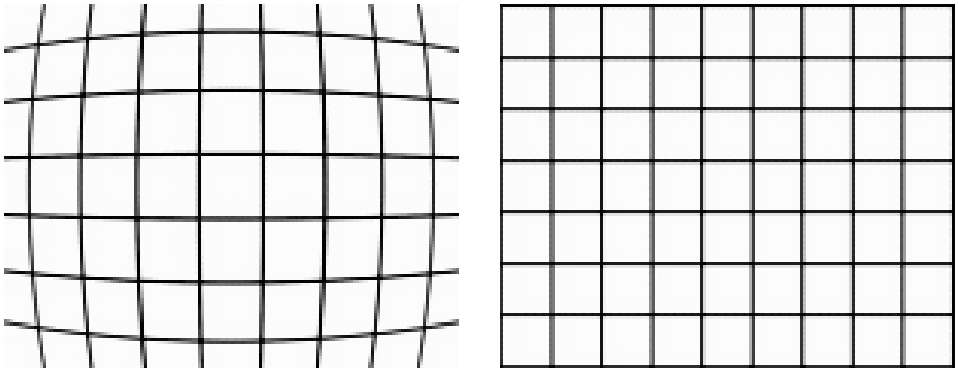


Figure 5.6: Barrel distortion (simulation). Uncorrected view (left); and the corrected view (right).

corrected image coordinates. The mapping is applied to every image. Figure 5.6 demonstrates the correction of barrel distortion.

Next, the real-world rigid transformations between camera positions must be determined. This may be done by a number of methods. Visual techniques such as the *scale-invariant feature transform* (SIFT) algorithm [Se *et al.* (2001)] or Harris corner detection [Harris & Stephens (1988)] can be used to identify features common to each camera view, and thereby infer the geometry. Alternatively, encoders can be used to measure angular rotations. A combination of visual and encoder techniques could also be adopted to obtain the camera relationships to a more exacting degree. Once the extrinsic geometric relations between any number of cameras is known, we can determine the epipolar geometry. We can then calculate the transformation that projects the camera images onto the virtual plane, for each camera.

### 5.3.1 The Rectifying Projection

CeDAR is a stereo head configuration so we consider here the case of two cameras, although any number may be used as long as the transformations between cameras are known. The process involves projecting camera images into a plane parallel to the baseline whose orthogonal vector points in the z-direction. This projection transforms the images such that they appear to come from parallel aligned cameras pointing in a directional orthogonal to the baseline, as shown in Figures 5.4 and 5.7. The sequential steps involved in the rectification process are

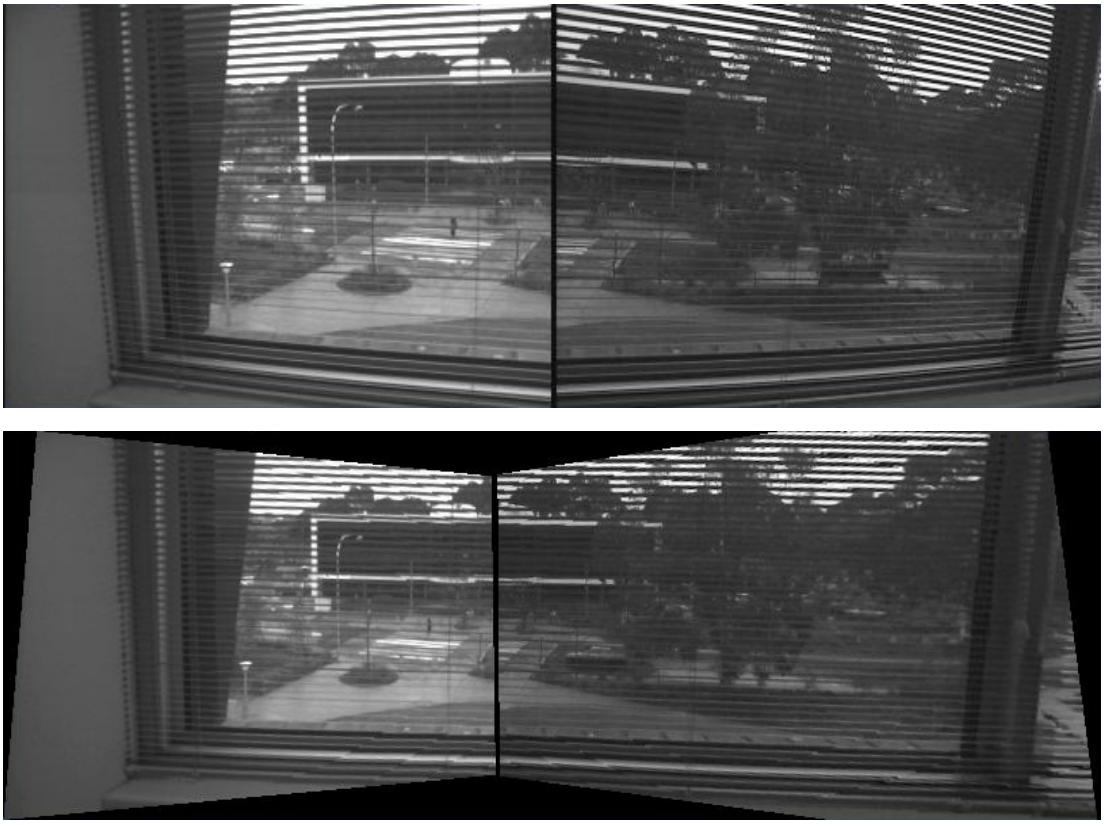


Figure 5.7: Scene viewed through blinds demonstrating the output of the active rectification process. The original right and left images respectively (top); and the same images rectified such that parallel epipolar geometry is enforced (bottom).

now presented.

#### 5.3.1.1 Intrinsic Parameters

We assume that the focal length of each camera remains constant throughout use so that the intrinsic parameters need not be recalculated. We obtain the intrinsic camera parameters for left and right cameras, $A_l$ and $A_r$, from Matlab Camera Calibration Toolbox single camera calibrations. Example parameters obtained from our cameras were:

$$A_l = \begin{vmatrix} 373.45 & 0.00 & 145.81 \\ 0.00 & 374.30 & 128.26 \\ 0.00 & 0.00 & 1.00 \end{vmatrix}$$

$$A_r = \begin{vmatrix} 368.20 & 0.00 & 152.45 \\ 0.00 & 370.90 & 131.81 \\ 0.00 & 0.00 & 1.00 \end{vmatrix}$$

These parameters are used to determine the rectifying transformation to correct camera barrel distortion, as per Matlab Camera Calibration Toolbox. Barrel rectification is a static transformation that only needs to be determined once, and applied identically to all camera images. The rectification is sped up by creating a lookup table of transformed pixel relocations, and is applied to all images obtained from the camera.

#### 5.3.1.2 Extrinsic Parameters

Extrinsic parameters are those that depend upon head geometry. For processor economy, we are prepared to sacrifice a minimal reduction in accuracy in favour of real-time performance. For CeDAR and many other binocular platforms, the camera translations are kept constant. Encoders are used to measure camera rotations about their optical axes. This eliminates the computational costs associated with image-based methods of extracting more precise extrinsic parameters. Experimentation has shown that the encoder resolution is sufficiently accurate for us to assume that systematic errors, such as encoder drift, are for our purposes insignificant.

For two cameras, rectifying the images to a plane parallel to the baseline ensures that parallel epipolar geometry is enforced. For more than two cameras in a configuration where there is no single baseline, we need to declare a baseline and rectify the camera views to this line. Since we are considering the case of a stereo configuration, a common baseline exists and rotations around the optical centres are sufficient to align retinal planes and enforce parallel geometry. For multiple camera configurations where there are more than two cameras and no common baseline, rotations around camera optical centres will enforce parallel epipolar geometry but will not ensure that rows in each image align. In this case, the scaling effect of translations perpendicular to the baseline would also have to be accounted for. In the case of a stereo rig such as CeDAR, this problem does not exist.

We proceed to build $R_l$ and $R_r$ from the extrinsic parameters $\theta_x, \theta_y, \theta_z$ read from the encoder data at the time the images were obtained. Since our configuration has a common baseline, translations $t_l, t_r$ are not required for rectfication and are set to zero vectors. We assume the focal lengths, components in matrix A, to be constant, but where it is possible to read the focal lengths from the cameras, they can be entered directly into matrix A. Beginning from the static stereo rectification method outlined by Fusiello *et al.* (2000), we first create the current left and right projection matrices $\tilde{P_{ol}}, \tilde{P_{ol}}$ according to:

$$\begin{aligned} \tilde{P_{ol}} &= A_l[R_l|t_l] \\ \tilde{P_{or}} &= A_r[R_r|t_r] \end{aligned} \tag{5.5}$$

#### 5.3.1.3 Determine Desired Projection Matrices $\tilde{P_{nl}}, \tilde{P_{nr}}$

Parallel epipolar recitification involves rectifying images to a plane aligned with the baseline whose normal vector points in the z-direction. In this case, angles $\theta_x, \theta_y, \theta_z$ are zero in the desired rotation matrices $R_{l0}, R_{r0}$. Desired translations $t_{l0}, t_{r0}$ are also zero. We can then create the desired new left and right projection matrices $\tilde{P_{nl}}, \tilde{P_{nl}}$:

$$\begin{aligned} \tilde{P_{nl}} &= A_l[R_{l0}|t_{l0}] \\ \tilde{P_{nr}} &= A_r[R_{r0}|t_{r0}] \end{aligned} \tag{5.6}$$

#### 5.3.1.4 Determine Rectification Transformations $T_l, T_r$

Now that the current and desired projection matrices are known for each camera, the transformation T mapping $\tilde{P}_o$ onto the image plane of $\tilde{P}_n$ is sought.

Each projection matrix $\tilde{P}$ can be written in the form [Fusiello *et al.* (2000)]:

$$\tilde{P} = \left[\begin{array}{c|c} q_1^\top & q_{14} \\ q_2^\top & q_{24} \\ q_3^\top & q_{34} \end{array}\right] = [Q|q] \tag{5.7}$$

substituting this form of $\tilde{P}$ into Equation 5.1 gives

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \left[\begin{array}{c|c} q_1^\top & q_{14} \\ q_2^\top & q_{24} \\ q_3^\top & q_{34} \end{array}\right] \tilde{w} \tag{5.8}$$

This can be rearranged to its Cartesian form:

$$\begin{aligned} u &= \frac{q_1^\top w + q_{14}}{q_3^\top + q_{34}} \\ v &= \frac{q_2^\top w + q_{24}}{q_3^\top + q_{34}} \end{aligned} \tag{5.9}$$

From Equation 5.8, the Cartesian coordinates $c$ of the optical centre C is reduced to:

$$\begin{bmatrix} u_o \\ v_o \\ 1 \end{bmatrix} = \left[\begin{array}{c|c} q_1^\top & q_{14} \\ q_2^\top & q_{24} \\ q_3^\top & q_{34} \end{array}\right] \tilde{c} \tag{5.10}$$

where $(u_o, v_o)$ is the image frame origin $(0, 0)$ and $\tilde{c}$ is the homogeneous coordinate of the optical centre. We rearrange the above to obtain the Cartesian form:

$$c = -Q^{-1}q \tag{5.11}$$

So $\tilde{P}$ can be written:

$$\tilde{P} = [Q| - Qc]. \tag{5.12}$$

In parametric form, the set of 3D points $w$, associated with image point $\tilde{i} \cong \tilde{P}\tilde{w}$ becomes:

$$w = c + \lambda Q^{-1}\tilde{i} \tag{5.13}$$

where $\lambda$ is a scale factor. From Equation 5.7 we can write for $\tilde{P_o}$ and $\tilde{P_n}$:

$$\begin{aligned} w &= c + \lambda_o Q_o^{-1} \tilde{i_o} \\ w &= c + \lambda_n Q_n^{-1} \tilde{i_n} \end{aligned} \tag{5.14}$$

hence:

$$\tilde{i_n} = \lambda Q_n Q_o^{-1} \tilde{i_o} \tag{5.15}$$

and so:

$$T = Q_n Q_o^{-1} \tag{5.16}$$

$T$ is determined for each camera.

#### 5.3.1.5 Apply Rectification

$T_l$ is then applied to the original left image, and $T_r$ to the original right image. If the camera is not in the home position, transform $T$ transforms the camera image to a location outside of the frame of the original image. To save memory, we first apply $T$ to the corner points of the original image to find the expected size and location of the transformed image. We can then allocate memory for the extent of the transformed image only (rather than using mosiac-sized memory allocations for each frame), and apply an offset translation to $T$ such that the transformed image has the origin at [0,0]. In this manner, we contain the transformed image in a minimal amount of memory. The transformed image can then be augmented into a common mosaic memory space (or displayed on screen within a frame that represents the mosaic) at the location according to the original transform as follows.

#### 5.3.1.6 Mosaic Images

The location of the projected image in the mosaic is determined by transforming the principal (central) point $(x_p, y_p)$ of the original image under $T$ to obtain the location$(x_T, y_T) = T[x_p, y_p]$. While the head is in the initial home position, the principal (central) point of the mosaic provides the coordinates at which images coming from the head are to be augmented. This corresponds to $T = I$ where $I$ is the identity matrix (no apparent rectifying transformation).

**Active Rectification Algorithm:**

1. Determine intrinsic parameters to remove lens distortions.
2. Determine extrinsic parameters (head geometry).
3. Determine desired projection matrices $\tilde{P_{nl}}, \tilde{P_{nr}}$.
4. Determine rectification transformations $T_l, T_r$.
5. Apply rectification transformation to images.
6. Mosaic images.

Figure 5.8: Summary: active rectification algorithm.

We chose a mosaic frame size that is large enough to display the region of the scene in which we are interested. Figure 5.8 summarises the active rectification algorithm. Figure 5.1 is an example of output from the mosaicing process.

## 5.4 Results

Figure 5.7 shows input images and non-mosaiced rectified images with enforced parallel epipolar geometry. Figure 5.9 shows further non-mosaiced output where horizontal lines have been drawn to highlight the enforced parallel epipolar geometry.

It can be seen that horizontal scanlines in the images become aligned such that the images are then suitable for depth analysis from pure horizontal disparity. They are also rectified for insertion into mosaics. Figure 5.10 shows the definition of rectification result parameters. Such parameters include the left and right coordinates for the positioning of images in the mosaic that share a common vertical coordinate due to the presence of a common tilt axis; and $d$, the convergence disparity (horizontal distance in pixels) between left and right mosaic positioning of rectified images. The convergence distance permits conversion from relative image disparity to absolute disparities for egocentric depth perception (Chapter 6).

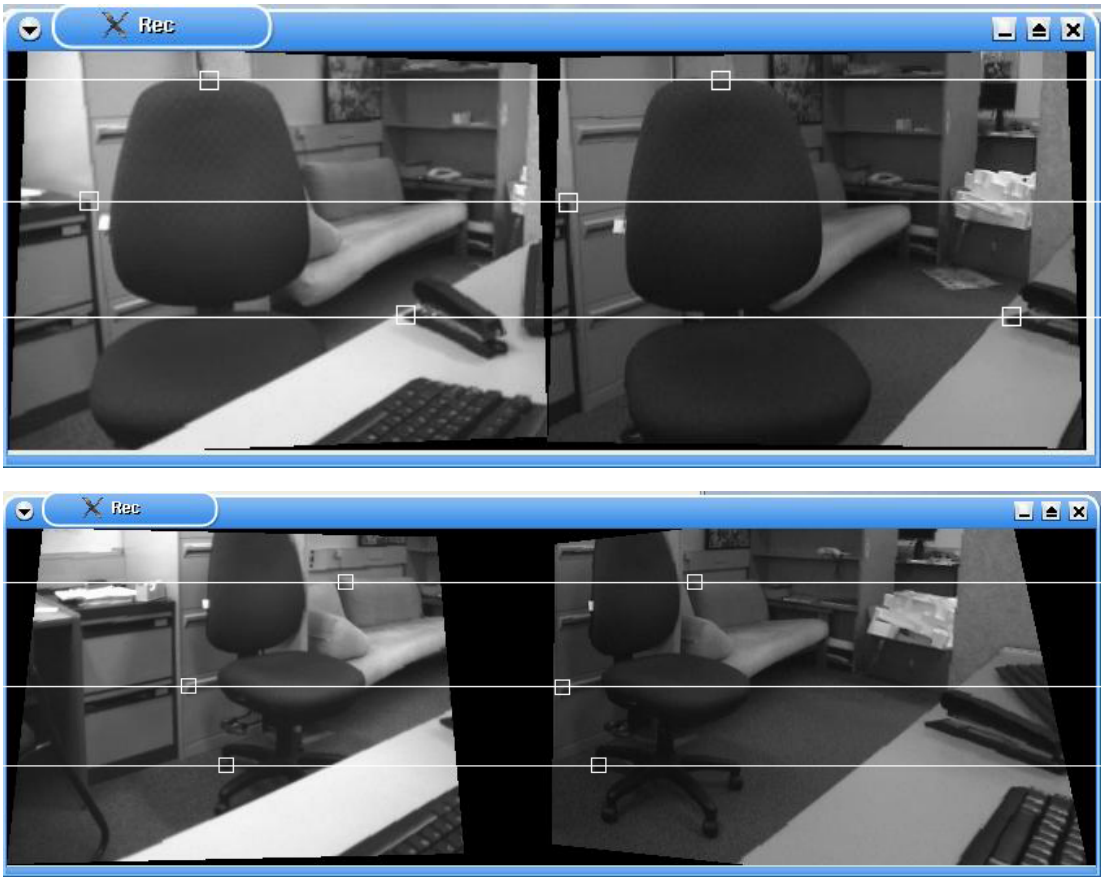



Figure 5.9: Online output of active rectification. The annotations show the enforced alignment of the top of the chair, corner of the cabinet and base of the stapler (top). Later, both cameras have moved but parallel epipolar geometry remains enforced via online rectification (bottom); the alignment of the top of the couch, the cabinet handle and the base of the chair are annotated.

The movies in Figures 5.11 and 5.12 show output of the mosaicing process. In these examples, operation incorporates network processing: images are obtained from the cameras by the video server; the camera geometry is obtained from the motion control server; a client node determines the rectification transformation, acquires greyscale images, applies the rectifying transform and displays the results in a mosaic at the coordinates determined during rectification. In this example, output display was achieved at full 30Hz camera frame rate, including saving of the display buffer for creation of the demonstration movie.

The movie in Figure 5.13 shows networked processing output: images are obtained from cameras by the video server; the camera geometry is obtained from the motion control server; a secondary server node determines the rectification transformation, acquires greyscale images, applies the rectifying transform and distributes the rectification results; a final node (the client) receives the rectified images and parameters, does additional cue processing (in this instance the cue is a form of saliency, but any cue processing could be substituted), and displays the results in a mosaic. In this example, output display was achieved at 24Hz, the additional cue processing requiring extra CPU cycles.

## 5.5 Discussion

We rely on encoder data to obtain the camera geometry. Therefore, accuracy in rectification performance relies on initial homing accuracy and encoder calibration accuracy. Homing involves setting the cameras to the position where they are parallel and pointing perpendicularly away from the baseline.

Inaccuracies in determining camera geometry may also be introduced where the cameras do not rotate around the camera centres. This may be invoked by poor initial alignment of the cameras within the head, or by variations in focal length during camera operation, which is often unavoidable while viewing scenes with significant depth changes. Timing is also a potential source of error. If the head angles are not recorded at the same time that the images are captured, the image contents may not be rectified accurately. As we shall see, we use algorithms that perform robustly despite such error.

When in the home position, the mosaic location of camera images corresponds to the centre of the mosaic. If the contents of the cameras correspond to objects

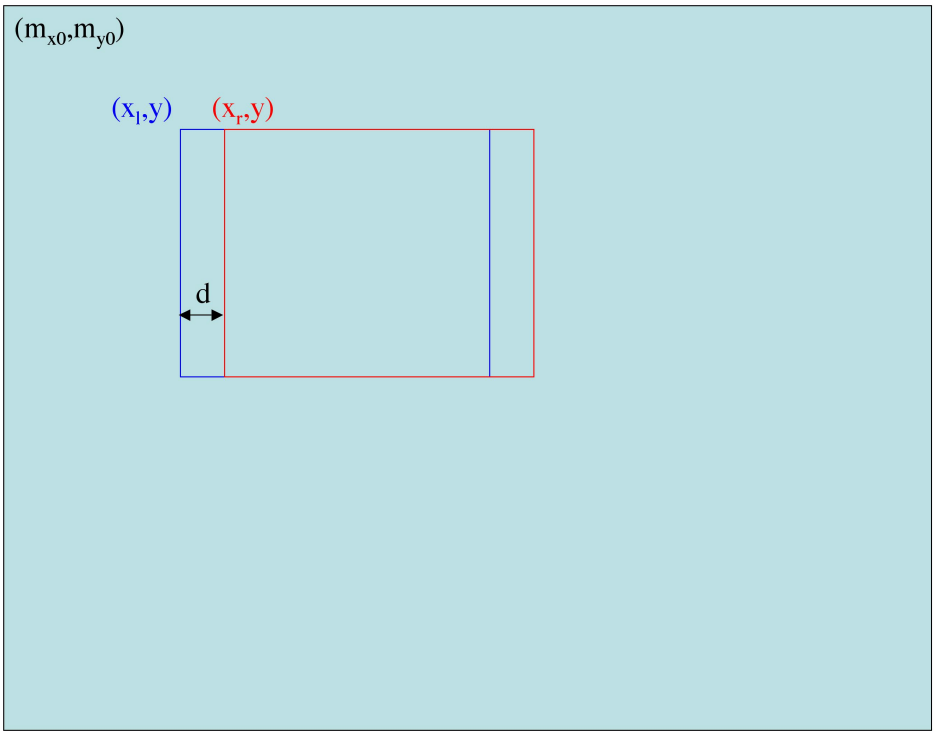



Figure 5.10: Rectification result parameter definitions. The shaded frame represents the mosaic. The red and blue frames contain the rectified right and left camera images respectively.


Figure 5.11: Online image rectification and mosaicing, left camera. In this demonstration, the camera was moved by hand while the images were automatically rectified (*snapshot - see Appendix C for full video*).

at large scene depths, it is possible to augment images from both the left and right cameras into a single mosaic. If scene objects are closer, perspective disparities are significant such that augmentation into a single mosaic would place near scene objects at different locations in the mosaic, the disparity induced depending on the scene depth of the object. We therefore maintain separate left and right mosaics.

Integration of images over time and space into mosaics is similar to the manner in which humans assemble images into a broad and resolute perception. Humans combine variable resolution (foveal) imaging and broad binocular peripheries into a high *equivalent resolution* perception (hundreds of megapixels, discussed in Chapter 2). The synthetic active rectification system incorporates (non-zoomed) images of constant spatial resolution into an approximately one megapixel resolution mosaic, significantly lower than the equivalent resolution of humans. Variable resolution cameras (log polar cameras for example), or increased foveal sampling and peripheral down-sampling could be used to increase the system's equivalent resolution and/or to provide a broader periphery, while


Figure 5.12: Online image rectification and mosaicing, left camera. The sinusoidal motion in this demonstration is effected by computer control (*snapshot - see Appendix C for full video*).


Figure 5.13: Online cue mosaic construction of saliency cue (*snapshot - see Appendix C for full video*).

not incurring additional computational expense. Increasing the size of the mosaic in the current constant resolution implementation does not increase processing requirements (though it may increase memory use), so it is likely that such a regime could achieve equivalent resolutions an order of magnitude closer to that of human vision.

For creating absolute disparity maps, standard image frame disparity maps are created, then parameter $d$ is added to every disparity estimate. Alternatively, the disparities can be estimated directly from the left and right mosaic contents. If optical flow is calculated within the mosaic reference frame the result is pure scene flow. Mosaicing accounts for the effect of deliberate camera motion on the contents of the original camera image. In this manner, the rectification process converts relative retinal camera images to an absolute, egocentric reference frame ready for spatial awareness (Chapter 6).

We wrap the algorithm in a CORBA server that takes images from the video server and head parameters from the motion control server. Rectified Y,U, or V[1] images and rectification parameters such as mosaic coordinates for left and right rectified image positions are distributed to subsequent processing nodes in the processing network. Rather than distributing the entire mosaics, we save bandwidth by distributing only the rectified images and mosaicing parameters. Subsequent processing nodes, such as those used to determine depth and flow can obtain and preserve an egocentric representation from this data. Epipolar alignment is preserved in any subsequent cue maps produced by servers operating on the rectified images. In fact, we separate, rectify, then distribute rectified Y,U and V images independently. We decouple channels because some subsequent servers only require one of the channels, saving bandwidth by only distributing channels or subsequent cues required by a dependent node.

## 5.6 Summary

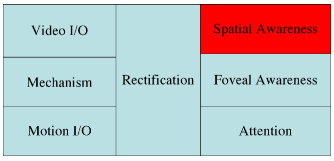
We have provided biological evidence in support of an egocentric scene perception. We have outlined a synthetic approach based on combining epipolar rectification and mosaicing. We have detailed the steps involved in implementing such a

[1]YUV is a colour space where channel Y contains greyscale intensities and U and V are intensity normalised chrominance channels.

regime. We have used the method to rectify images for online mosaic construction and display, and to transfer them, together with the mosaicing parameters, to a client PC for remote processing and/or display.

# Chapter 6

# Spatial Perception



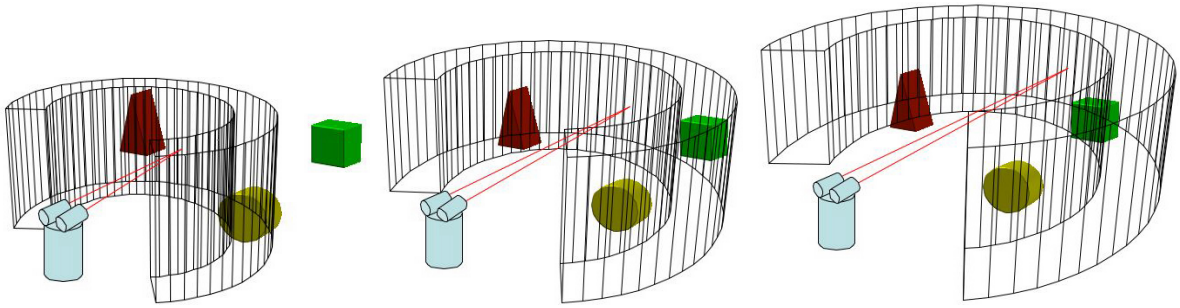



Figure 6.1: Building spatial perception by scanning the fixation point over the scene. For a given camera geometry, searching for pixel matches between the left and right stereo images over a small disparity range defines a volume about the *horopter*. By varying camera geometry, this measurable volume can be scanned over the scene. Initially, only the circle lies within the measurable volume (left). As the cameras diverge, the triangle (middle), then the cube (right), become detectable.

In this chapter we describe spatial awareness. We develop a biologically-inspired framework particularly suited for real-time synthetic active stereo vision.

## 6.1 Introduction

Over recent years, calibrated stereo vision has proved an economical sensor for obtaining 3Dd range information [Banks & Corke (1991)]. Traditionally, stereo sensors have used fixed geometric configurations. This passive arrangement has proven effective in obtaining range estimates for regions of relatively static scenes. In reducing processor expense, most depth mapping algorithms match pixel locations in separate camera views within a small disparity range, for example, ±32 pixels. This means that depth maps obtained from static stereo configurations are often dense and well populated over portions of the scene around the fixed horopter, but they are not well suited to dynamic scenes or tasks that involve resolute depth estimation over larger scene volumes. As discussed, an active stereo vision approach can offer many benefits over static stereo approaches.

Transferring images from active cameras into a static egocentric reference frame, as described in the previous chapter, facilitates head-centred (egocentric) perception. We now consider the development of egocentric spatial awareness using active vision. Components of spatial perception include an awareness of scene structure and scene motion.

As discussed, various cues contribute to the human perception of scene depth. Many of these cues are complex and may only provide relative depth information. We therefore concentrate on depth from disparity estimations. For a stereo camera pair observing a 3D environment, disparity can be defined as the retinal displacement between the matching projections of scene points in the left and right images (Figure 6.2).

In undertaking task-oriented bahaviours it may be necessary to give attention to a subject that is likely to be moving relative to us. By actively varying our gaze geometry it is possible to place our resolute foveas over any of the locations of interest in a scene, thereby obtaining maximal resolution in the vicinity of those locations, including maximising the resolution of depth estimates. Where a subject is moving, gaze can be made to follow the subject such that this information is continually maximised. Figure 6.1 shows how the *horopter* can be scanned over the scene by varying camera geometry for a synthetic stereo configuration. This approach is potentially more efficient than methods that use static cameras because a small constant disparity range scanned over the scene is


Figure 6.2: An example of binocular disparity. The insets show (from left) the left and right camera views and an overlay of left and right views. Gaze is fixated upon the mannequin's forehead. Vector A-B represents disparity for the ball. The images have not undergone parallel epipolar rectification, so there is a vertical component to the disparity vector. A wide baseline introduces parallax disparity at the manequine's face.

computationally cheaper and obtains more dense results than a search over large disparity range (or variable ranges) from a static camera configuration. Peripheral lens distortions mean that disparities determined in the periphery of such a static system may be prone to inaccuracies if such distortions are not accounted for. Active mechanisms can obtain disparity estimates from the camera optical centres where such distortions are minimal. Placing the optical centres over an object of interest also maximises contextual depth information about an object, whereas the quantity of such information may be reduced when analysing the disparity of objects situated near the periphery of a static camera configuration. Additionally, multiple views of the scene from different depth mapping geometries can be combined to reinforce the certainties associated with an estimate of scene depths. Varying the camera geometry not only helps to improve the resolution of range information about a particular location, but by scanning the horopter, it also increases the volume of the scene that may be densely depth mapped.

We begin by seeking evidence for spatial computations in the primate brain. We then discuss existing methods to synthesise the computation of retinal disparity. The brain augments spatial estimates into an egocentric perception. We propose a method to augment active vision disparity data into an egocentric, unified occupancy grid representation. Finally, we extend the presented occupancy grid framework and present results that demonstrate how the occupancy grid can

be used to extract information about the surroundings. In particular, we focus upon 3D scene motion and 3D cue-surface correspondences.

### 6.1.1 Retinal Disparity in Biology

The two different perspectives from the eyes of the human vision system leads to slight displacements of objects (disparities) in the two monocular views of the scene (Figure 6.2). The human vision system is able to use these disparities (amongst other cues) for depth estimation and to merge both views into a fused cyclopean view, a 3D representation of the scene.

Disparity estimations occur across around 80% of the area of the human visual field [Rodieck (1998)], corresponding to the majority of the area of the left and right views that overlap. The point we are gazing upon will appear once, crisp, and in focus in our perceived cyclopean image (overlapping visual fields). Points further away from and closer to our gaze point, though not in focus, will appear twice in our cyclopean view due to disparity. If the point is closer than the gaze point, it will appear once displaced further to the right (contributed from our left eye's view) and again further to the left (from our right eye's view). This is how we perceive *crossed*, or negative disparity. Points beyond the gaze point will appear once displaced further to the left (contributed from our left eye's view) and a second time further to the right (from our right eye's view). This is how we perceive *uncrossed* or positive disparity. Figure 6.3 describes crossed and uncrossed disparity. Though we are not physically aware of "retinal displacements", whether an object's retinal image on the two eyes has crossed or uncrossed disparity immediately tells us whether the object is in front of or beyond the fixation point.

Evolution has given most land-dwelling animals capable of disparity estimation a fronto-parallel vision geometry, probably due to the fact that, we live and interact in a predominantly horizontally planar world where objects of interest tend lie to the left and right of each other rather than above or below. The geometry of such biological vision systems simplifies the disparity problem. Vertical disparities are largely eliminated by assuming a fronto-parallel camera geometry. Though the visual cortex does not perform one-dimensional scanline analysis, stereoscopic depth estimation can be reduced to a one-dimensional spa-

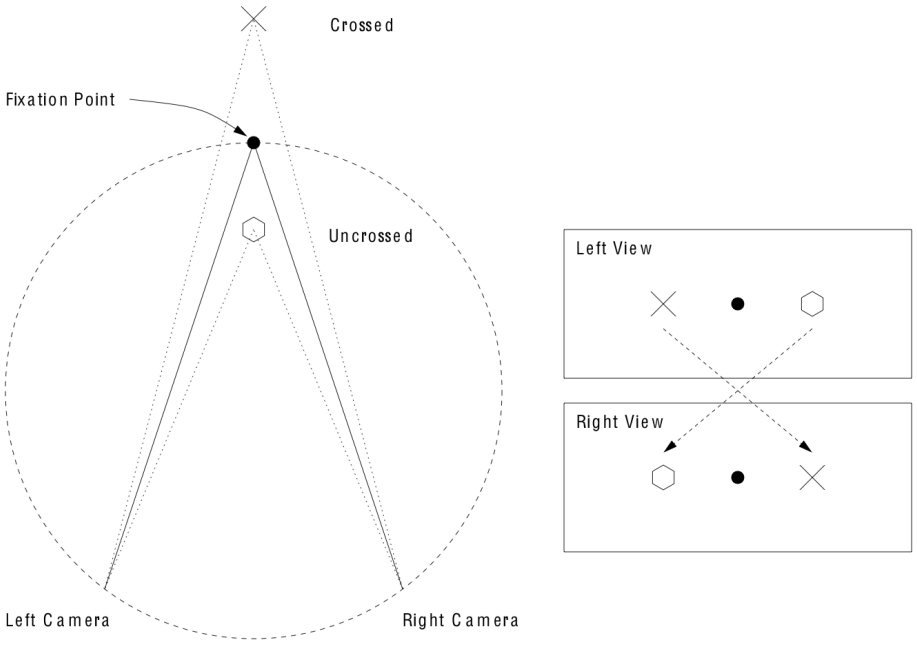



Figure 6.3: The horopter and crossed disparity. A plan view showing the approximate location of the horopter for the given camera geometry (dashed line). Objects in front of (hexagon) and beyond (cross) the fixation point exhibit uncrossed and crossed disparity respectively.

tial problem. For static synthetic systems with a strict fronto-parallel camera configuration, parallel epipolar geometry is enforced and it is then sufficient to analyse corresponding scanlines of both images to estimate disparity.

### 6.1.2 Evidence of Spatial Perception in the Brain

Not all animals are able to detect binocular disparity; an example is the rabbit. When studying cats [Ohzawa *et al.* (1997)] and monkeys [Poggio *et al.* (1988)], scientists presented evidence of neural mechanisms for binocular depth discrimination based on disparity sensitive cells in the visual cortex. All healthy primates exhibit the ability to discriminate disparities. We briefly review spatial perception in the primate brain. We look for evidence of the perception of scene structure and motion.

#### 6.1.2.1 Scene Structure

Although numerous cues provide spatial information, primates mainly interpret scene depth from estimates of binocular disparity. Early visual areas (V1, V2, V3) and ventral areas (hV4, V8, V4) are likely to be involved in processing both absolute and retinal disparities [Lamme *et al.* (2000); Neri *et al.* (2004); Poggio *et al.* (1988)]. Absolute scene depth is likely to be interpreted in dorsal areas (V3A, MT/V5, V7). Gaze convergence, focal length and prior familiarity with an object's size can provide information for conversion from relative to absolute depth distances. Gaze convergence stretches extraocular muscles. Signals originating from kinesthetic sensations in these extraocular muscles are known to be passed to the visual cortex, where they play a role in absolute depth perception [Zajac (1960)].

#### 6.1.2.2 Scene Motion

While the eye is stationary, primates can estimate relative scene velocities with high accuracy, however, during eye movements accuracy reduces. Additionally, when an object moves directly towards or away from an observer there is minimal eye movement occurring. In this instance, the ability to discern absolute and relative speeds is still present via disparity. As described previously, individual neurons in early visual areas (LGN, V1 and V3) respond to motion that occurs

locally within their receptive field. A global perception of motion appears to occur in area MT/V5 in human visual cortex [Kandel *et al.* (2000)].

### 6.1.3 Synthesising Disparity Estimation

The correct and fast estimation of disparity is non-trivial. Sensor noise and different transfer functions of the left and right imaging system introduce stochastic signal variations. Left-right perspective differences lead to a variety of systematic image variations, including occlusion effects and foreshortening. In addition, since most object surfaces in real-world scenes display specular reflection[1], the intensities observed by the imaging systems are not directly correlated with the object surfaces, but nearly always have a viewpoint dependent component which moves independently of the surface in question.

Disparity in stereo image pairs has been computed using area and feature matching techniques that try to counteract the set of distorting signal variations. Features are detected in one image and searched for in the other. Disparity has also been recovered from frequency and phase-based calculations. Other methods include image interleaving and coherence-based detection. All of these methods have their intrinsic problems caused by the various assumptions inherent in their approach. We now review common techniques to recover image disparity.

#### 6.1.3.1 Feature-Based

In this scheme, features such as edges, corners, contours or patches are identified in both images. Intensity information is converted to a set of features assumed to be a more stable image property than raw intensity data. The matching stage operates only on these extracted image features. Of course, only a discrete number of specific feature-classes can be utilised. Therefore a significant area of the image can be identified as containing no matchable features and is not considered further in the matching process. This approach can be fast as features

[1]Specular reflection occurs on glossy or shiny objects where light is reflected from its source without being affected by the surface of the object it is reflecting off. This contrasts with Lambersian reflection where the reflected light is altered by the reflecting surface to give that surface its textual appearance and colour. Most reflections contain both a specular and Lambersian component.

Figure 6.4: Example disparity map. One of the input camera views (left), and a dense disparity map output (right). Distant scene regions are represented by darker intensities.

usually appear in limited numbers, allowing significant data reduction. It is insensitive to lighting conditions or small image deformations. Since distinctive features only form a small part of an image, this method produces sparse disparity estimates as it is impossible to determine the disparity for featureless regions. Interpolation is necessary to populate the missing regions. Also the choice and localisation of features can be difficult since they are strongly related to the image content. Potentially, every feature detected in one image can be matched with every feature of the same class in the second image. This false matching problem can be reduced by the addition of constraints to the solution, such as restricting the search to be only along epipolar lines.

#### 6.1.3.2 Area-Based

This technique uses raw image data and epipolar geometry in binocular image pairs to compute disparity. Image intensity values within small discrete patches of one view are compared to identically sized patches in the same vicinity in the second view. The task is to match patches of maximum correlation and note their displacement. This method can produce dense disparity maps but reliance on image intensities means it is often sensitive to lighting conditions and geometrical deformations.

This scheme can be implemented in many different ways, depending upon

the chosen similarity measure, the algorithmic solution, and on the complexity of the modeled disparity field. Similarity measures used can include: sum of products, covariances, sum of squared differences, sum of absolute differences and cross-correlation. Algorithmic solutions range from complete search to iterative least squares, simplex algorithms and dynamic programming, and can be highly dependent on the *a priori* knowledge of the scene and the similarity measure. This method is also susceptible to the false matching problem of feature-based techniques. To ensure stable performance, area-based algorithms need suitably chosen correlation measures and a sufficiently large patch size. They are hence often computationally expensive.

#### 6.1.3.3 Phase-Based

This approach uses the fact that the disparity from bandpass signals is equivalent to the local phase difference between the signals. From a theoretical point of view, it can be seen as a direct application of the Fourier shift theorem. Using this theorem, the phase shift (a measure of local disparity) between horizontal scanlines in pairs of images can be derived from local frequency and phase. In this manner, Fourier phase images are extracted from the raw intensity data. The Fourier phase may exhibit phase wrap-around, making it necessary to employ hierarchical methods. The Fourier shift theorem cannot be directly applied to images because it is used to determine the overall shift between two signals, whereas pixel displacements in a stereo image pair are fundamentally local. Thus, as mentioned, a coarse-to-fine heirachical method is essential, and increases computation significantly. Although similar to area-based techniques, this method is search-free. Phase is amplitude invariant, so the method is robust to intensity and small image distortions and produces dense disparity maps. It is important that the two images come from calibrated cameras that have had an epipolar constraint applied, so that horizontal lines of pixels coincide.

#### 6.1.3.4 Coherence-Based

In this method, disparity calculations and fusion of a pair of images into a cyclopean view is performed simultaneously. Henkel (1999) utilised a network calculation structure as shown in Figure 6.5. Simple disparity estimators are arranged

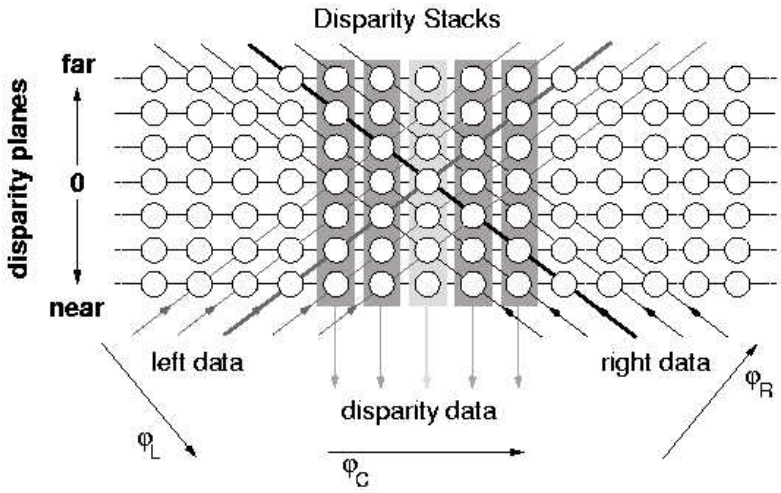



Figure 6.5: Coherence-based disparity stack network [Henkel (1999)].

in horizontal layers which have slightly overlapping working ranges. Image data is fed into the network along the diagonally running data lines. Within each of the vertical disparity stacks, coherently coding sub-populations of disparity units are employed, and the average disparity value of these pools can be read out of the network. This method is an attempt to reproduce the operation of complex cells in the human visual cortex. Though quite successful, it is complex and computationally expensive.

### 6.1.4 Depth From Disparity

Once image disparity has been calculated, camera calibration data and geometry information can relate image disparity directly to absolute scene depths. For cameras whose axes are strictly fronto-parallel, scene depths $Z$ can be computed from disparity $D$ according to $Z = fB/D$.

Few existing methods are tailored to cope with the arbitrary motion of active cameras. To our knowledge, none convert relative disparities to absolute disparities during cameras motion, in real time (motion may originate from deliberate camera movement or perturbations such as moving cameras by hand).

In primates, kinesthetic sensations are known to be involved in converting retinal disparities from varying viewing geometries into an egocentric perception of scene depth. For our active vision framework, encoders provide the equivalent of the kinesthetic feedback signal at the active rectification stage. As described in the previous chapter, active rectification projects images from active cameras into a static fronto-parallel reference frame (the mosaics), where the above equation holds.

## 6.2 Spatial Representation

Initial efforts in computer vision attempted to identify scene structure and objects from features such as lines and vertices in images. Stereo disparity maps are still created from stereo images by identifying patches of object surfaces in multiple views of scenes. Traditionally, somewhat sparse and noisy stereo depth data has been used to judge the existence of surfaces at a location in the scene. Decisions based directly on such unfiltered data could adversely affect the sequence of future events reliant upon such a decision. In previous use of stereo range data, only a few attempts were made to strengthen or attenuate a belief in the location of mass in the scene [Moravec (1996)]. Occupancy grids can be used to accumulate diffuse evidence about the occupancy of a grid of small volumes of space from individual sensor readings and thereby develop increasingly confident and detailed maps of a scene [Elfes (1989)].

As well as addressing the above issues, an occupancy grid allows the integration of data according to a sensor model. As we shall see, each pixel in the disparity map is considered as a single measurement for which a sensor model is used to fuse data into the occupancy grid. Not only is uncertainty in the measurements considered in the sensor model, but it is also partially absorbed by the granularity of the occupancy grid. Bayesian updating of cell occupancy status can be used to integrate sensor data.

### 6.2.1 Occupancy Grids

Occupancy grids were first used in robotics to generate accurate maps from simple, low resolution sonar sensors [Elfes (1989)]. Occupancy grids were used to

accumulate diffuse evidence about the occupancy of a grid of small volumes of nearby space from individual sensor readings and thereby develop increasingly confident and detailed maps of a robot's surroundings. The use of occupancy grids has been applied to range measurements from other sensing modalities static stereo vision on mobile platforms [Murray & Little (2000)], laser and millimeter wave range scanners [L Matthies (1988)].

Representing the scene by a grid of small cells enables us to represent and accumulate the diffuse information from depth data into increasingly confident maps. Belief in any data point can then be related to that point's surroundings. This approach reduces the brittleness of the traditional methods.

The occupancy grid approach represents the robot's environment by a 2D or 3D regular grid. An occupancy grid cell contains a number representing the probability that the corresponding cell of real-world space is occupied, based on sensor measurements. Sensors usually report the distance to the nearest object in a given direction, so range information is used to increase the probabilities in the cells near the indicated object and decrease the probabilities between the sensed object and the sensor. The exact amount of increase or decrease to cells in the vicinity of a ray associated with a disparity map point forms the sensor model.

Combining information about a scene from other sensors with stereo depth data is usually a difficult task. Another strength of the occupancy grid approach is that it facilitates such integration. A Bayesian approach to sensor fusion enables the combination of data, independent of the particular sensor used [Moravec (1989)]. A single occupancy grid can be updated by measurements from sonar, laser or stereoscopic vision range measurements. In this approach, the sensors are able to complement and correct each other, when inferences made by one sensor are combined with others. For example, sonar provides good information about the emptiness of regions, but weaker statements about occupied areas. It can also recover information about featureless areas. Conversly, stereo vision provides good information about textured surfaces in the image.

### 6.2.2 Bayesian Occupancy Grids

We use a Bayesian methods [Moravec (1989)] to integrate sensor data into the occupancy grid. Sensor models are used to incorporate the characteristics of error

for the particular sensor being used.

Let $s[x, y]$ denote occupancy state of cell $[x, y]$. $s[x, y] = occ$ denotes an occupied cell and $s[x, y] = emp$ denotes an empty cell. $P(s[x, y] = occ)$ denotes the probability that cell $[x, y]$ is occupied. $P(s[x, y] = emp)$ denotes the probability that cell $[x, y]$ is empty.

Given some measurement $M$, we use the incremental form of Bayes Law to update the occupancy grid probabilities [Elfes (1989)]:

$$\begin{aligned} P(occ)_{k+1} &= \frac{P(M \mid occ)}{P(M)} P(occ)_k \\ P(emp)_{k+1} &= \frac{P(M \mid emp)}{P(M)} P(emp)_k \end{aligned} \tag{6.1}$$

where $emp$ denotes $s[x, y] = emp$, $occ$ denotes $s[x, y] = occ$, and

$$\begin{aligned} P(M) &= P(M \mid occ)P(occ) \\ &+ P(M \mid emp)P(emp) \end{aligned} \tag{6.2}$$

#### 6.2.2.1 Sensor Models

Let $r$ denote the range returned by the sensor and $d[x, y]$ denote the distance between the sensor and the cell at $[x, y]$. For a *real* sensor, we must consider Kolmogoroff's theorem [Moravec (1989)] where localisation due to a measurement produces a continuous PDF (left, Figure 6.6). For an *ideal* sensor (right, Figure 6.6), we have:

$$\begin{aligned} P(r \mid occ) &= \begin{cases} 0 & \text{if} \quad r < d[x, y] \\ 1 & \text{if} \quad r = d[x, y] \\ 0.5 & \text{if} \quad r > d[x, y] \end{cases} \\ P(r \mid emp) &= 0 \end{aligned} \tag{6.3}$$

We adopt the 1D ideal sensor model for integrating data into the occupancy grid. In this case, the occupancy of the cell that a measurement corresponds to is incremented. The occupancy of cells in front of this cell are decremented, and the cells behind it are tended towards ambient levels using Bayesian updating as follows.

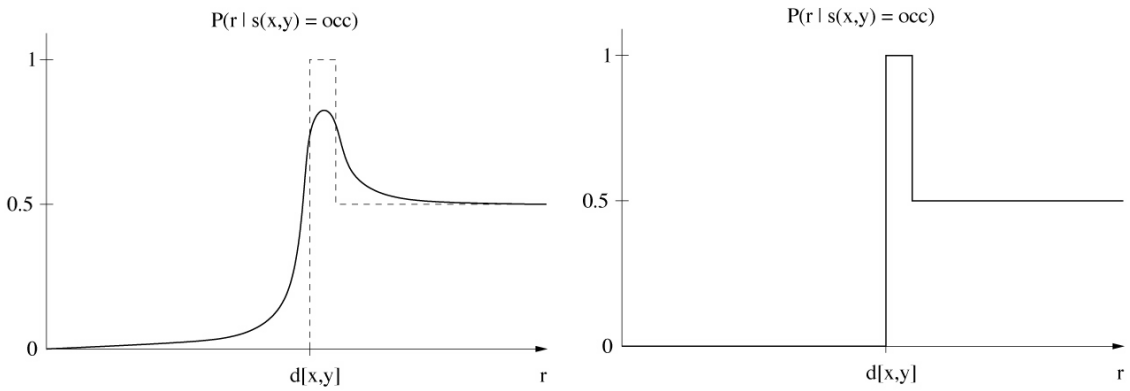



Figure 6.6: Example sensor profiles. The 1D profile of a real sensor (left), and an ideal sensor model (right) [Moravec (1989)].

#### 6.2.2.2 Updating the Occupancy Grid

We can re-write Equation 6.1:

$$\frac{P(occ)}{P(emp)} \leftarrow \frac{P(M \mid occ)}{P(M \mid emp)} \frac{P(occ)}{P(emp)} \tag{6.4}$$

In terms of likelihoods this becomes:

$$L(occ) \leftarrow L(M \mid occ) L(occ) \tag{6.5}$$

Taking the log of both sides:

$$\log L(occ) \leftarrow \log L(M \mid occ) + \log L(occ) \tag{6.6}$$

Log-likelihoods thus provide a more efficient implementation for incorporating new data into the occupancy grid by reducing the update to an addition [Elfes (1989)].

## 6.3 An Occupancy Grid for Active Vision

We now develop a 3D occupancy grid specifically designed for the integration of active vision data into an egocentric 3D representation of scene structure and motion.

### 6.3.1 A Space Variant Occupancy Grid Representation of the Scene

The occupancy grid is constructed such that the size of a cell at any depth corresponds to a constant amount of pixels of disparity at that depth. It is also constructed such that rays emanating from the origin pass through each layer of the occupancy grid in the depth direction at the same X, Y coordinates.

The width and height of the occupancy grid correspond to the size of the full mosaic, as defined in the rectification process. At farther scene depths, pixel disparities correspond to larger changes in scene depth. For example, 10 pixels of disparity for an object at 1m scene depth corresponds to a much smaller depth variation than does 10 pixels of disparity for an object at 50m scene depth. Accordingly, the occupancy grid configuration exhibits cell cube sizes that increase with depth. Cell cube sizes are defined by their x, y and z edge lengths. The z-length of a cell cube in the occupancy grid corresponds to the effect of a specific amount of pixels of disparity (n) at the depth the cell exists.

The camera images are projected into the mosaic reference frame where parallel epipolar geometry has been enforced (synthesising a fronto-parallel arrangement). A hypothetical *absolute* disparity $D$ at mosaic coordinates $(u, v)$ can therefore be mapped to 3D world coordinates according to:

$$\begin{aligned} Z &= \frac{fB}{D} \\ X &= \frac{uZ}{f} \\ Y &= \frac{vZ}{f}, \end{aligned} \tag{6.7}$$

where $B$ is the length of the baseline and $f$ is the focal length of the cameras (assumed constant and equal).

The space-variant active vision occupancy grid (Figure 6.8) is constructed and subdivided into cells according to the algorithmic summary in Figure 6.7. The space-variant approach significantly reduces the number of cells at larger depths where high depth resolution is not usually available anyway, improving processor performance. It also increases resolution in the grid at nearer depths where we are more interested in an accurate estimation of the location of objects.

The space variant occupancy grid formulation reduces ray tracing computations associated with sensor model integration of range measurements. At 1m

**Active Vision Space-Variant Occupancy Grid Construction**
For each pair of camera images:

1. Select the minimum and maximum distances from the active head origin that the occupancy grid will represent, $Z_{min}$ and $Z_{max}$.

2. Set the width and height of the occupancy grid $(W, H)$ to the mosaic width and height used in the rectification process, so that the same visual space is represented.

3. Select the cell cube edge length $n$ in terms of pixels; it should be a factor of the width and height.

4. The first slice (Z=0) of the occupancy grid is drawn as follows:
   a) transfer the corners of the mosaic to 3D coordinates. That is, find X and Y from Equation 6.7 using $Z = Z_{min}$ for corners $(0,0), (0,H), (W,0), (W,H)$.
   b) subdivide the face into a grid of $W/n$ squares in the x-direction and $H/n$ squares in the y-direction.
   c) set the z-length $z_{Z=0}$ of the cubes in this slice to the same as its x and y lengths.

5. The next slice is drawn as follows:
   a) increment slice reference Z.
   b) transfer the corners of the mosaic to 3D coordinates. That is, find X and Y from Equation 6.7 using $Z = Z_{min} + z_{Z-1}$ for corners $(0,0), (0,H), (W,0), (W,H)$.
   c) subdivide the face into a grid of $W/n$ squares in the x-direction and $H/n$ squares in the y-direction.
   d) set the z-length $z_Z$ of the cubes in this slice to the same as its x and y lengths.

6. Repeat step 5 until occupancy grid z-size exceeds $Z_{max}$.

Figure 6.7: Summary: Active vision occupancy grid construction procedure.

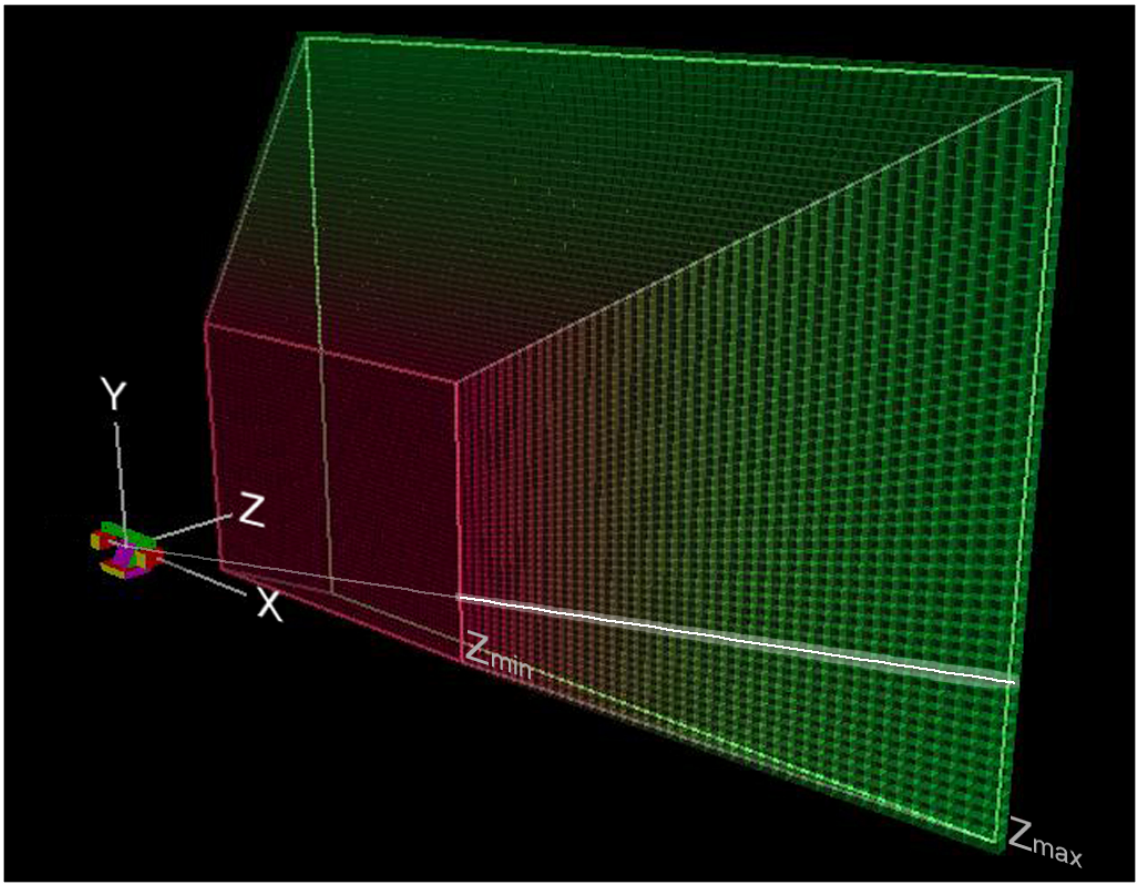



Figure 6.8: Occupancy grid configuration. The active head is shown at the origin. The X, Y and Z cell directions are defined. A ray projected back onto the occupancy grid (white highlighted cells) passes through all slices of the occupancy grid in the Z direction at identical X, Y cell coordinates.

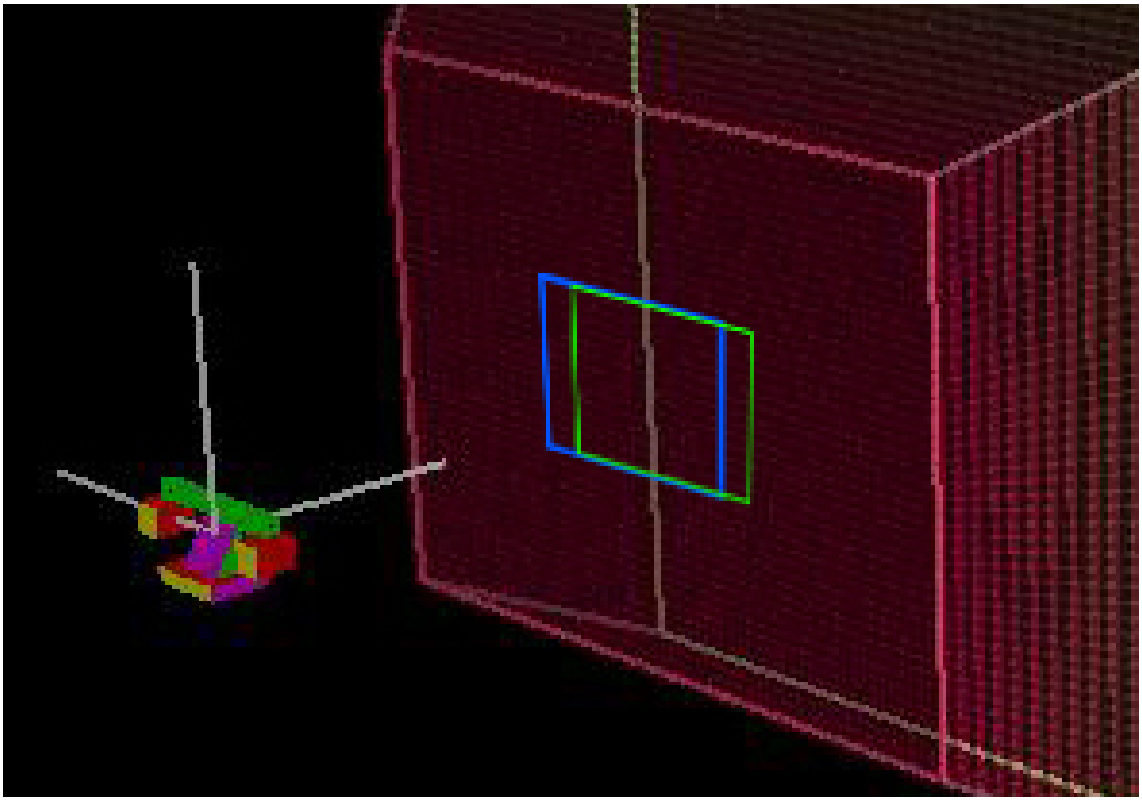


Figure 6.9: Occupancy grid showing the re-projection of the left (blue) and right (green) camera frames back onto the front face of the occupancy grid. The size of all slices through the occupancy grid in the z direction corresponds to the mosaic size set in the rectification stage.

depth, a slice through the occupancy grid contains a fixed number of cells in the horizontal direction and another fixed number of cells in the vertical direction. At any other depth, the number of vertical and horizontal cells in the slice are the same respectively, with the central cells aligned with the origin at the location of the sensor. This means that a ray emanating from the origin and passing through a cell at 1m with slice coordinates (x, y) also passes through all other slices at coordinates (x, y). This configuration means that ray-tracing through the occupancy grid for sensor integration becomes trivial. Figure 6.8 shows the construction of the occupancy grid and rays of cells emanating from the origin.

### 6.3.2 Populating the Occupancy Grid

Active rectification provides epipolar rectified images and the convergence disparity parameter $d$. The overlapping regions of the rectified left and right images can then be used as input for estimation of horizontal disparities. In this manner, a relative horizontal disparity map is obtained using the rectified images as input. The relative disparities can then be converted to absolute (mosaic reference frame) disparities by simply adding convergence disparity parameter $d$ to all disparities. Each pixel in the resulting map can be converted to an absolute 3D scene location using Equation 6.7. The cell that each estimate from each disparity entry corresponds to in the occupancy grid is thus determined.

A sensor model is used to fuse disparity data from the active cameras into the 3D occupancy grid. Each pixel in the absolute disparity map can be thought of as a single measurement along a ray emanating from the origin located midway between the baseline connecting the camera centres. For processor economy we use the ideal sensor model in Figure 6.6. Applying this simple sensor model involves increasing the occupancy of the cell that the pixel corresponds to, and decreasing the likelihood of all cells along the occupancy grid ray in front of that cell. Occupancy of all cells behind the occupied cell are tended towards the ambient level (or the value that corresponds to "don't know"). Thus ray tracing in the occupancy grid is trivial, because rays pass through all cells with identical X, Y coordinates. For example, a ray passing through layer Z=10 at cell X=5, Y=5 also passes through all other Z layers at X=5, Y=5.

We combine all disparity matches in the disparity image into the occupancy grid by applying the sensor model. Figure 6.12 shows an example of an occupancy grid pupulated by this process. For a given camera geometry, the locations of the left and right rectified images within the mosaic defines the area that may be disparity-mapped. Figure 6.10 shows the calculation of the disparity-measurable area.

The limits of the measurable volume (defined by eight vertices) for a given camera geometry and disparity search range can be found by using Equation 6.7 to project a disparity of $\pm 16$ at each of the corners of the output area of the disparity map onto the occupancy grid. That is, set $D = d + 16, d - 16$ (where $d$ is the rectification convergence parameter and the disparity search is over $\pm 16$

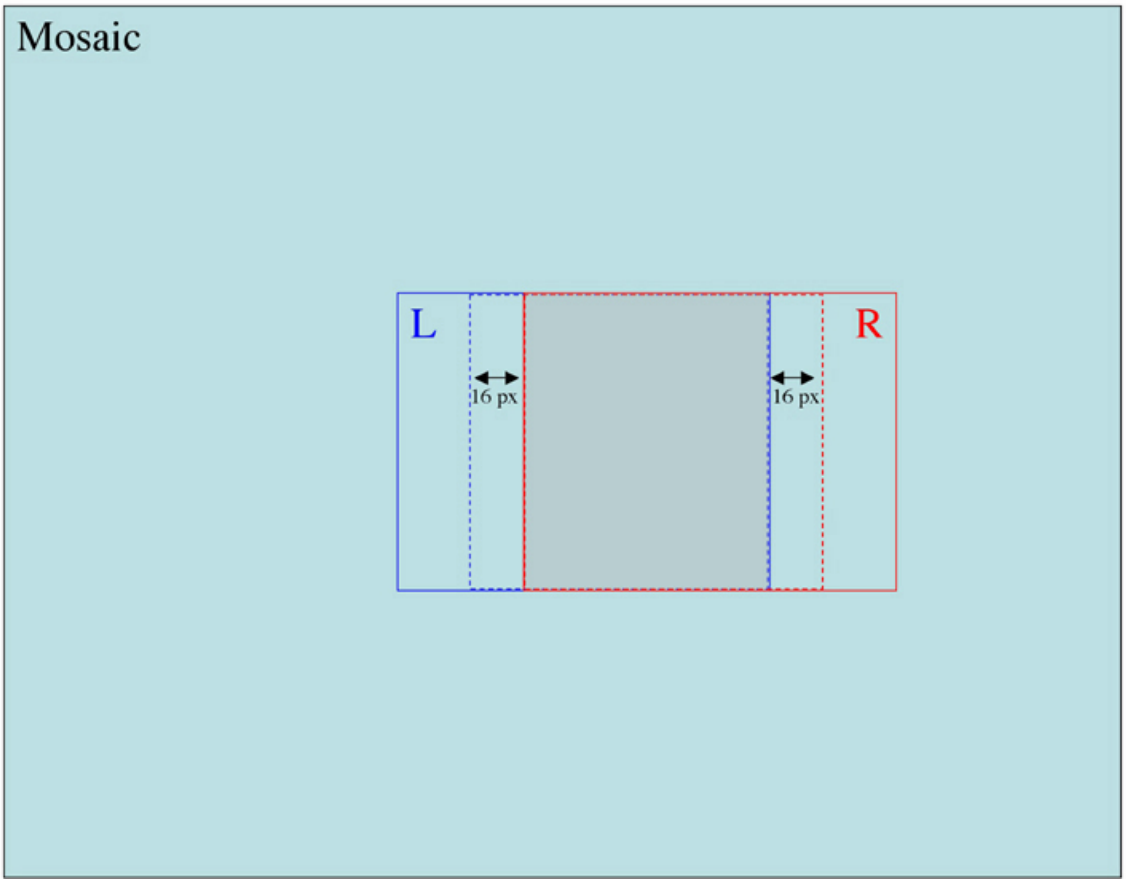



Figure 6.10: Disparity estimation coverage in mosaic space. At any instant in time, the locations of the left and right rectified images within the mosaic defines the area that may be disparity-mapped. In this example, disparities are searched for over a range of ±16 pixels, which means we may find matches up to 16 pixels outside the overlapping region. Therefore the input images for disparity search are defined by the areas surrounded by the blue and red dotted lines. The size of the output disparity map is the size of the shaded area of left-right overlap. The position of the output area in the mosaic, the camera geometry, and the disparity search range define the measurable volume in the occupancy grid (Figure 6.12).

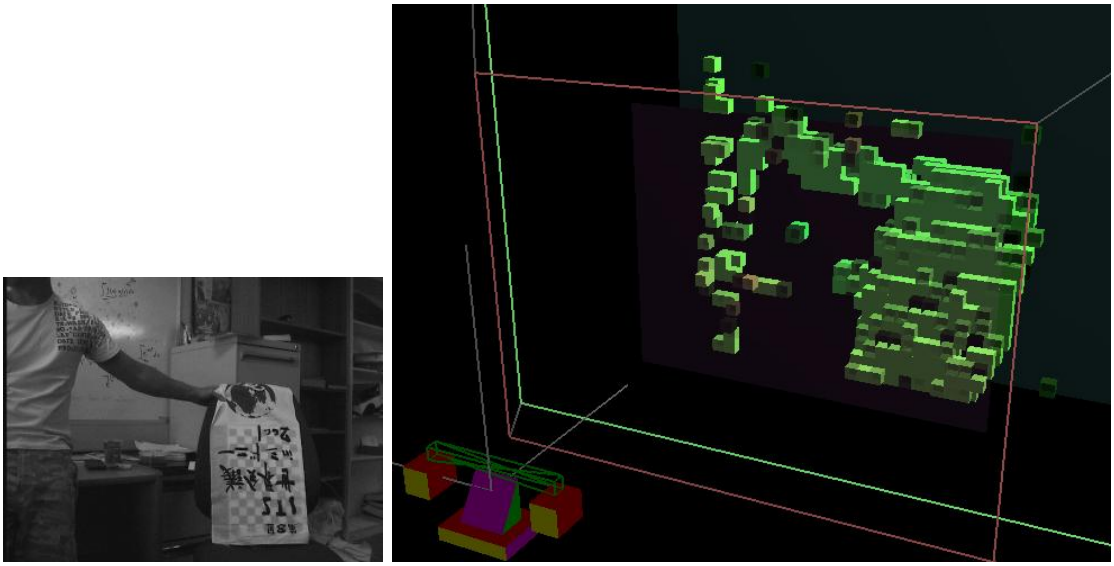


Figure 6.11: Online snapshot of raw occupancy grid contents. The rectified left camera image (left), and occupancy grid (right) are shown. The two semi-transparent vertical planes in the occupancy grid show the near and far bounds of the region in which depth measurement is possible, given the current camera geometry and disparity search range (±16 pixels).

pixels) and set $(u, v)$ as each of the four corners of the left-right overlap area in mosaic coordinates. Figure 6.12 shows the online result of this process - near and far planes showing the limits of the measured volume defined by the eight resulting vertices.

The simplicity in incorporating data into the structure enables us to construct an occupancy grid model of the relevant volume of the scene by scanning the horopter over it. We do not just obtain an instantaneous impression of the region of the scene for which we presently have a depth map, instead we accumulate evidence about each cell in the occupancy grid. We are able to accumulate information about occupied cells and retain a memory of where mass was previously observed in the scene, even if we are not viewing that region of the scene anymore.

We may define a task-oriented occupancy grid volume and resolution. For example, in the laboratory or for object manipulation, the selected occupancy grid volume is small, and so are cell sizes. For high-speed outdoor navigation,

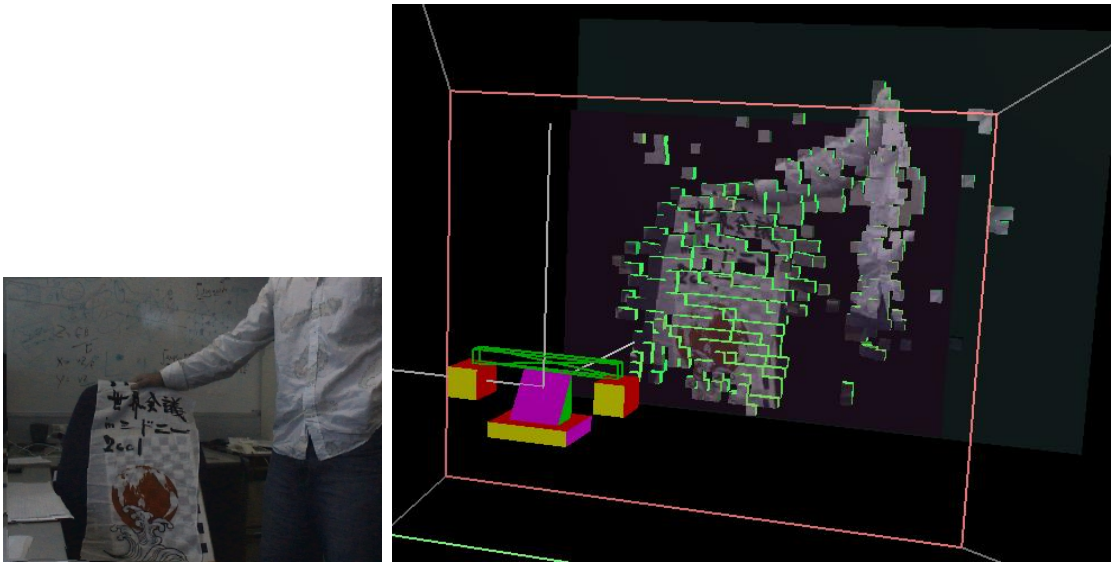


Figure 6.12: Online snapshot showing extent of depth-measurable volume. The rectified left camera image (left), and the measurable volume for the current camera geometry and a disparity search range of ±16 (right) are shown . The left-right overlap (see Figure 6.10) of the currently viewed regions in the left and right mosaics sets the width of the disparity image and measurable volume. The disparity search range sets the depth. The vertical purple and green planes show the front and rear limits of the measurable volume. The wireframe shows the limits of the entire occupancy grid. We have enabled online re-projection of the scene onto the occupancy grid for ease of interpretation.

the desirable sensing volume is large, and at distance so too are cell sizes. We only update the cells in the occupancy grid that represent the region of the scene relevant to our task-oriented behaviors. Information about the scene that falls outside this bound is suppressed, including data from depth map images that falls beyond the defined occupancy grid volume.

We may choose a threshold above which cells are considered occupied at any point in time. Figures 6.11 and 6.13 show example occupancy grid output. A demonstration movie showing online occupancy grid population is shown in Figure 6.13. The method to update the contents of the space-variant occupancy grid is summarised in Figure 6.14.

### 6.3.3 Dealing with Dynamics

Updates to the occupancy grid occur at a frequency high enough for us to effectively analyze dynamic scenes. So that previously occupied cells do not remain flagged as occupied once a moving object has moved away, we incorporate the use of a decay rate applied to all cells. We decay the occupancy of all cells towards ambient levels over time. This of course means that cell occupancies may "linger" once an object has moved away until the occupancy decays to ambient levels. It also instantiates a trade off between accumulated confidences and the ability to accumulate occupancy in cells through which a moving object passes.

A high decay rate should be used where linger is undesirable. A lower decay rate should be used where more confidence about static scene regions is desirable. The decay rate may vary between these extremes automatically, according to preference associated with the desired task. The alternative is to separately consider evidence from each disparity map over time, resetting the entire occupancy grid every time a disparity map is obtained. For coarse occupancy grids this is a reasonable solution because occupancy evidence can be accumulated over the correspondingly larger $n$ by $n$ image areas, rather than time. For finer occupancy grids where the cells correspond to small $n$ by $n$ image regions, there is more of a reliance on accumulating evidence over time rather than image area. In this case, such a solution is not ideal because evidence may be too scarce on a frame by frame basis to render enough cells as *occupied* for an accurate scene representation. Resetting the grid is equivalent to a very high decay rate.

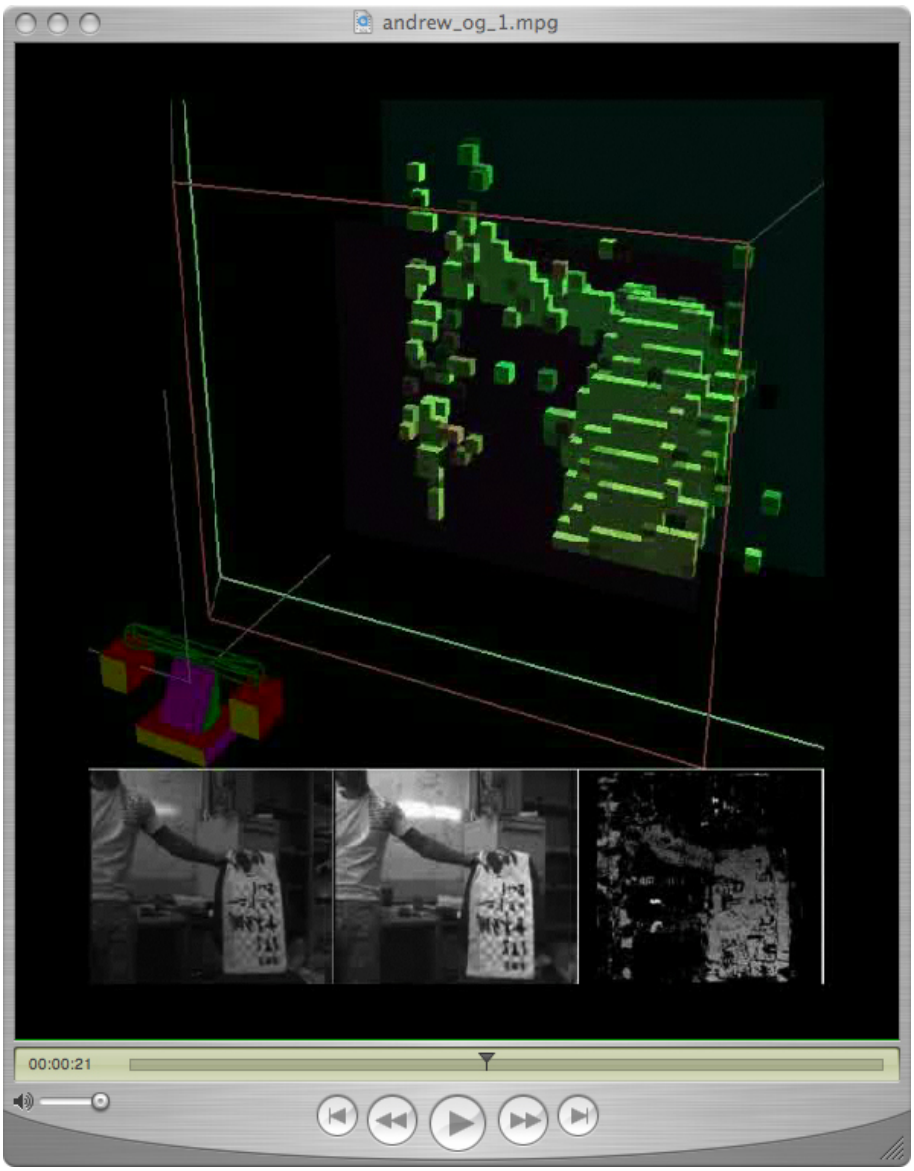



Figure 6.13: Online population of occupancy grid demonstration (*snapshot - see Appendix C for full video*).

**Active Vision Space-Variant Occupancy Grid Update**
For each pair of camera images:

1. Active rectification provides epipolar rectified images and convergence disparity parameter $d$.
2. Obtain relative disparity map from rectified images.
3. Convert to absolute disparities by simply adding $d$ to all disparities.
4. Convert each pixel to an absolute scene location using Equation 6.7.
5. Find corresponding occupancy grid cell for each pixel.
6. For each pixel, apply ideal sensor model along ray passing through the occupancy grid cell using Bayesian update. The ray passes through all occupancy grid cells at same x, y coordinates.
7. Tend occupancy of all cells in occupancy grid towards ambient level.

Repeat from 1.

Figure 6.14: Summary: active vision occupancy grid update procedure.

### 6.3.4 Dealing with Error

Errors associated with extrinsic parameter estimation during rectification will affect the accurate construction of an occupancy grid from each pair of stereo images. Bayesian integration of many such occupancy grids over time and from many different viewing geometries will reduce the effect of inaccurate extrinsic parameter measurement. The Bayesian approach means we assume the error in the estimates approximates zero mean Guassian error. Active rectification calibration inaccuracies may mean that noise is not zero mean but systematic, which may affect conversion from relative to absolute occupancy grid depths. The granularity of the occupancy grid can help to absorb such error, and despite such systematic calibration errors, or other sources such as the error due to variations in focal lengths, the occupancy grid preserves relative locations of surfaces in the egocentric representation.

The occupancy of a cell is effectively accumulated over disparity pixels in a $n$ by $n$ sized square in the absolute disparity image. Depending on the threshold selected above which cells are considered occupied, a cell will require numerous "hits" before it can be considered occupied. Using the described sensor model to integrate data, this condition would require multiple disparity pixels of the same value in the same $n$ by $n$ square, and few hits in other cells in the same occupancy grid ray. In this manner, Bayesian updating helps to identify the cell in each ray that is most likely to be occupied, if any.

If numerous hits come from a small object it may render a cell occupied. We cannot identify where within the cell the object is likely to be by looking at the occupied cell alone - reconsideration of the original disparity maps would be required. For the purpose of rapid spatial perception, it suffices to say that there is likely to be a surface within the cell, and that there is free space in front of that cell - which is nonetheless useful information for obstacle avoidance and navigation, and may be the basis of further exploration. It may be the case that a small object renders a cell occupied yet it is small enough that objects behind it in the real scene render a second more distant cell in the occupancy grid ray occupied. Alternatively, using the ideal sensor model, a strong response at a distant cell along a ray could reduce the response of such a small but closer object below the occupancy threshold. If it is a priority to detect *all* cells that

may be occupied, rather than the *most likely* cell in a ray that is occupied, it may be more beneficial to use cell accumulation rather than Bayesian sensor models in incorporating disparity data into the occupancy grid. Unlike the sensor model, this would not tend occupancy towards a single dominant cell in each ray. In our laboratory experiments, we are usually surrounded by walls and we are interested in scanning the broad structure of such surroundings rather than detecting and avoiding tiny objects, so we choose the ideal sensor model for data integration.

Aside from fronto-parallel calibration in the rectification step, good performance depends mainly upon the accuracy and density of disparity maps. We use an area-based SAD disparity estimation algorithm, but any disparity estimation algorithm may be used instead.

### 6.3.5 Performance

The regime is biased towards a coarse real-time perception, rather than accuracy. It operates continually, over the entire field of view. In this sense it may be likened to a *peripheral* response. Depth maps are produced using a processor economical area-based technique via a SAD correlation metric. Difference-of-Gaussian pre-filtering is incorporated to reduce the effect of intensity variation [Banks & Corke (1991)]. Rectification, pre-filtering, disparity mapping, occupancy grid management, display and logging were achieved at 18Hz on a single processor hyper-threaded 3.0GHz PC using only 48% average CPU load[1]. Hardware support for rendering display is achieved using OpenGL function calls. Memory usage involves storage of $M = W/n \cdot H/n \cdot Z$ integers representing the log likelihood occupancies of all grid cells, plus minor overheads. The use of an occupancy grid representation of spatial information as a component of a larger system does not usually involve rectification, display or data logging at the same processing node. Information distributed to the wider system would usually only include the $M$ bytes of occupancy data, or where the ideal sensor model is used, only $W/n \cdot H/n$ bytes of data containing only the Z-coordinate of the first occupied cell along each occupancy grid ray.

[1] Significant idle time due to network latency in image acquisition causing a largely idle image acquisition thread. Stand-alone implementation for demonstration purposes only. Processing network version incorporates optimised image acquisition.

## 6.4 Use of Occupancy Grid in Synthetic Perception

We now extend upon the presented occupancy grid framework. Once an occupancy grid is constructed and populated, it can be used for coarse spatial awareness in navigation, mapping, obstacle detection and obstacle avoidance. By re-projecting stimulus from which the spatial perception originated back onto the occupancy grid itself, approximate 3D cue-surface correspondences are cheaply computed. By projecting 2D cues such as optic flow back onto the occupancy grid, a visualisation of 3D scene flow may be obtained. We now look at cue-surface correspondences, and the extraction of coarse 3D information using the occupancy grid, such as scene motion and object segmentation.

### 6.4.1 Cue-Surface Correspondence

Projecting the rectified image back onto the occupancy grid is straightforward because the corners of the image can be projected onto the front face of the occupancy grid according to Equation 6.7. A cell at coordinates $(x, y)$ in the front slice corresponds to the same sized image area $(W/n \cdot H/n)$ as the cells at coordinates $(x, y)$ in any other slice in the z-direction. It is merely a matter of scaling the projection that would exist on the front cell at $(x, y, z_{min})$ to the face size of the first occupied cell along the ray of cells corresponding to that region of the image $(z, y, z_{occ})$. Simple re-projection by scaling is made possible by the space-variant configuration of the occupancy grid. Figure 6.12 shows re-projection of the image back onto the occupancy grid.

In this manner, a perception of where surfaces are, and how they appear can be obtained. Re-projecting the original image back onto the occupancy grid is not the only way to use the occupancy grid in scene perception. We may also project *any* cue map back onto the occupancy grid. For example, if edge detection is computed on the rectified camera images, we may re-project the edge map (it has the same frame as the rectified image) onto the occupancy grid to obtain a coarse perception of 3D edge structure. A 3D perception of cue-surface correspondences for any number of cues can be maintained with a low bandwidth representation by keeping only the 2D cue maps and the contents of the occupancy grid in memory.

### 6.4.2 3D Scene Motion

An important case of cue-surface correspondence using the occupancy grid is the perception of scene motion. Few methods calculate absolute 3D scene flow in a head-centred coordinate frame. Even fewer (perhaps none) account for the effect of deliberate camera motions on the perception of flow. Most methods deal with retinal flow, rather than absolute scene flow. The method of Kagami, though not tailored for active vision or an egocentric perception [Kagami *et al.* (2000)], seems most promising for real-time 3D scene motion estimation.

As with disparity estimation, there are various ways to calculate image frame optic flow from multiple camera views of a scene. The main criterion for the selection of a suitable synthetic method is real-time performance. For real-time performance, we choose an area-based SAD method that searches up to 4 pixels of optic flow between frames, outputting an estimate of the $x$ and $y$ components of optic flow at each pixel location in the image. Again, we operate on intensity normalised DOG images. Confidence in calculations therefore depends a lot on texture. As described, we work in mosaic space so that the effect of camera rotations is accounted for.

From consecutive left and right rectified DOG images in mosaic space we obtain X and Y component optic flow maps $fx_l, fx_r, fy_l, fy_r$. As the location of the current and previous frame in the mosaic from a single camera is known, we calculate optical flow on the overlapping region of consecutive frames in the mosaic. In the same manner, the overlapping regions of consecutive depth maps are subtracted and a depthflow map $fd$ is obtained. Equation 6.7 is used to convert all five maps from pixel flows to absolute scene flows. The cues can then be projected onto the occupancy grid. In this manner, multiple estimates of the X, Y and depth flow components of each occupancy grid cell can be obtained. The flow maps provide the x, y and depth components of flow and the occupancy grid cell location grounds the vector to an approximate spatial location. For display purposes, we have averaged each of the flow components at each occupied occupancy grid cell (the x-component at an occupancy grid cell is obtained by averaging all $fx_l$ and $fx_r$ pixels that project to that occupancy grid cell, the y-component at a cell is obtained by averaging all $fy_l$ and $fy_r$ pixels that project to that occupancy grid cell, and the depth component is obtained by averaging

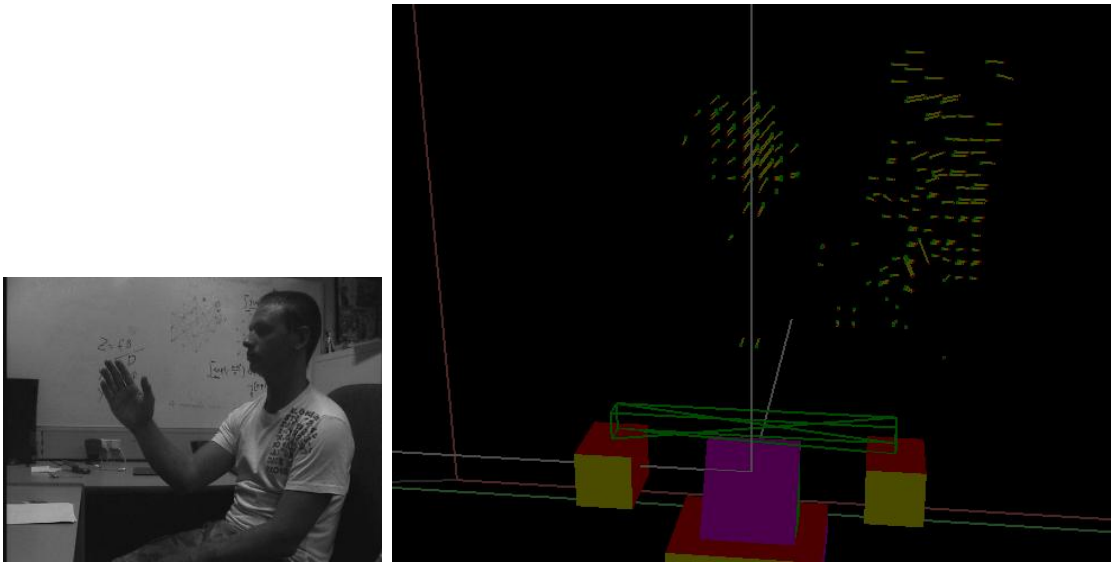


Figure 6.15: Occupancy grid vectors representing 3D motion of visual surfaces in the scene.

all $fd$ pixels that project to that occupancy grid cell). In this manner we are able to assign sub-cell sized 3D motion vectors to the occupied cells in the occupancy grid. We only consider an occupancy grid over a finite region of the scene. Some flow may be detected along a ray which does not have a correspondingly occupied grid cell. Occupied cells that are not assigned a velocity from the flow calculations are assigned a zero velocity.

Figure 6.15 shows online output. In the depicted example, a single computer was used for all processing including rectification, occupancy grid operations, and display. The process operates at approximately 17Hz on a single computer with 40% CPU idle time[1].

Using this technique, a (coarse) perception of where surfaces are, how they are moving, and how they appear visually can be obtained. A demonstration movie showing online cell flow estimation is available as shown in Figure 6.16.

[1]Idle time due to un-optimised network latency upon request of images from cameras - example was for demonstration purposes only.

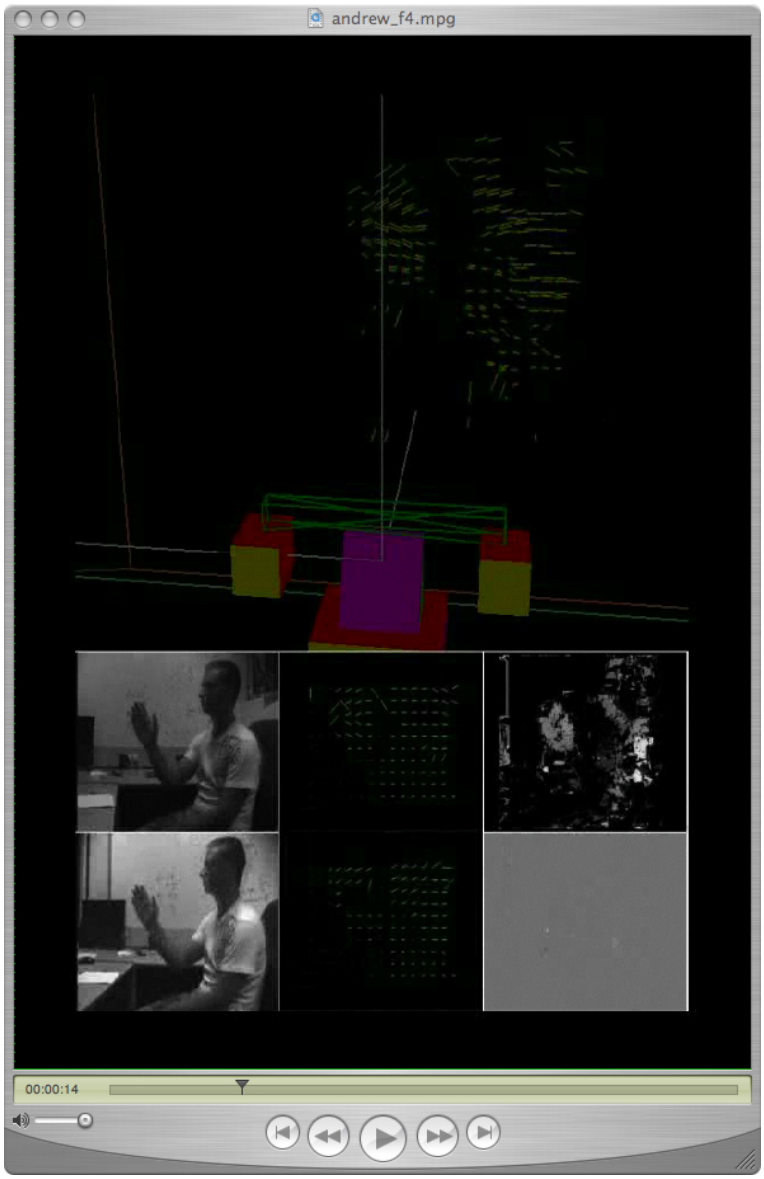



Figure 6.16: Online estimation of 3D flow of occupancy grid cells (*snapshot - see Appendix C for full video*).

### 6.4.3 Ground Plane Extraction from the Occupancy Grid

Ground plane extraction directly from the occupancy grid is similar to that of a *v-disparity* analysis [Labayrade *et al.* (2002)]. A 2D histogram plotting the number of occupied cells in each row X at its Y and Z cell coordinates is constructed (essentially a side-on density projection of the occupancy grid). A Hough transform [Tian & Shah (1997)] is then applied to find the most dominant line in this density side view of the occupancy grid. We search for the line within reasonable bounds of where the road is likely to be to reduce the computational expense of the Hough transform. In this manner, we are able to extract a planar approximation to the location of the ground plane in terms of altitude and attitude. We assume the sensor is situated at a roll angle that is parallel to the road, and that the road is planar. We do not consider any non-zero roll angle of the road relative to the sensor. The granularity of the occupancy grid is such that small violations of this assumption are absorbed. Any systematic misalignment can be removed by calibration. Figure 6.17 shows an image from the online output of the occupancy grid, including the location of the ground plane. A demonstration movie showing online ground plane estimation is available as shown in Figure 6.18.

#### 6.4.3.1 Ego-motion from Ground Plane Motion

We wish to infer the motion of the vehicle relative to the road from an analysis of the flow grid. Preferrably, the analysis would not consider regions of the scene that are likely to be moving in a manner dissimilar to that of the road. Hence, we only consider regions in the vicinity of the ground plane to extract the vehicle velocity. Histograms of the velocity components of all the cells adjacent to the previously detected ground plane are constructed. At present, we use the histogram mean and associated 95% confidence interval as a measurement of the vehicle velocity. Once the velocity of the vehicle relative to the road has been calculated, we can remove the velocity of the vehicle from calculations of the velocity of objects in the scene.

Figure 6.19 shows a plot of the vehicle velocity as determined by unfiltered 3D flow data. Only the flow in the z-direction (directly towards the cameras) is considered. Although the velocity of the vehicle was not logged, the fluctuation of the velocity about a value of approximately 30km/h fits well with the fact that

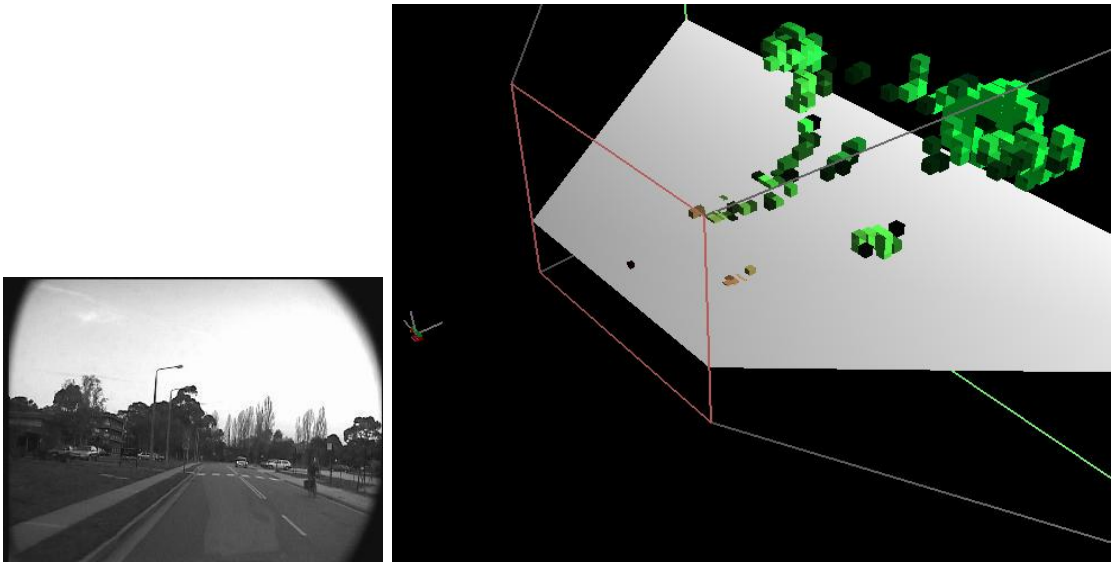


Figure 6.17: Ground plane extraction using occupancy grid. The inset shows the view from the left camera. Detection of the cyclist, light pole, and trees in the background are also evident on the occupancy grid.

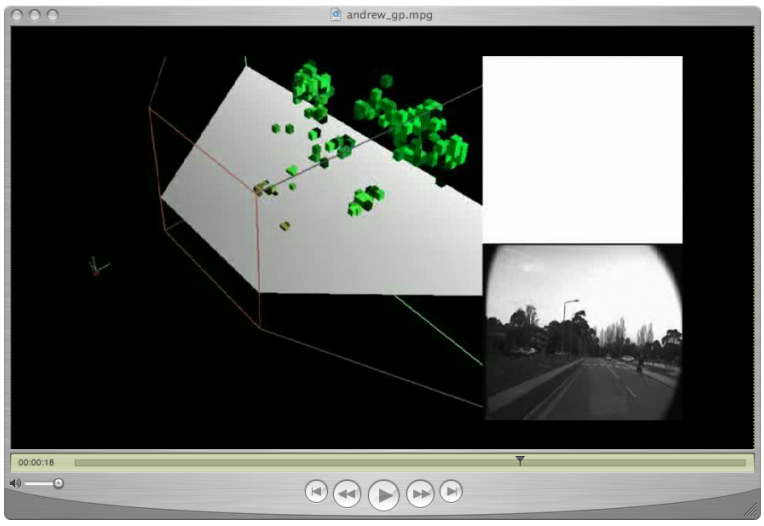



Figure 6.18: Online ground plane detection from occupancy grid demonstration (*snapshot - see Appendix C for full video*).

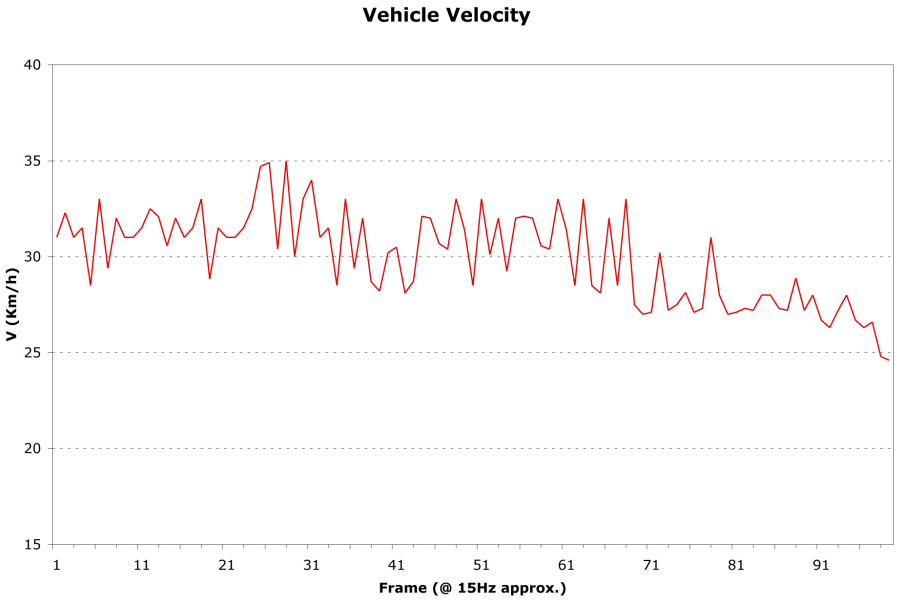



Figure 6.19: Vehicle velocity according to unfiltered 3D flow data.

the vehicle was decelerating in a designated 40km/h zone on the ANU campus. The data velocity from flow was determined for the same sequence of footage as that shown in Figure 6.17.

### 6.4.4 Object Segmentation from the Occupancy Grid

An *object* in the occupancy grid is considered to be a group of *26-connected* (neighbouring) cells located above the ground plane. After an occupancy grid, and cell flows have been calculated, we can segment objects in the flow grid in a manner similar to that of the occupancy grid. A 3D raster scan labels adjacent 26-connected cells whose velocities are similar. Where available, we use information about the location of the ground plane from the previous step to limit the search for objects to the region above the ground plane. Essentially, if a cell has a flow estimate assigned to it, and its velocity is not significantly different to that of an adjacent cell with an estimated velocity, it is assigned the same unique object identity as that cell. The use of velocity information enables us to distinguish, for example, a hand from a chair, despite them being labelled as the same object in the occupancy grid segmentation (Figure 6.20).

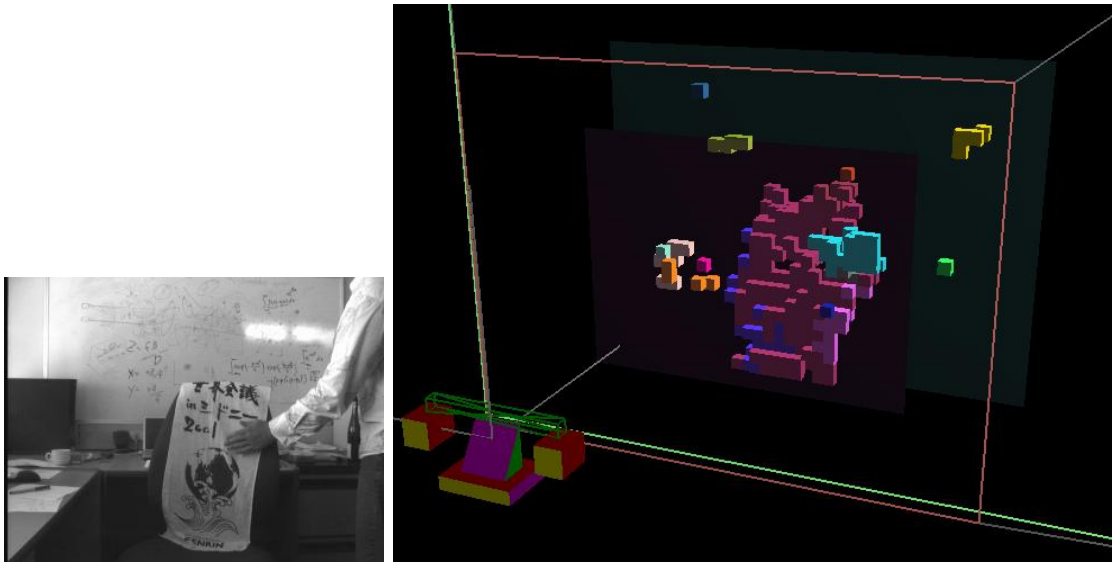


Figure 6.20: Object segmentation using occupancy grid. The hand (blue) is touching the chair (burgundy), but is segmented as a separate object because it is moving.

### 6.4.5 Tracking Objects in the Occupancy Grid

Tracking an object in occupancy grid space involves finding object correspondences in consecutive frames. Data associated with each object in the occupancy grid includes its volume and centre of gravity. Cell flow information also provides velocity information for adjacent occupied cells that form an object. By considering each object in the current frame and comparing the location of the centre of gravity and average velocity of all adjacent object cells with objects in the previous frame, it is possible to estimate likely object correspondences over time. Objects are considered to correspond if the Malhonobis distance between their volume, centre of gravity and velocity (the combined average velocity of all joined cells that constitute the object) is below a threshold.

A demonstration movie showing online object segmentation in the occupancy grid using volume, centre of gravity Malhonobis distance only is shown in Figure 6.21. Different colours represent different objects. If an object has a small Malhonobis distance across consecutive frames, it is considered a correspondence and its colour is preserved. If this correlation is lost between consecutive frames, the object will be assigned a different colour. A demonstration movie showing

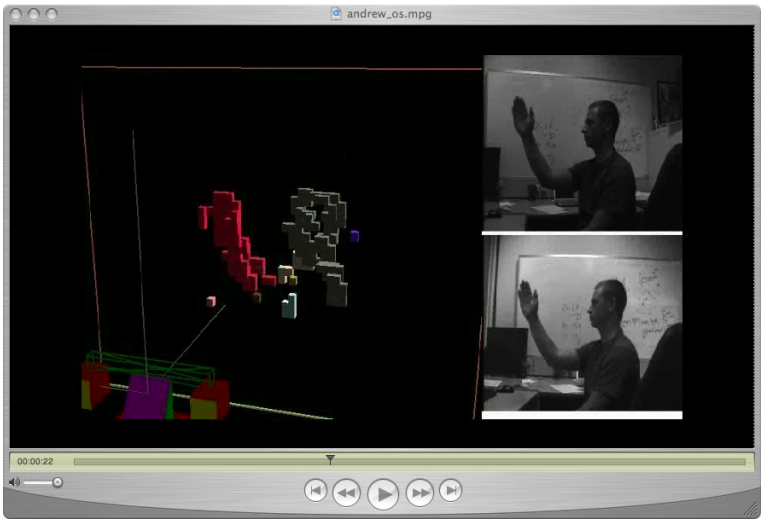



Figure 6.21: Online object segmentation using occupancy grid demonstration (*snapshot - see Appendix C for full video*).

online object segmentation and correspondence across consecutive frames in the occupancy grid using volume, centre of gravity and velocity in the Malhonobis distance measurement is shown in Figure 6.22.

## 6.5 Summary

We have shown that animals perceive scene depths using, amongst other cues, retinal disparity. We have discussed various existing methods to synthesise the computation of retinal disparity. We have provided evidence for the existence of spatial perception in the primate brain. Many brain areas are involved in spatial perception, but it appears that egocentric spatial perception occurs mainly in later areas such as MT/V5. The brain augments spatial estimates into an egocentric perception and we have accordingly presented a method to augment active vision disparity data into an egocentric, unified, space-variant occupancy grid representation. The occupancy grid has been explicitly designed for integrating data from active vision, and for providing low-bandwidth and useful representations useful for perception in real time. We have shown how the occupancy grid can be used to extract information about the surroundings such as 3D scene motion and 3D cue-surface correspondences. For these reasons, we find that the space

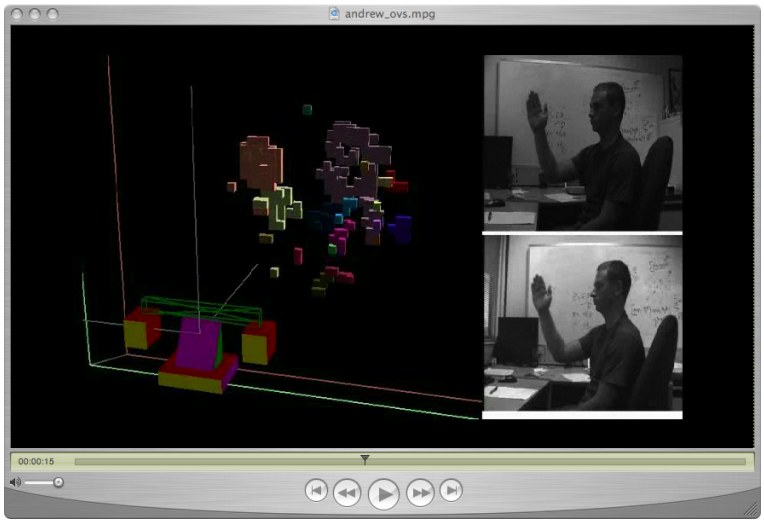



Figure 6.22: Online object segmentation using occupancy grid and cell flow demonstration (*snapshot - see Appendix C for full video*).

variant occupancy grid is particularly suited for spatial perception in synthetic primate vision. The regime provides coarse but real-time perception. It operates continually, over the entire field of view. In this sense it may be likened to a *peripheral* response.

# Chapter 7

# Coordinated Fixation

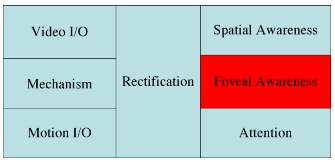



Figure 7.1: Foveal object segmentation and coordinated fixation. The left and right input (respectively), and the output from the Markov random field zero disparity filter (MRF ZDF).

In this chapter we synthesise coordinated stereo fixation. The approach enables real-time tracking of fast-moving objects and simultaneous segmentation of the tracked object or surface from image background.

## 7.1 Introduction

In the previous chapter, we developed a coarse spatial awareness based on continual processing over the entire visual field. In this sense, the spatial perception can

be considered a peripheral response. As such it operates regardless of gaze geometry. There is no concept of fixation upon surfaces in the scene. Introspection of human vision provides motivation for coordinated foveal fixation. Although peripheral processing occurs continually in the visual cortex irrespective of fixation, humans find it difficult to fixate on *unoccupied space*. Empty space contains little information; we are more concerned with resolute and focussed fixation upon objects or surfaces. The human visual system exhibits its highest resolution at the fovea. The extent of the fovea covers a retinal area of approximately the size of a fist at arms length [Wandell (1995)], conceptually in line with the task-oriented interactions of humans with the real world.

We limit foveal processing resources to the region of the images immediately surrounding the image centres. The region beyond the fovea is considered only for an estimate of where the foveas are to fixate next (for tracking purposes). For the resolution of our cameras, the fovea corresponds to a region of about 60x60 pixels and an approximate area of $0.5m^2$ at $2m$ distance. Actively moving this region over the scene facilitates coverage of a large visual workspace.

For humans, the boundaries of an object upon which we have fixated emerge effortlessly because the object is centred and appears with similar retinal coverage in our left and right eyes, whereas the rest of the scene usually does not. For synthetic vision, the approach is the same. The object upon which fixation has occurred will appear with identical pixel coordinates in the left and right images, that is, it will have zero disparity. For a pair of cameras with suitably similar intrinsic parameters, this condition does not require epipolar or barrel distortion rectification of the images. Camera calibration, intrinsic or extrinsic, is not required.

We aim to develop the propensity for the system to fixate upon objects in a manner that allows the segmentation of a spatially coherent target object from its surroundings, despite its colour or form. Further, such segmentation should permit tracking of arbitrary targets. We begin by introducing existing methods suitable for target fixation and tracking. Having considered such methods, we outline our approach. We develop a *maximum a posterior probability* (MAP) approach. We provide segmentation and tracking results using the regime and compare its performance to that of existing tracking methods.

### 7.1.1 Existing Fixation Methods

When tracking objects under real-world conditions, three main problems are encountered: ambiguity, occlusion and motion discontinuity. Ambiguities arise due to distracting noise, mismatching of the tracked objects and the potential for multiple targets or target-like distractors, to overlap the tracked target. Occlusions are inevitable in realistic scenarios where the subject interacts with the environment. Certainly, in dynamic scenes, the line of site between the cameras and target is not always guaranteed. At usual frame rates ($\sim 30fps$), the motion of agile subjects can seem erratic or discontinuous and motion models designed for tracking such subjects may be inadequate.

Existing methods for markerless visual tracking can be categorised according to the measurements and models they incorporate [Gavrila (1999)].

#### 7.1.1.1 Cue-Based Methods

In terms of cue measurement methods, tracking usually relies on either intensity information such as edges [Blake & Isard (1998); Cham & Rehg (1999); Gavrila & Davis (1996); Metaxas (1999)], skin colour, and/or motion segmentation [Imagawa *et al.* (1998); Jojic *et al.* (1999); Martin *et al.* (1998); Wren *et al.* (2000)], or a combination of these with other monocular cues [Loy *et al.* (2002); Toyama & Horvitz (2000); Triesch & von der Malsburg (2000)] or depth information [Azoz *et al.* (1998); Jennings (1999); Ong & Gong (1999); Wren *et al.* (2000)]. Fusion of cues at low levels of processing can be premature and may cause loss of information if image context is not taken into account. For example, motion information may occur only at the edges of a moving object, making the fused information sparse. Further, for non-spatial cue-based methods, occlusions by other target-like distractors may become indistinguishable from the tracked target.

*MeanShift* and *CamShift* methods are enhanced manifestations of cue measurement techniques that rely on colour chrominance-based tracking. For real-time performance, a single channel (chrominance) is usually considered in the colour model. This heuristic is based on the assumption that skin has a uniform chrominance. Such trackers compute the probability that any given pixel value corresponds to the target colour. Difficulty arises where the assumption of a single chrominance cannot be made. In particular, the algorithms may fail to track

multi-hued objects or objects where chrominance alone cannot allow the object to be distinguished from the background, or other objects.

The MeanShift algorithm is a non-parametric technique that ascends the gradient of a probability distribution to find the mode of the distribution [Cheng (1995); Fukunaga (1990)]. Particle filtering based on colour distributions and Mean Shift was pioneered by Isard and Blake [Isard & Blake (1998)] and extended by Nummiaro et al [Nummiaro *et al.* (2002)]. CamShift was initially devised to perform efficient head and face tracking Bradski (1998). It is based on an adaptation of MeanShift where the mode of the probability distribution is determined by iterating in the direction of maximum increase in probability density. The primary difference between the Cam Shift and the Mean Shift algorithms is that Cam Shift uses continuously adaptive probability distributions (recomputed each frame) while MeanShift is based on static distributions. More recently, Shen developed *Annealed MeanShift* to counter the tendency for MeanShift trackers to settle at local rather than global maxima [Shen *et al.* (2005)].

Although very successful in tracking the vicinity of a known chrominance, MeanShift methods are not designed for direct target segmentation and background removal (for classification enhancement). In terms of output, they provide an estimation of a tracked target bounding box, in the form of an estimate of the 0th and 1st moments of the target probability distribution function. They are also not typically capable of dealing with instantaneous or unexpected changes in the target colour model (such as, for example, when a hand grasps another object such as a mug or pen). They do not incorporate spatial constraints when considering a target in a 3D scene, and are not inherently intended to deal with occlusions and other ambiguous tracking cases (for example, a tracked target passing in front of a visually similar distractor). In such circumstances, these trackers may shift between alternate subjects, select the centre of gravity of the two subjects, or track the distracting object rather than the intended target. To preclude such ambiguities, motion models and classifiers can be incorporated, but they may rely upon weak and restrictive assumptions regarding target motion and appearance.

#### 7.1.1.2 Spatiotemporal Methods

Spatial techniques use depth information and/or temporal dynamic models to overcome the occlusion problem [Jojic *et al.* (1999); Wren *et al.* (2000)]. The use of spatial (depth) information can introduce problems associated with multiple camera calibration, and depth data is notoriously sparse, computationally expensive to recover, and can be inaccurate. Spatiotemporal continuity is not always a strong assumption. At frame rates, agile target motion may appear discontinuous or undergo occlusion. Methods such as Kalman filtered tracking [J.Joseph & LaViola (2003)] that are strongly reliant upon well-defined dynamics and temporal continuity may prove inadequate. Traditional segment-then-track (exhaustive search methods, for example, dynamic template matching) approaches are subject to cumulative errors where inaccuracies in segmentation affect tracking quality, which in turn affect subsequent segmentations.

Model-based methods incorporating domain knowledge such as tracking a hand which is part of an articulated entity (the human body), can be used to resolve some of the ambiguities. Joint tracking of articulated parts can be performed with an exclusion principle on observations [MacCormick & Blake (1999); Rasmussen & Hager (1998)] to alleviate such problems. *A priori* knowledge such as 2D target models may be used [Imagawa *et al.* (1998); Martin *et al.* (1998)]. Alternatively, a 3D model of the target and any associated articulated entity may be used such that kinematic constraints can be exploited [Cham & Rehg (1999); Ong & Gong (1999); Wren *et al.* (2000)]. 2D projections of deformable 3D models can be matched to observed camera images [Gavrila & Davis (1996); Metaxas (1999)]. Unfortunately, these methods can be computationally expensive, do not always resolve projection ambiguities, and performance depends heavily upon the accuracy of complex, subject dependent, articulated models and permitted motions.

#### 7.1.1.3 Zero Disparity Methods

Methods exist that do not require *a priori* models or target knowledge. Instead, the target is segmented using an uncalibrated semi-spatial response by detecting regions in images or cue maps that appear at the same pixel coordinates in the left and right stereo pairs. That is, regions that are at zero disparity. To overcome

pixel matching errors associated with gain differences between left and right views, these methods traditionally attempt to align vertical edges and/or feature points.

The work of Coombs and Brown involved the implementation of a simple *zero disparity filter* (ZDF) for the Rochester robot head [Coombs & Brown (1992)]. This basic method used the extraction of edge detail from image pairs to form binary images. These images were simply *and*-ed to extract potential zero disparity regions. The robot head then fixated its gaze upon the centre of gravity of the output.

Rougeaux, Kita, Kuniyoshi and Sakane [Rougeaux & Kuniyoshi (1997a); Rougeaux *et al.* (1994)] also investigated the use of virtual horopters to test whether the tracked subject was moving away from or towards the cameras. One of the stereo pair images (for example, the left image) was *virtually* shifted (in memory only) horizontally by a single pixel to the left (by purging the leftmost column of pixels) and then to the right (by adding an extra column at the left of the image), and the zero disparity response determined between each new image and the unaltered (right) image, for both cases. The virtual shift that yields the largest zero disparity response area was deemed the correct tracking direction, and the cameras were then verged or diverged accordingly such that the horopter best aligned with the tracked subject.

Oshiro applied a similar edge extraction method to foveal log-polar cameras [Oshiro *et al.* (1996)]. Yu used a wavelet representation to match broader image regions [Yu & Baozong (1996)]. Rougeaux later revisited the approach, combining the edge-based ZDF with optical flow for broader segmentation [Rougeaux & Kuniyoshi (1997b)]. Rae combined edge-based techniques with additional aligned point features such as corners, symmetry points and cue centroids [Rae & Ritter (1998)].

Rougeaux also implemented a method to compute disparity from phase difference using the output of complex band-pass filters as suggested by Sanger. Regions at zero disparity were extracted from the disparity maps. After several convolutions with a Symmetric Nearest Neighbour filter to enhance region boundaries and smooth areas of homogeneous grey level, a fast morphological algorithm created a binary mask for the target which was then fixated upon. The algorithm ran at approximately 30Hz frame-rate on two Intel i860 DSPs. However, the overall resolution was low - the disparity maps were sized only 32x32

pixels.

A multiple cue object tracking algorithm has been implemented on CeDAR that incorporates four simple cues: colour, edge detection, texture detection and motion [Dankers (2002)]. A cue voting scheme was adopted to identify pixel locations in the view frame that appeared target-like with respect to each cue. A simple zero disparity filter using virtual horopters then visually extracted the object from its surroundings, as well as mapping its position in three-dimensional space. The processing of all visual information took, on average, only $8ms$ per frame on a dual Pentium III computer. Unfiltered operation was susceptible to distractions due to target-like regions in the camera's views. Kalman Filtering reduced the effect of these distractions significantly. The algorithm allows successful real-time tracking of arbitrary objects through a cluttered environment (Figure 7.2).

Unfortunately, existing zero disparity methods do not cope well with bland subjects or backgrounds, and perform best when matching textured sites and features on textured backgrounds. Nevertheless, the zero disparity class of segmentation forms the base upon which we develop our approach.

### 7.1.2 Our Approach

We aim to ensure coordinated active stereo fixation upon a target, and to facilitate its robust pixel-wise segmentation. We propose a biologically inspired, conceptually simple method that segments and tracks the subject in parallel, eliminating problems associated with the separation of segmentation and tracking. The method inherently incorporates spatial considerations to disambiguate between, for example, multiple overlapping targets in the scene such that occlusions or distractions induced by non-tracked target-like distractors do not affect tracking of the selected target. As we shall see, the method does not rely on imposing motion models on the target trajectory, and can cope with gross partial occlusions. In this regard, the three common problems of ambiguity, occlusion and motion discontinuity are addressed. Despite using stereo vision, the approach does not require stereo camera calibrations, intrinsic or extrinsic. The method utilises dynamic stereo foveal scene analysis, and we choose an active implementation that has the benefit of increasing the volume of the visual workspace

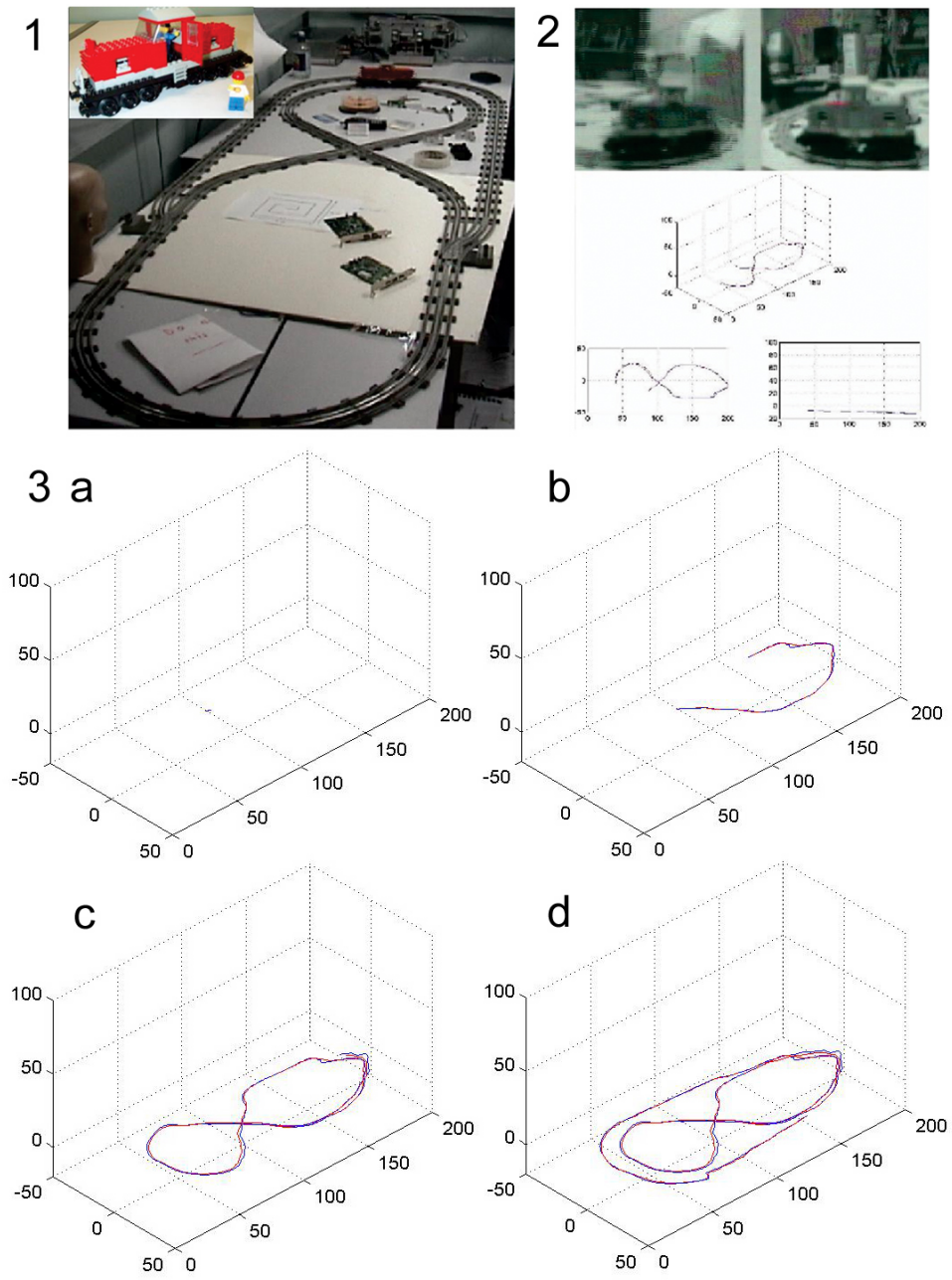



Figure 7.2: Multiple cue horopter tracking. 1) Scenario; 2) online screenshot showing left and right camera views and trajectory; 3) absolute target trajectory at *a* t=0*s*, *b*10*s*, *c* 20*s* and *d* 30*s* (units in *cm*).

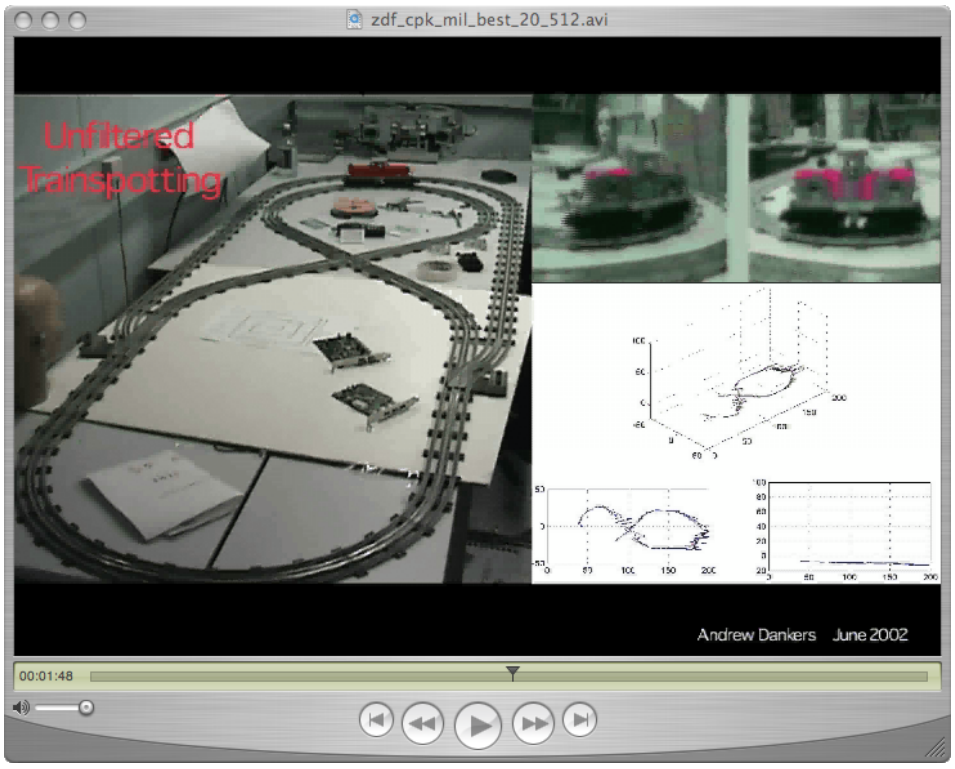



Figure 7.3: Multiple cue horopter tracking demonstration (*snapshot - see Appendix C for full video*).

Figure 7.4: Correlation-based ZDF output. NCC of $3 \cdot 3$ pixel regions at same coordinates in left and right images. Higher correlation values are shown more white.

## 7.2 Coordinated Fixation With Simultaneous Segmentation

We begin by assuming short baseline stereo fixation upon a target object. A probabilistic ZDF is formulated to identify the projection of the target object as it maps to identical image frame pixel coordinates in the left and right foveas. Simply comparing the intensities of pixels in the left and right images at the same coordinates is not adequate due to inconsistencies in, for example, saturation, contrast and intensity gains between the two cameras, as well as focus inconsistencies and noise. Figure 7.9 shows example correlation-based ZDF output where regions, rather than pixels, are compared providing somewhat improved results.

A human can easily distinguish the boundaries of the object upon which fixation has occurred even if one eye looks through a tinted lens. Accordingly, the regime should be robust enough to cope with these types of inconsistencies. One approach is to *normalised cross-correlate* (NCC) small templates in one image with pixels in the same template locations in the other image. The NCC function is shown in Equation 7.1:

$$NCC(I_1, I_2) = \frac{\sum_{(u,v)\in W} I_1(u,v) \cdot I_2(x+u, y+v)}{\sqrt{\sum_{(u,v)\in W} I_1^2(u,v) \cdot \sum_{(u,v)\in W} I_2^2(x+u, y+v)}}, \tag{7.1}$$

where $I_1, I_2$ are the compared left and right image templates of size $W$ and $u, v$ are coordinates within the template. Figure 7.4 shows the output of this approach. Bland areas in the images have been suppressed (set to 0.5) using *difference of Gaussians*[1] (DOG) pre-processing. The 2D DOG kernel is constructed

[1]The *difference of Gaussians* function approximates the *Laplacian of Gaussians* function. Convolving a 2D DOG kernel with an image suppresses bland regions.

using symmetric separable 1D convolutions. The 1D DOG function is shown in Equation 7.2:

$$DOG(I) = G_1(I) - G_2(I), \tag{7.2}$$

where $G_1(), G_2()$ are Gaussians with different standard deviations $\sigma_1, \sigma_2$ according to:

$$G(x) = \frac{e^{-x^2}}{2\sigma^2}, \tag{7.3}$$

DOG pre-processing is used to suppress untextured regions that always return a high NCC response whether they are at zero disparity or not. As Figure 7.4 shows, the output is sparse and noisy. The palm is positioned at zero disparity but is not categorised as such.

To improve results, image context needs to be taken into account. Contextual information can assist by assigning similar labels to visually similar neighbourhoods. Most importantly, contextual refinement allows slight relaxation of the zero disparity assumption such that non-planar surfaces, or surfaces that are not perpendicular to the camera optical axes – but appear visually similar to the dominantly zero disparity region – can be segmented as the same object. For these reasons, we adopt a Markov random field [Geman & Geman (1984)] (MRF) approach.

### 7.2.1 MRF ZDF Formulation

The MRF formulation defines that the value of a random variable at the set of sites (pixel locations) $S$ depends on the random variable configuration field $f$ (labels at all sites) only through its neighbours $N \in S$. For a ZDF, the set of possible labels at any pixel in the configuration field is binary, that is, sites can take either the label *zero disparity* ($f(S) = l_z$) or *non-zero disparity* ($f(S) = l_{nz}$). For an observation $O$ (in this case an image pair), Bayes' law states that the *a posterior* probability $P(f \mid O)$ of field configuration $f$ is proportional to the product of the likelihood $P(O \mid f)$ of that field configuration given the observation and the prior probability $P(f)$ of realisation of that configuration:

$$P(f \mid O) \propto P(O \mid f) \cdot P(f). \tag{7.4}$$

The problem is thus posed as a MAP optimisation where we want to find the configuration field $f(l_z, l_{nz})$ that maximises the *a posterior* probability $P(f \mid O)$.

In the following two sections, we adapt the approach of Boykov *et al.* (1997) to construct the terms in Equation 7.4 suitable for ZDF tracking.

#### 7.2.1.1 Prior term $P(f)$

The *prior* term encodes the properties of the MAP configuration we seek. It is intuitive that the borders of zero disparity regions coincide with edges (or intensity transitions) in the image. The Hammersly-Clifford theorem, a key result of MRF theory, is used to represent this property:

$$P(f) \propto e^{-\sum_C V_C(f)}. \tag{7.5}$$

*Clique potential* $V_C$ describes the prior probability of a particular realisation of the elements of the clique $C$. For our neighbourhood system, MRF theory defines cliques as pairs of horizontally or vertically adjacent pixels. Equation 7.5 reduces to:

$$P(f) \propto e^{-\sum_p \sum_{q \in N_p} V_{p,q}(f_{p,f_q})}. \tag{7.6}$$

In accordance with Boykov *et al.* (1997), we assign clique potentials using the *Generalised Potts Model* where clique potentials resemble a well with depth $u$:

$$V_{p,q}(f_p, f_q) = u_{p,q} \cdot (1 - \delta(f_p - f_q)), \tag{7.7}$$

where $\delta$ is the unit impulse function. Clique potentials are isotropic ($V_{p,q} = V_{q,p}$), so $P(f)$ reduces to:

$$P(f) \propto e^{-\sum_{\{p,q\} \in \varepsilon_N} \{2u \quad \forall f_p \neq f_q, 0 otherwise.}} \tag{7.8}$$

$V_C$ can be interpreted as a cost of discontinuity between neighbouring pixels $p, q$. In practice, we assign the clique potentials according to how continuous the image is over the clique using the Gaussian function:

$$V_c = \frac{e^{-\Delta I_C^2}}{2\sigma^2}, \tag{7.9}$$

where $\Delta I_C$ is the change in intensity across the clique, and $\sigma$ is selected such that $3\sigma$ approximates the minimum intensity variation that is considered smooth.

Note that at this stage we have looked at one image independently of the other. Stereo properties have not been considered in constructing the prior term.

#### 7.2.1.2 Likelihood term $P(O \mid f)$

The *likelihood* term describes how likely it is that an observation $O$ matches a hypothesised configuration $f$ and involves incorporating stereo information for assessing how well the observed images fit the configuration field. It can be equivalently represented as:

$$P(O \mid f) = P(I_A \mid f, I_B), \tag{7.10}$$

where $I_A$ is the primary image and $I_B$ the secondary (chosen arbitrarily) and $f$ is the hypothesised configuration field. In terms of image sites $S$ (pixels), Equation 7.10 becomes:

$$P(O \mid f) \propto \prod_S g(i_A, i_B, l_S), \tag{7.11}$$

where $g()$ is some symmetric function [Boykov *et al.* (1997)] that describes how well label $l_S$ fits the image evidence $i_A \in I_A$ and $i_B \in I_B$ corresponding to site $S$. It could for instance be a Gaussian function of the difference in observed left and right image intensities at $S$; we evaluate this instance (Equation 7.15) and propose alternatives later.

To bias the likelihood term towards a specific type of object, we can include an *a priori* target appearance term $H_S$, Equation 7.12. This term is not required for the system to operate, it merely provides a greater propensity for the MRF ZDF detector to track specific properties of objects based upon *a priori* knowledge as required by the task. The term enumerates how target-like a pixel site is in terms of its colour and texture (by assigning a probability to site S in each image). It may be formulated to best suit the task, or it may be modulated autonomously. For now, we ignore the term.

$$P(O \mid f) \propto \prod_S g(i_A, i_B, l_S, H_S) \tag{7.12}$$

#### 7.2.1.3 Energy Minimisation

We have assembled the terms in Equation 7.4 necessary to define the MAP optimisation problem:

$$P(f \mid O) \propto e^{-\sum_p \sum_{q \in N_p} V_{p,q}(f_p, f_q)} \cdot \prod_S g(i_A, i_B, l_S). \tag{7.13}$$

Maximising $P(f \mid O)$ is equivalent to minimising the energy function:

$$E = \sum_{p} \sum_{q \in N_p} V_{p,q}(f_{p,f_q}) - \sum_{S} ln(g(i_A, i_B, l_S)). \tag{7.14}$$

### 7.2.2 Optimisation

A variety of methods can be used to optimise the above energy function including *simulated annealing* and *graph cuts*. For active vision, high-speed performance is a priority. At present, a graph cut technique is the preferred optimisation technique, and is validated for this class of optimisation as per Kolmogorov & Zabih (2002b). We adopt the method used in Kolmogorov & Zabih (2002a) for MAP stereo disparity optimisation (we omit their use of $\alpha$*–expansion* as we consider a purely binary field). In this formulation, the problem is that of finding the *minimum cut* on a *weighted graph*.

A weighted graph $G$ comprising of vertices $V$ and edges $E$ is constructed with two distinct terminals $l_{zd}, l_{nzd}$ (the source and sink). A cut $C = V^s, V^t$ is defined as a partition of the vertices into two sets $s \in V^s$ and $t \in V^t$. Edges $t, s$ are added such that the cost of any cut is equal to the energy of the corresponding configuration. The cost of a cut $|C|$ equals the sum of the weights of the edges between a vertex in $V^s$ and a vertex in $V^t$.

The goal is to find the cut with the smallest cost, or equivalently, compute the *maximum flow* between terminals according to the Ford Fulkerson algorithm [Ford & Fulkerson (1962)]. The minimum cut yields the configuration that minimises the energy function. Details of the method can be found in Kolmogorov & Zabih (2002a). It has been shown to perform (at worst) in low order polynomial time, but in practice performs in near linear time for graphs with many short paths between the source and sink, such as this [Kolmogorov & Zabih (2002b)].

### 7.2.3 Robustness

In the following sections we use a hand as the target object to evaluate MRF ZDF segmentation and tracking performance. A hand was selected because it is agile, rapidly deformable and often non-planar. Hands are skin-coloured, so tracking is often complicated by the presence of additional regions of skin in the image frames. We deliberately complicate matters further by introducing a

second distracting hand. The direct and continuous attachment of the hand to the arm may also complicate segmentations based on appearance. Hands can be made to easily come in contact with, or grasp, objects. All of these factors mean that a hand often constitutes a difficult target for real-time tracking and segmentation algorithms.

Using a hand as a target, we now look at the situations where the MRF ZDF formulation performs poorly and provide methods to combat these weaknesses. Figure 7.9*a* shows ZDF output for typical input images where the likelihood term has been defined using intensity comparison. Output was obtained at approximately 27fps for the 60x60 pixel fovea on a standard 3GHz single processor PC. For this case, $g()$ in Equation 7.11 has been defined as:

$$g(i_A, i_B, f) = \{\, e^{-(\Delta I_C)^2} / 2\sigma^2 \forall f = l_z, 1 - (e^{-(\Delta I_C)^2} / 2\sigma^2) \forall f = l_{nz}. \qquad (7.15)$$

The variation in intensity at corresponding pixel locations in the left and right images is significant enough that the ZDF has not labelled all pixels on the hand as being at zero disparity. To combat such variations, NCC is instead used (Figure 7.9*b*). Whilst the ZDF output improved slightly, processing time per frame was significantly increased ($\sim 12fps$). As well as being slow, this approach requires much parameter tuning. Bland regions return a high correlation whether they are at zero disparity or not, and so the correlations that return the highest results cannot be trusted. A threshold must be chosen above which correlations are disregarded. This also has the consequence of disregarding the strongest valid correlations. Additionally, a histogram of correlation output results is not symmetric (left, Figure 7.6). There is difficulty in converting such output to a probability distribution about a mean of 0.5, or converting it to an energy function penalty.

To combat the thresholding problem with the NCC approach, the images can be pre-processed with a DOG kernel. The output using this technique (Figure 7.9*c*) is good, but is much slower than all previous methods ($\sim 8fps$) and requires yet more tuning at the DOG stage. It is still susceptible to the problem of non-symmetric output.

We prefer a comparator whose output histogram resembles a symmetric distribution, so that these problems could be alleviated. For this reason we chose a simple *neighbourhood descriptor transform* (NDT) that preserves the relative

intensity relations between neighbouring pixels (in a fashion similar to but less rigidly than that of the *Rank* transform), and is unaffected by brightness or contrast variations between image pairs. Figure 7.5 depicts the definition of the NDT transform.

In this approach, we assign a boolean descriptor string to each site and then compare the descriptors. The descriptor is assembled by comparing pixel intensity relations in the 3x3 neighbourhood around each site (Figure 7.5). In its simplest form, for example, we first compare the central pixel at a site in the primary image to one of its four-connected neighbours, assigning a *'1'* to the descriptor string if the pixel intensity at the centre is greater than that of its northern neighbour and a *'0'* otherwise. This is done for its southern, eastern and western neighbours also. This is repeated at the same pixel site in the secondary image. The order of construction of all descriptors is necessarily the same. A more complicated descriptor would be constructed using more than merely four relations[1]. Comparison of the descriptors for a particular site is trivial, the result being equal to the sum of entries in the primary image site descriptor that match the descriptor entries at the same positions in the string for the secondary image site descriptor, divided by the length of the descriptor string.

Figure 7.6 shows histograms of the output of individual neighbourhood comparisons using the NCC DOG approach (left) and NDT approach (right) over a series of sequential image pairs. The histogram of NDT results is a symmetric distribution about a mean of 0.5, and hence is easily converted to a penalty for the energy function.

Figure 7.9*d* shows NDT output for typical images. Assignment and comparison of descriptors is faster than NCC DOG ($\sim 27fps$), yet requires no parameter tuning. In Figure 7.9*e*, the left camera gain was maximised, and the right camera contrast was maximised. In Figure 7.9*f*, the left camera was defocussed and saturated. Segmentation performance remained good under these artificial extremes.

[1]Experiment has shown that a four neighbour comparator gives results that compare favorably (in terms of trade-offs between performance and processing time) to more complicated descriptors.

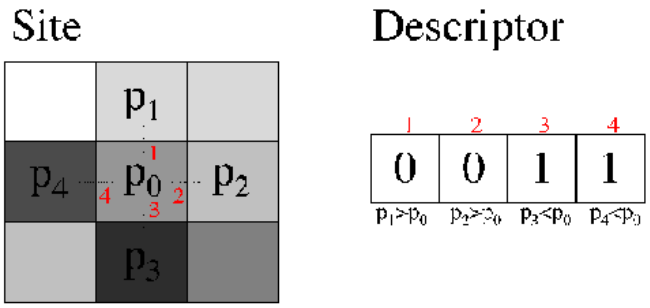



Figure 7.5: NDT descriptor construction. An example four-entry descriptor string is shown for the adjacent $9 \cdot 9$ pixel neighbourhood.

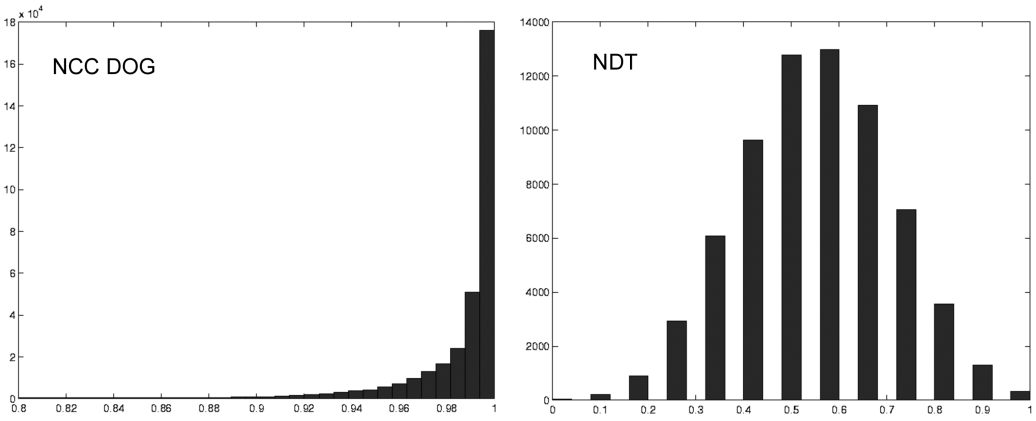



Figure 7.6: Histograms of individual NCC DOG (left) and NDT (right) neighbourhood comparisons for each entire frame over a series of frames.

### 7.2.4 Incorporating Colour

So far we have considered the intensity (Y) channel only in our formulation. Colour is a cue whose regional consistency tends to correspond well with the edges of objects. For example, in an image, the borders of a yellow tennis ball lying on a court correspond to the area of yellow pixels. This obvious relationship can be exploited to zero disparity on a target object and to distinguish it from the background.

We obtain YUV colour space images from the cameras, where Y encodes the intensity channel, and U and V encode the intensity-normalised colour chrominance channel values. To incorporate colour into the formulation, we modulate the likelihood term (the output of each NDT left-right intensity configuration comparison that operates on the Y channels) by a measure of left-right colour chrominance similarity at the same image locations. The similarity measure is obtained by calculating the Mahlonobis distance between colour chrominances ($\Delta C$) at the compared neighbourhood locations in the left and right images according to:

$$\Delta C = \sqrt{(L_u - R_u)^2 + (L_v - R_v)^2}, \tag{7.16}$$

where $L_u$,$L_v$ and $L_u$,$L_v$ are the U,V colour chrominance components at the compared neighbourhood in the left and right images respectively.

We convert the measured colour chrominance distances to a colour chrominance similarity modulation factor $m_c$ using a Gaussian lookup function:

$$m_c = m_{c,min}(1 + \frac{e^{-\Delta C^2}}{2\sigma^2}), \tag{7.17}$$

where $m_{c,min}$ is the minimum desired modulation factor, and $\sigma$ is selected such that $3\sigma$ approximates the maximum passable chrominance variation. For example, if the colour chrominance distance is 0.0 the corresponding colour modulation $m_c$ would be 1.0. As the colour distance increases, the modulation factor tends towards $m_{min}$ according to the above Gaussian function.

In this manner, neighbourhood regions whose intensity configuration looks the same across the left and right images, but whose colour chrominance varies significantly, would have their zero disparity likelihood suppressed accordingly.

### 7.2.5 Computational Pipelining

Rather than computing the likelihood and prior terms as they are indexed, we can pre-compute lookup maps for the prior (mono) term and likelihood (stereo) terms at all foveal locations upon image acquisition. We then index to these maps at runtime. In this manner, performance gains are possible due to MMX vector pre-computation of these lookup maps.

## 7.3 Incorporating Tracking

Target tracking is implemented using a combination of virtual and physical retinal shifts. Figure 7.7 describes the four steps of the tracking algorithm. Initialisation of the system is simple. The operator merely passes their hand through the area surrounding the arbitrary initial stereo fixation point. At a fixation point $2m$ from the cameras, the initial search window defines a receptive volume of about $0.5m^3$. Once tracking begins, segmentation of the zero disparity region induced by the hand is followed by continual NCC alignment of the horopter such that the zero disparity segmentation area is maximised. The NCC search window is sufficient to cope with the upper limits of typical hand motions between successive frames. The MRF ZDF process reduces the segmented area to that associated with a 2D projection of the object on the horopter, such that occlusions or secondary hands do not distract track unless they are essentially touching the tracked hand (see Section 7.5.2.1). If track is lost, it will resume on the zero disparity region induced by the subject closest to the fixation point. In this manner, if track is lost, the subject need only return their hand to the volume surrounding the current fixation point (where track was lost).

The method of virtual verification followed by physical motion copes with rapid movement of the hand, providing an awareness of whether the hand has moved towards or away from the cameras, so that the physical horopter can be shifted to the location that maximises the zero disparity area associated with the hand. It is emphasised that template matching is not used to track the hand; it is only used to estimate the pixel shift required to align the virtual horopter over the hand. Tracking is performed by extracting the zero disparity region at the virtual horopter, and physically moving the cameras to point at the centre of gravity of

**MRF ZDF Tracking Algorithm:**

1. Determine virtual shift required to approximately align virtual horopter over subject: the pixel distance $d$ between a small template (approximately 30x30 pixels) at the centre of the left image and its location of best match in the right image is determined using NCC. We conduct the search in a window a few pixels above and below the template location in the left image and up to 10 pixels to the left and right in the right image. In this manner, the NCC will only return a high correlation result if the subject in the template is located near the 3D scene fixation point.

2. Perform a virtual shift of the left fovea by $d/2$ and the right fovea by $-d/2$ to approximately align the location of best correlation in the virtual centre of the left and right foveas. If the NCC result is not sufficiently high, no physical shift is conducted and the process returns to the first step.

3. MRF ZDF segmentation extracts the zero disparity pixels associated with a 2D projection of the hand from the virtually aligned foveas. If there is indeed a hand at the virtual fixation point, the area of the segmented region will be significantly beyond zero.

4. If the area is greater than a minimum threshold, the virtual shift has aligned the centre of the images over the hand. In this case, a physical movement of the cameras is executed that reduces the virtual shift to zero pixels, and aligns the centres of the cameras with the centre of gravity of the segmented area. If the area is below the threshold, there is little likelihood that a hand or object is at the virtual fixation point, and no physical shifting is justified.

The process then cycles, continuing from step 1.

Figure 7.7: MRF ZDF tracking algorithm.


Figure 7.8: Online coordinated foveal fixation, tracking and object segmentation. An agile, rapidly deformable object - a hand - provides a formidable target (*snapshot - see Appendix C for full video*).

the segmented zero disparity region, if it is significantly non-zero. Thus, virtual horopter alignment is generally successful if any part of the hand is selected as the template, and does not depend on the centre of the hand being aligned in the template. Figure 7.8 shows a demonstration of real-time coordinated foveal fixation, tracking and segmentation of an agile, deformable target - a hand.

## 7.4 Results

Target tracking and segmentation for the purpose of real-time HCI gesture recognition and classification must exhibit robustness to arbitrary lighting variations over time and between the cameras, poorly focussed cameras, hand orientation, hand velocity, varying backgrounds, foreground and background distractors including non-tracked hands and skin regions, and hand appearance such as skin or hand covering colour. System performance must also be adequate to allow natural hand motion in HCI observations. The quality of the segmentation must be sufficient that it does not depart from the hand over time. Ideally, the method should find the hand in its entirety in every frame, and segment adequately for gesture recognition. For recognition, segmentation need not necessarily be perfect for every frame because if track is maintained, real-time classification is still possible based on classification results that are validated over several frames. Frames that are segmented with some error still usually provide useful segmentation in-

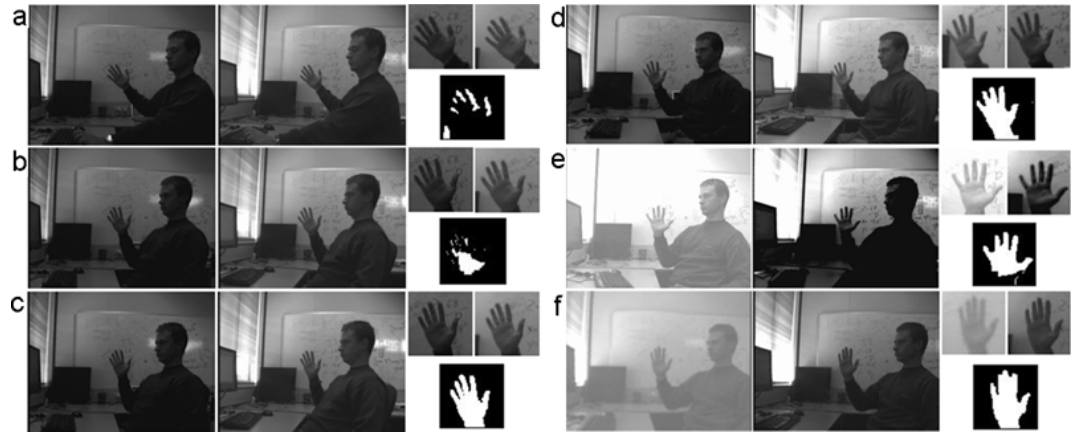



Figure 7.9: MRF ZDF hand segmentation. The left and right images and their respective foveas are shown with ZDF output (bottom right) for each case *a-f*. Result *a* involves intensity comparison, *b* involves NCC, and *c* DOG NCC for typical image pairs. Results *d-f* show superior NDT output for typical images *d*, and extreme adverse conditions *e,f*.

formation to the classifier.

Figure 7.9 shows snapshots from online MRF ZDF hand segmentation sequences. Segmentations on the right (*d-f*) show robust performance of the NDT comparator under extreme lighting, contrast and focus conditions. Figure 7.10 shows the robust performance of the system in difficult situations including foreground and background distractors. As desired, segmentation of the tracked hand continues. Figure 7.11 shows a variety of hand segmentations under typical circumstances including reconfiguring, rotating and moving hands in real time.

Figure 7.11 shows various segmentations for conceivable symbolic gestures. Segmentation quality is such that the target is extracted from its surroundings which has significant benefits in classification processes because the operation is not tainted by background features. The last two examples in Figure 7.11 show the segmentation of a hand holding a set of keys, and a hand holding a stapler. In these two cases, the conjoined hand and object form a volume that is segmented as the same object. Such volumetric segmentations may be useful for examining the contextual interaction of objects. Foveal segmentation of such interaction events from the background may be of great benefit to tasks such as object manipulation, or the inspection of connected objects.

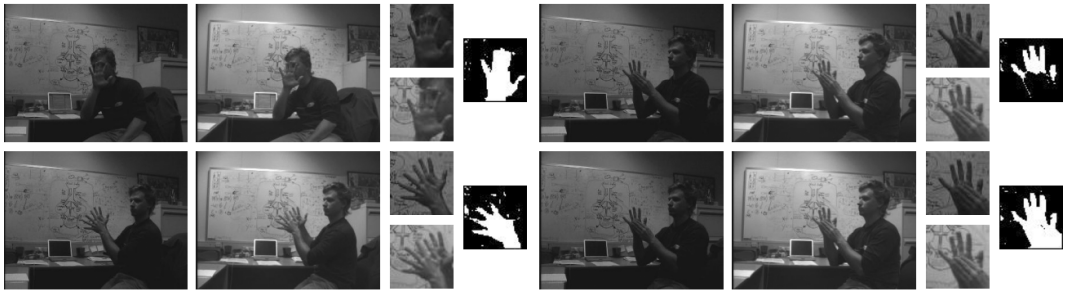


Figure 7.10: Robust performance in difficult situations. Segmentation of a tracked hand from a face in the near background (top left); from a second distracting hand in the background (bottom left); and from a distracting occluding hand in the immediate foreground, a distance of $3cm$ from the tracked hand at a distance of $2m$ from the cameras (top right). Once the hands are closer together than $3cm$, they are segmented as the same object (bottom right).

Figure 7.11: Segmentation of objects with intricate borders. The last two frames show composite objects - a hand grasping a stapler, and a hand grasping a bunch of keys.

## 7.5 Performance

### 7.5.1 Speed

On average, the system is able to fixate upon and track subjects at 27fps, including display. Acquiring the initial segmentation takes a little longer ($\sim 23-25fps$ for the first few frames) after which successive MRF ZDF optimisation results do not vary significantly so using the previous segmentation as an initialisation for the current frame accelerates MRF labeling. Similarly, the change in segmentation area between consecutive frames at 30fps is typically small, allowing sustained high frame rates after initial segmentation. The frame rate remains above 20fps and is normally up to the full 30fps camera frame rate.

### 7.5.2 Quality

In typical tracking of a reconfiguring moving hand over 100 consecutive frames, inaccurate segmentation of the hand typically occurs in around 15 frames. We describe a frame as *inaccurate* if the segmentation result has incorrectly labelled more than 10% of the pixels associated with the hand segmentation (either misslabeling pixels on the hand as not being on the hand, or vice versa). These figures have been determined by recording segmentation output for typical gesturing sequences and having a human arbitrator review and estimate the percentage of inaccurate pixels in each frame.

Segmentation success also depends on the complexity of hand posture. For example, if the hand is posed in a highly non-planar fashion or a pose whose dominant plane is severely non-perpendicular to the camera optical axes, non-successful segmentation can degrade to up to around 50 frames in 100. In these situations, the zero disparity assumption is violated over some parts of the hand. The induced relaxation of the zero disparity assumption due to MRF contextual refinement is not always sufficient to segment the hand. Under such circumstances, methods reliant on prior knowledge could conceivably assist segmentation - for example, if the colour or appearance of the hand was known prior to segmentation and incorporated using the $H_S$ term from Equation 7.12. Nevertheless, despite some inaccurately segmented frames, track is rarely lost for natural motions and gestures.

The approach compares favorably to other ZDF approaches that have not incorporated MRF contextual refinement, allowing relaxation and refinement of the zero disparity assumption such that surfaces that are not perpendicular to the camera axis can be segmented.

#### 7.5.2.1 Foreground and Background Robustness

Figure 7.10 shows examples of segmentations where subject-like distractors such as skin areas, nearby objects, or other hands are present. For the case where the tracked hand passes closely in front of a face (that has the same skin colour and texture as the tracked hand), the system successfully distinguishes the tracked hand from the nearby face distractor (Figure 7.10, top left). Similarly, when the tracked hand passes in front of a nearby hand, segmentation is not affected (Figure 7.10, bottom left). Cue or model-based methods are likely to have difficulty distinguishing between the tracked hand and the background hand.

The right side images in Figure 7.10 show the case where a tracked hand is occluded by an incoming distractor hand. The hands are located approximately $2m$ from the cameras in this example. Reliable segmentation of the tracked hand (behind) from the occluding distractor hand (in front) remains until the distractor hand is a distance of approximately $3cm$ from the tracked hand. Closer than this the hands are segmented as a connected object, which is conceptually valid.

### 7.5.3 Tracking Constraints

An object can be tracked as long as it does not move entirely out of the fovea between consecutive frames. This is because no predictive tracking is incorporated (such as Kalman filtering). In practice, we find that objects must move fast enough that they leave the fovea completely between consecutive frames. Tracking a target as it moves in the depth direction (towards or away from the cameras) is sufficiently rapid that loss of track does not occur. In interacting with the system, we find that track was not lost for natural hand motions (Figure 7.8).

The visual workspace for the system remains within a conic whose arc angle is around $100^o$. Performance remains effective up to a workspace depth (along the camera axis) of $5m$, for the resolution, baseline and zoom settings of our stereo

apparatus. Higher resolution or more camera zoom would increase disparity sensitivity, permitting zero disparity filtering at larger scene depths.

#### 7.5.3.1 Segmentation for Recognition

Primates bind the different visual attributes of an object, such as colour or form, into a unitary precept [Trieisman & Gelade (1980); Reynolds *et al.* (2000)]. The MRF ZDF segmentation may facilitate the identification of a segmented object by removing background information. A recognition step could be applied to the segmented output. The detailed contours of the segmentation may contribute to object identification.

A hypothetical example is now considered: a black cross located some distance directly behind a black circle, for example, would elicit a camera image in which the circle partially occludes the cross. If such an image was passed to a two-class classifier designed to identify crosses or circles, the image may be classified as containing either. Based on spatial constraints, however, a MRF ZDF segmentation would return either *only* pixels on the cross or those *only* on the circle, depending upon which object was tracked at the stereo fixation point. The segmentation identifies pixels not on the object at fixation. In this manner, spatial information (and the image frame disparities elicited in binocular images) largely eliminates classification ambiguity.

### 7.5.4 Comparison to State-of-Art

Our method is based on active vision hardware, and as such, it is difficult to find a performance metrics that compare the MRF ZDF method with methods that do not use active vision mechanisms. Additionally, implementation details for other ZDF methods are difficult to obtain, and are usually hardware and calibration dependent, such that reproduction is not viable. Methods that do not use contextual refinement for direct segmentation cannot be party to a segmentation performance comparison. Having said that, we provide samples of output from other implementations to allow the reader to assess performance visually.

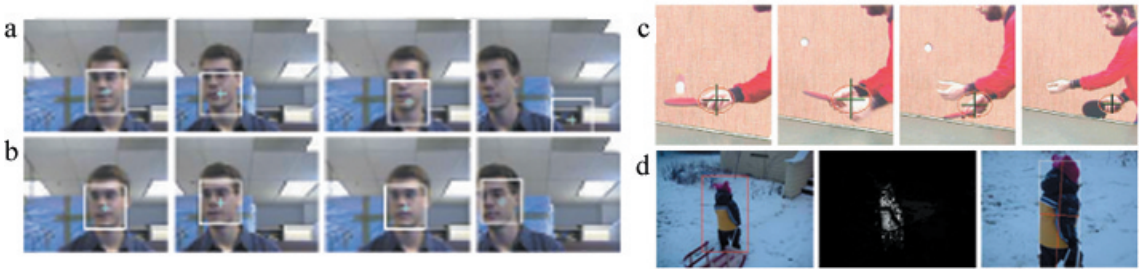



Figure 7.12: Comparison to other methods. Example output, images reproduced with permission from *a* Shen (MeanShift [Shen *et al.* (2005)]), *b* Shen (Annealed MeanShift [Shen *et al.* (2005)]), *c* Comaniciu (CamShift [Comaniciu *et al.* (2003)]), *d* Allen (CamShift [Allen *et al.* (2003)]).

#### 7.5.4.1 Comparision to Colour-Based Methods

We provide tracking output from recent methods for empirical evaluation (Figure 7.12). These methods provide bounding box output only, and as such do not deal with segmenting, for example, two overlapping hands (Figure 7.12*c*).

#### 7.5.4.2 Comparision to ZDF-Based Methods

Figure 7.13 shows sample ZDF output from existing methods for comparison. These methods provide probability distribution and bounding box outputs. The underlying probability maps may be suitable for MRF refinement such as ours, but they do not inherently provide segmentation.

#### 7.5.4.3 Comparison to Non-MRF Methods

Figure 7.4 shows sample ZDF output from our system without the incorporation of MRF contextual refinement. Figure 7.9*c* shows output using the same algorithm as in Figure 7.4, but incorporates MRF contextual refinement from the original images. Any attempt to use the output in Figure 7.4 alone for segmentation (via any, perhaps complex, method of thresholding), or for tracking, would not yield results comparable to those achievable by using the output in Figure 7.9*c*. The underlying non-MRF processes may or may not produce ZDF probability maps comparable to those produced by others (Section 7.5.4.2). However, the tracking quality achievable by incorporating MRF contextual image

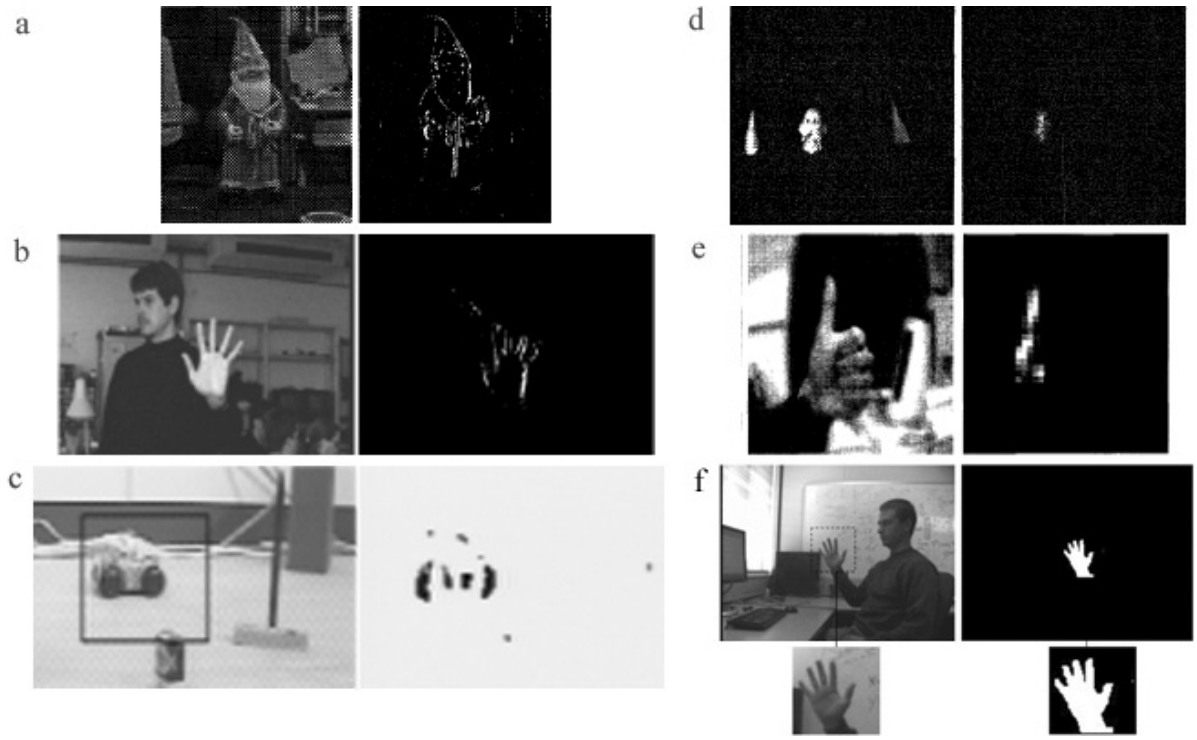



Figure 7.13: ZDF performance comparison. Images reproduced with permission from: *a* Oshiro [Oshiro *et al.* (1996)], *b* Rae [Rae & Ritter (1998)], *c* Rougeaux [Rougeaux *et al.* (1994)], *d* Yu [Yu & Baozong (1996)], *e* Rougeaux [Rougeaux & Kuniyoshi (1997b)], *f* the presented MRF ZDF algorithm.

information refinement is better than is possible by the underlying ZDF process.

## 7.6 Discussion

It is critical that the MRF ZDF refinement operates consistently at, or near, frame rate. This is because we consider only the 60x60 pixel fovea when extracting the zero disparity region. At slower frame rates, a subject could more easily escape the fovea, resulting in loss of track. Increasing the fovea size could help prevent this occurring, but would have the consequence of increasing processing time per frame.

Our method uses all image information, it does not match only edges, features or blobs extracted from single or multiple cues. The strongest labeling evidence does indeed come from textured and feature rich regions of the image, but the Markov assumption propagates strongly labelled pixels through pixel neighbourhoods that are visually similar, until edges or transitions in the images are reached. The framework deals with the trade-off between edge strengths and neighbourhood similarity in the MRF formulation.

In contrast to many motion-based methods, where motion models are used to estimate target location based on previous trajectories and motion models (for example, Kalman filtering), the implementation does not rely upon complex spatiotemporal models to track objects. It merely conducts a continual search for the maximal area of ZDF output, in the vicinity of the previous successful segmentation. The segmentations can subsequently be used for spatial localisation of the tracked object, but spatiotemporal dynamics do not form part of the tracking mechanism.

### 7.6.1 Incorporation with Synthetic Perception

The ability to fixate upon, segment, and track scene surfaces in the fovea can operate in parallel with the coarse peripheral spatial perception described in the previous chapter. Primates combine foveal and peripheral perception into a unified scene representation. Accordingly, when operating in parallel with peripheral spatial perception, target foveal fixation and segmentation facilitates investigations into primate-like perception.

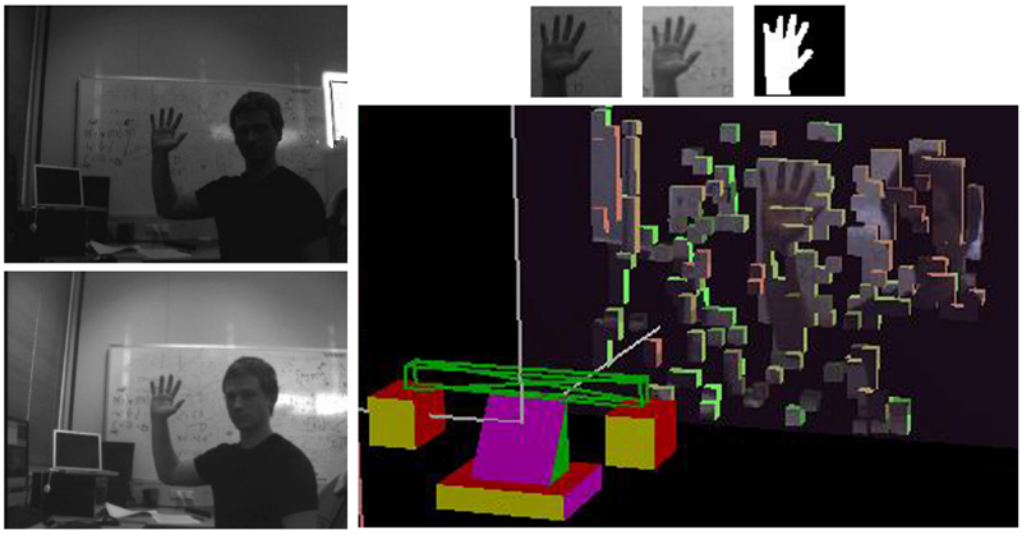


Figure 7.14: Bimodal system operation. Left: left (top) and right (bottom) input images. Right: Foveal perception (top) and peripheral perception (bottom). Foveal segmentation enhances the coarse perception of mass in the scene.

Once the peripheral mode has provided a rough perception of where mass is in the scene, the foveal mode allows coordinated stereo fixation upon mass/objects in the scene, and enables extraction of the object or region of mass upon which fixation occurs. By adjusting the camera geometry, the system is able to keep the object at zero disparity and centred within the foveas. Moreover, while the target is tracked in the foveal mode, the peripheral mode continually provides spatial information about the object's surroundings. This combined ability is potentially useful for examining how a tracked target interacts with its surroundings.

Figure 7.14 shows a snapshot of output of the foveated and peripheral perception modes operating in parallel. Bimodal perception operates at approximately 15Hz on the 3GHz single processor PC. Figure 7.15 shows a demonstration movie of bimodal perception.

Obtaining a peripheral awareness of the scene and extracting objects within the fovea permits experimentation in fixation and gaze arbitration. Prioritised monitoring of objects in the scene is the next step in our work towards artificial scene awareness. In the next chapter, we investigate attention and autonomous target selection.

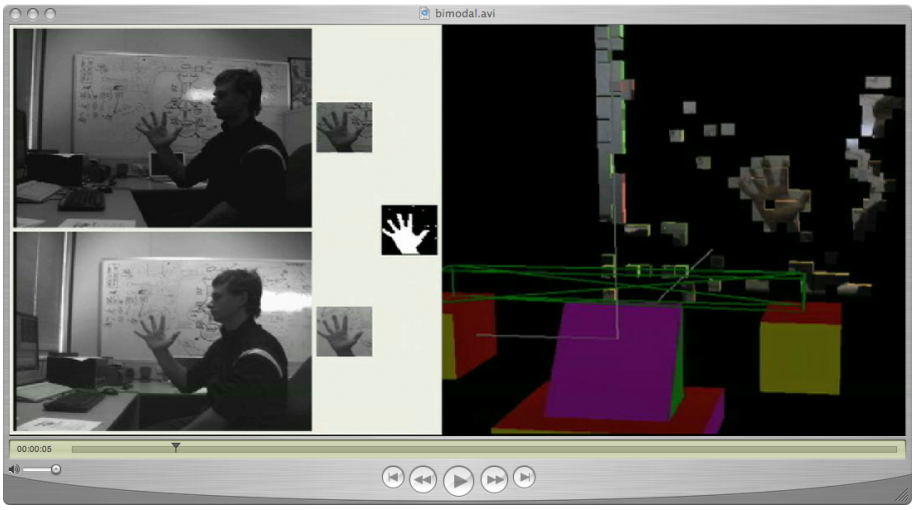



Figure 7.15: Online bimodal perception demonstration (*snapshot - see Appendix C for full video*).

#### 7.6.1.1 Processing Network Integration

When implemented as a processing node in the processing network, the coordinated fixation and segmentation algorithm requires image input. We take rectified image input from the rectification server. The MRF ZDF node requests Y, U, and V rectified channels from the left and right cameras. Mosaic parameters are not required. The node outputs the segmentation mask or the multiplication of the mask with the original image, as selected by client processes.

The node calculates axis shift distances for maintaining zero disparity tracking. It can be set to communicate directly with the motion control server for direct, automatic gaze control, or it can pass the parameters to client nodes for their arbitration. In the instance that the MRF ZDF node controls motion directly, other nodes may still send control commands which override MRF ZDF control. MRF ZDF gaze control can be reinstated after any such interruptions.

## 7.7 Summary

A Markov random field zero disparity filter (MRF ZDF) has been formulated and used to fixate upon, segment and track arbitrarily moving, rotating and re-configuring objects, performing accurate marker-less pixel-wise segmentation.

Target extraction is robust to lighting changes, defocus, target appearance, foreground and background clutter including non-tracked distracting (visually similar) targets, and partial or gross occlusions including those by distracting targets. Tracking is performed at approximately 27fps on a 3GHz single processor PC. We have provided segmentation and tracking results and compared its performance to that of existing tracking methods.

# Chapter 8

# Active Attention

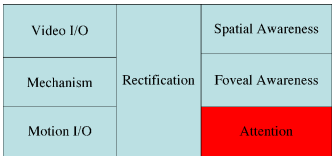



Figure 8.1: Attention.

In this chapter we develop an architecture for real-time saliency analysis of realistic, dynamic scenes using active vision. Using biological inspiration, we propose and implement an active attention framework that incorporates saliency, inhibition of return, task biasing, and moderation of covertly considered locations.

## 8.1 Introduction

We have considered biology and existing models of primate vision when selecting components of visual attention. We now incorporate active visual attention into the processing framework. We begin by implementing bottom-up centre-surround saliency cues in a manner similar to that of Itti & Koch (1998). We extend the model for use with an active vision platform by integrating the rectification and mosaicing process described in Chapter 5. We are able to use the occupancy grid to establish 3D cue-surface correspondences.

In monkeys, salient locations are retained across saccades by transferring activity among spatially-tuned neurons within the intraparietal sulcus [Merriam *et al.* (2003)]. A short-term inhibitory effect then prevents previously attended stimuli from being immediately re-attended. One reason for such a short term memory may be to help optimise search performance by inhibiting previously attended scene locations. Accordingly, our attentional system incorporates IOR to maintain an egocentric short term memory of previously attended scene locations. Further, we introduce methods to *covertly* propagate IOR in dynamic scenes according to the motion of scene objects.

As we shall see, image frame saliency is significantly affected by active camera motion. We do not select fixation locations based solely upon the saliency map. We modulate saliency by dynamic IOR bias, and a task-dependent spatial bias to obtain a fixation map. Finally, we covertly moderate fixation arbitration by accumulating evidence about the spatial coherence and strength of candidate peaks in the fixation map.

Attention is susceptible to online top-down modulation for assisting visual tasks. We incorporate attentional processing into the processing network. It is integrated with the coordinated fixation and segmentation component, and spatial awareness component, such that primate-like gaze behaviours emerge.

## 8.2 Synthesising Saliency

We begin by implementing cues known to contribute to the perception of attentional saliency in primates. As we shall see, cues are processed in real time on a network of computers.

### 8.2.1 Saliency Cues

The usefulness of cue synthesis is subject to real-time performance constraints, so cues are implemented with processor economy in mind. Pre-attentive feature computation occurs continually in primates across the entire visual field and takes around 25-50ms [Itti & Koch (2001)]. We process images in YUV colour space. Cues are processed in parallel; however, some serialisation in cue processing is required to meet cue dependencies (Figure 8.22*a*). Cue contrast is important in saliency, not local absolute cue levels [Nothdurf (1990)]. Accordingly, centre-surround spatial uniqueness in each synthetic cue is determined for incorporation into saliency perception.

Neurons at the earliest stages in the visual brain are known to be tuned to simple features like intensity contrast, colour opponency, orientation, motion and stereo disparity. These features contribute to the perception of attentional saliency. For synthetic saliency, we choose conceptually relevant and biologically plausible early visual cues including intensity and colour uniqueness, optical flow, depth and depth flow, orientation uniqueness, and collision path criticality.

#### 8.2.1.1 Intensity Uniqueness

Neurons tuned to intensity centre-surround produce a response that can be synthesised using a DOG approximation [Itti & Koch (2000)]. In a manner similar to Ude *et al.* (2005), we create a Gaussian pyramid from the intensity image. Successive images in the pyramid are down-sampled by a factor of two ($n$ times), and each is convolved with the same Gaussian kernel. To obtain DOG images, the Gaussian pyramid images are up-sampled (with bilinear interpolation) to the original image size and then combined. Combination involves subtracting pyramids at coarser scales $C_n$ from those at finer scale $C_{n-c}$. We consider two levels of interaction, immediate neighbours $C_n - C_{n-1}$, and second neighbours $C_n - C_{n-2}$, to obtain a DOG pyramid with $n-3$ entries. Finally, the $n-3$ entries are added to obtain a map where the most spatially unique region emerges with the strongest response. The borders of the image equate to an edge that would otherwise produce a significant step response in uniqueness computations, due to edge effects of convolution. Prior to conducting convolutions, images are padded with zeros beyond the image frame, and a smooth transition to zero response within

Figure 8.2: Example windowing function applied to images before centre-surround processing. This prevents step responses elicited by the image edges.

Figure 8.3: Intensity centre-surround uniqueness.

the image frame is enforced using a windowing function (Figure 8.2).

#### 8.2.1.2 Colour Uniqueness

Colour channels are sent to a separate server for processing in parallel with intensity information. Colour centre-surround uniqueness is computed as per intensity. In the retina, some ganglion cells produce a red-green centre-surround response, others exhibit the orthogonal blue-yellow centre-surround response. We process orthogonal U and V chrominance opponents (U is approximately a yellow-magenta response and V approximates an orthogonal cyan-pink response), and combine the centre-surround responses by addition. In this manner, the region

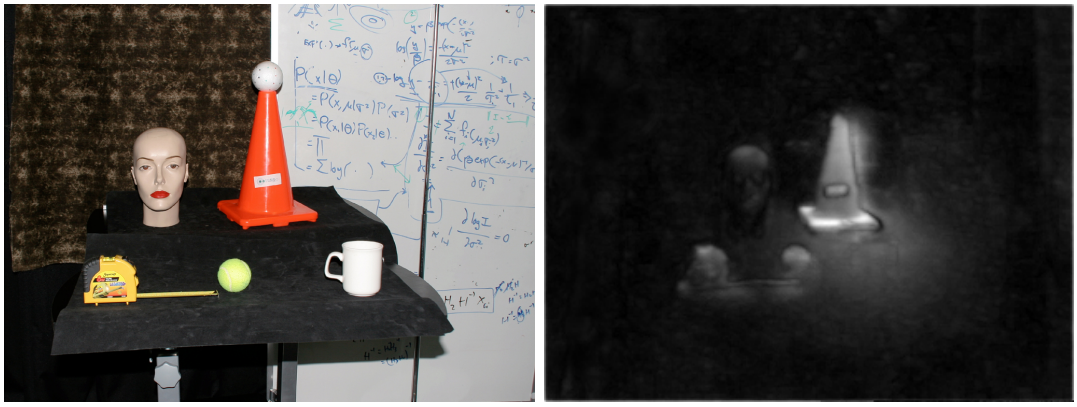


Figure 8.4: Colour centre-surround uniqueness.

with the most unique colour chrominance emerges with the strongest response. Colour uniqueness is calculated for both left and right image feeds at full frame rate. Figure 8.4 shows output for the colour uniqueness response.

#### 8.2.1.3 Chrominance Distance

Specific target chrominance(s) $T = (u, v)$ can be selected for detection in the left and right image views. At every pixel location in each image, the normalised Mahlonobis distance $|\Delta C|$ from the image pixel chrominance to the target chrominance is computed according to:

$$\Delta C| = \frac{\sqrt{(I_u - T_u)^2 + (I_v - T_v)^2}}{\Delta C_{max}}, \tag{8.1}$$

where $I_u$,$I_v$ and $T_u$,$T_v$ are the U,V colour chrominance components at the pixel location, and the sought target chrominance respectively. $\Delta C_{max}$ is the maximal possible distance from the sought chrominance in the 255 level U,V colour space:

$$\Delta C_{max} = \sqrt{(M_u)^2 + (M_v)^2}, \tag{8.2}$$

where $M_u = max(255 - T_u, T_u)$, $M_v = max(255 - T_v, T_v))$.

So that only chrominance distance responses in the vicinity of the sought chrominance are retained, we convert the measured chrominance distances to a chrominance similarity measure $S_c$ using a Gaussian lookup function:

$$S_c = \frac{e^{-|\Delta C|^2}}{2\sigma^2}, \tag{8.3}$$

where $\sigma$ is selected such that $3\sigma$ approximates the maximum passable chrominance distance.

For example, if the chrominance at an image location is a Mahlonobis distance of 0.0 from the sought chrominance, the chrominance similarity measure $S_c$ would be 1.0. As the chrominance distance increases, the modulation factor tends towards $3\sigma$, and $S_c$ correspondingly tends towards 0.0, according to the above Gaussian function. The method therefore passes pixels in the vicinity of the sought chrominance only.

#### 8.2.1.4 Optical Flow

The translation from the current to previous frame for each camera is known in mosaic coordinates. The rectification and mosaicing process removes the view-frame effect of any encoded camera geometry changes (pan, tilt). Once the location of the current and previous frame in the mosaic for each camera is known, we calculate optical flow only on the overlapping region of consecutive view frames in the mosaic. This process allows estimation of horizontal and vertical scene flow independent of the motion of the cameras (rather than flow relative to the camera image frame). A *sum of absolute differences* (SAD) flow operation [Banks & Corke (1991)] is used. We obtain four maps from the two cameras: horizontal and vertical flows in each camera. The responses are normalised and centre-surround uniqueness is determined for all four maps. In this manner, regions in view that are moving in a unique manner are extracted. We down-sample images before computing flow for processor economy. Figure 8.5 shows sample horizontal flow estimation.

#### 8.2.1.5 Disparity

The epipolar rectified mosaics allow us to search for pixel disparities along horizontal scan-lines only. We search the neighboring $\pm 16$ pixels in the second image for a correspondence to the candidate pixel location in the first image. We conduct a SAD disparity search in the overlapping region of current left and right frames only. Figure 8.6 shows sample disparity map output.

We consider that closer objects are more salient because they are more likely to interact with the apparatus. Therefore depth without any centre-surround

Figure 8.5: Optical flow, horizontal direction. The hand (light) moves left, the body (dark) moves right. Black areas represent those where no flow estimate is obtained.

Figure 8.6: Disparity cue. Left and right input images, and resulting disparity map (respectively).

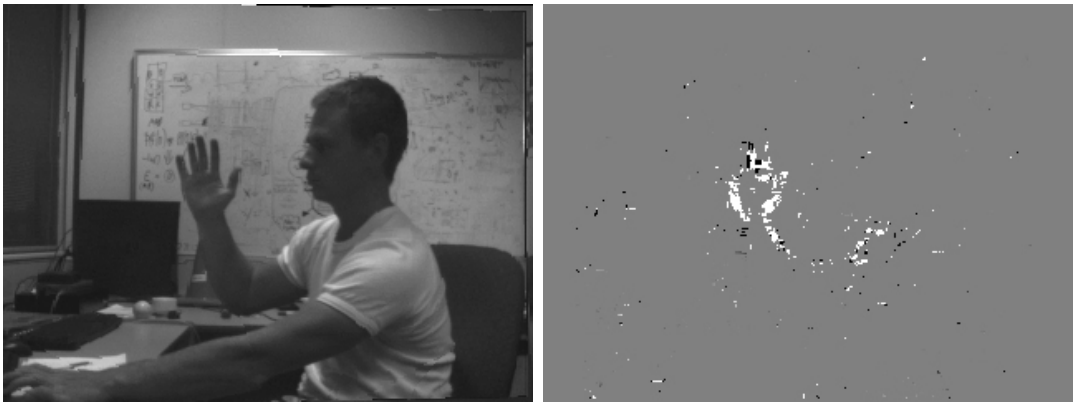


Figure 8.7: Depth flow cue. The hand moves towards cameras. SAD performance is best in textured regions, hence the response is sparse on the bland palm.

modulation is itself a saliency cue. In addition, centre-surround modulation is determined because a region that exhibits a different depth to its surroundings is also considered salient – it is likely to be an object or obstacle. The two maps are combined into one by weighted addition.

#### 8.2.1.6 Depth Flow

The velocities of visual surfaces in the depth direction are calculated using an approach similar to that of Kagami *et al.* (2000), by considering the overlapping regions of consecutive disparity maps. The centre-surround uniqueness algorithm is applied to the depth flow output. Figure 8.7 shows sample depth flow output.

#### 8.2.1.7 Orientation Uniqueness

Eye trackers have been used to observe that humans preferentially fixate upon regions with multiple orientations [Zetzsche (1998)]. A winner-take-all competition is activated amongst neurons tuned to different orientations and spatial frequencies within one cortical hypercolumn Carrasco *et al.* (2000). These observations suggest that responses to different orientations may be computed in parallel, somewhat separately, with integration and spatial competition occurring at the later stages.

We achieve a synthetic response using complex log-Gabor convolutions over multiple scales within each of the multiple orientations. The log-Gabor response

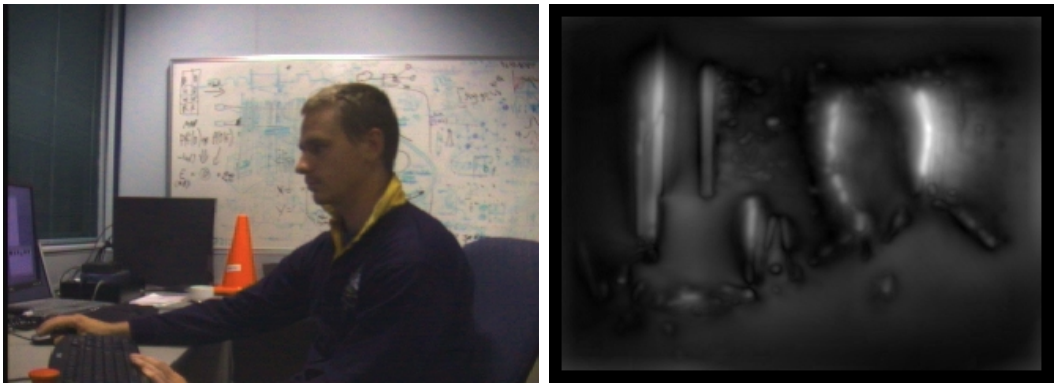


Figure 8.8: Orientation cue response, horizontal direction only.

models the impulse response observed in the orientation sensitive neurons in cats Sun & Bonds (1994). The log-Gabor kernel provides a broader spatial frequency response than the Gabor kernel, so fewer scale convolutions are necessary for the same spatial sensitivity. We compute the convolutions in Fourier space and obtain orientation response maps for each orientation and scale. Within each orientation, we sum all scale responses (Strong local interactions between separate orientation filters have been characterised via neuronal correlates [Itti & Koch (2000)]). Processing each orientation is a heavy operation, and because we have four virtual CPUs per processing node, we limit the operation to four orientations per camera. The associativity of convolution means that the subsequent orientation uniqueness operation (involving a series of convolutions) need not be done for each orientation separately. We can simply sum the orientation maps, and apply the centre-surround uniqueness operation to the result. We obtain orientation response maps for each orientation, a single map of the regions that respond to the most orientations (such as corners and edges, Figure 8.8), and an orientation uniqueness map (Figure 8.9) where the strongest response occurs at regions that contain orientations atypical to the rest of the image, regardless of scale.

#### 8.2.1.8 Critical Collision Cue

The critical collision cue responds to pixels on visual surfaces in the scene that are on an instantaneous trajectory leading towards the visual aparatus. A similar neural response has been observed in pigeons [Wylie *et al.* (1998)]. At each pixel

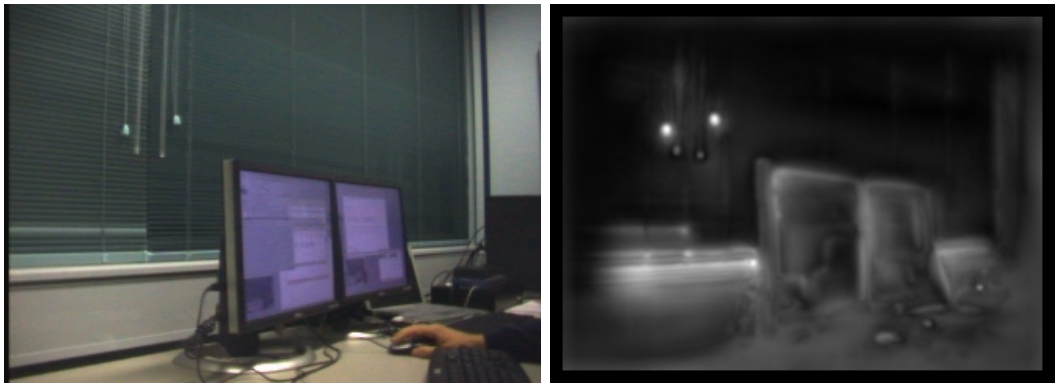


Figure 8.9: Orientation centre-surround uniqueness. The multiple orientation responses of the two bright dots stand out from the predominantly horizontal orientation of the blinds.

where the required measurements exist and are valid, we obtain a position vector $p = (x, y, depth)$ and a velocity vector $v = (flow_x, flow_y, flow_{depth})$. We obtain the collision criticality cue according to:

$$\frac{||p||}{||v||}(1 - (-nv \cdot np)), \tag{8.4}$$

where the dot represents the dot product, and $nv = v/|v|$, and $np = p/|p|$ are unit vectors. That is, the component of the velocity vector associated with a scene point in the direction of the negative distance vector to that scene point is calculated and modulated by the time ($||p||/||v||$) the scene point would take to get to the origin (the midpoint between the cameras) if it were to maintain the current trajectory. The calculation therefore highlights the scene areas whose trajectories are presently likely to collide with the vision system, and weights them according to which will collide first.

### 8.2.2 Cue Processing

Interdependencies exist in the extraction of cues. So that calculations are not conducted unnecessarily multiple times, cues that can incorporate intermediary maps calculated during other feature computations are serialised within a single processing node.

Figure 8.12 shows cue interdependencies. Serialisation of cue computation can

Figure 8.10: Critical collision cue. The head is moving towards CeDAR. Disparity and flow estimates on the sides of the head (response is better in textured regions) elicit a critical collision cue response.

be read off the graph. For example, the collision criticality cue depends on rectification, depth, flow, and depth flow ordered serial pathway. Such serialisation must be preserved in the processing network implementation. We therefore incorporate parallel and serial node processing as per Figure 8.11. One node (that we have named the $DFCS$ – depth flow centre surround – server) receives Y channel rectified images from both cameras, and processes intensity centre-surround, disparity, optical flow, depthflow, and the critical collision cue. This is because it is a very serialised pathway that operates on the same Y channel input. The node is a virtual quad processor computer, so it is able to parallelise much of the processing of these cues. For example, four optical flow maps can be computed simultaneously, corresponding to the left and right x and y flow.

Another two nodes (the $OCS_l$ and $OCS_r$ – left and right orientation centre surround – servers) also receive Y channel rectified images exclusively for orientation processing, a heavy process. The last cue processing node (the $CCS$ – colour centre surround – server) receives U and V colour chrominance channels from both cameras (4 channels) and processes centre-surround chrominance maps on these channels in parallel on its four virtual CPUs.

We combine centre-surround cues in a fashion similar to the winner-take-all method [Itti & Koch (2000)]. On each cue processing node, the outputs are weighted and combined into a single map (except the DFCS server, where three

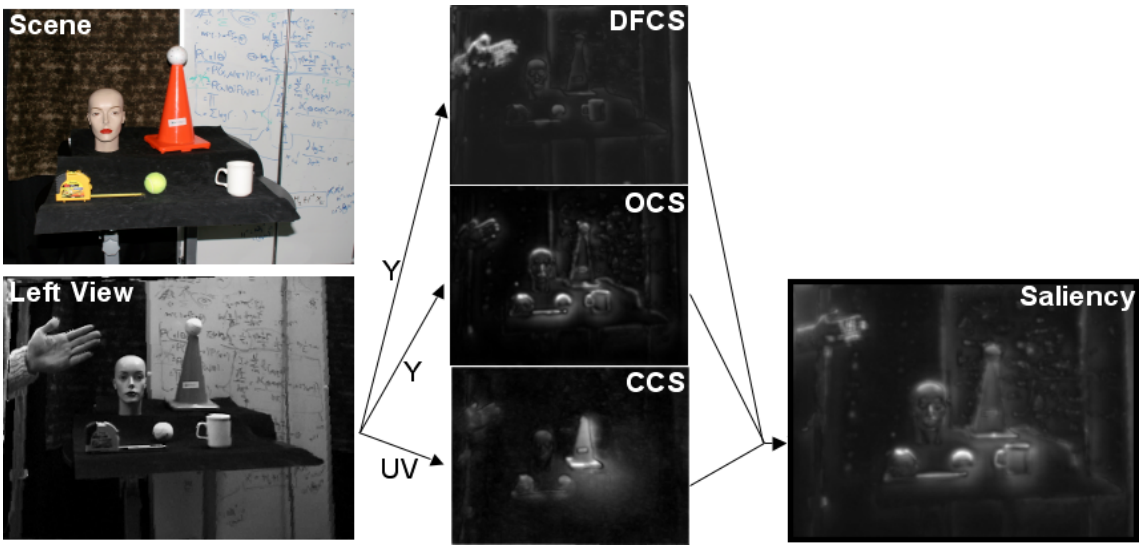



Figure 8.11: Processing node outputs are combined to contribute to saliency.

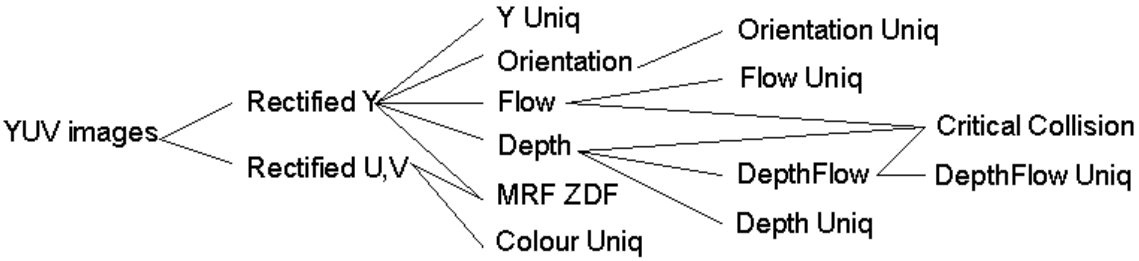



Figure 8.12: Synthetic cue dependencies.


Figure 8.13: Online perception of saliency demonstration (*snapshot - see Appendix C for full video*).

maps are distributed: left, right and stereo output) for distribution to a client node. A client node receives all outputs from the four cue processing nodes, applies weighting, and adds them into a single saliency map for each camera, and a stereo saliency map (Figure 8.11 shows combination into a single camera saliency map).

## 8.3 Active-Dynamic Attention

In monkeys, salient locations are retained across saccades by transferring activity among spatially-tuned neurons within the intraparietal sulcus [Merriam *et al.* (2003)]. Mechanisms of spatial updating maintain accurate representations of visual space across eye movements. Navalpakkham *et al.* hypothesise that because neurons involved in attention are found in different parts of the brain

that specialise in different functions, they may encode different types of salience: they propose that the posterior parietal cortex encodes a visual salience map; the pre-frontal cortex encodes a top-down task relevance map; and the final eye movements are subsequently generated by integrating information from both regions to form an attention guidance map possibly stored in the superior colliculus [Navalpakkam *et al.* (2005)].

Our implementation introduces three intermediary maps such that IOR can be dynamically and *covertly* propagated in dynamic scenes. The three maps include a Bayesian saliency mosaic, an IOR mosaic, and a TSB mosaic. The three maps are maintained in an egocentric mosaic reference frame so that they are not affected by deliberate pan and tilt camera motion, and such that the saliency/inhibition/bias of the current view can be related to previous views.

### 8.3.1 Bayesian Saliency Updates

We incorporate saliency maps into a saliency mosaic using the Bayesian update equation. For each camera, the response level of each pixel in each centre-surround cue for each image is used to increment the probability that the corresponding pixel in the mosaic is salient. Let $s[x, y]$ denote the cue response at pixel location $[x, y]$. Given a cue response measurement $M$ at a pixel $[x, y]$, we use the incremental log-likelihood form of Bayes' Law [Elfes (1989)] to update the saliency map at each pixel. We introduce cue weight $W_c$ corresponding to an empirical weighting of the cue compared to all other cues:

$$\log L(salient) \leftarrow \log L(M \mid salient) + W_c \log L(s) \tag{8.5}$$

Log-likelihoods provide an efficient implementation for incorporating new data into the saliency map by reducing the update to an addition. Gain $W_c$ may be autonomously updated by higher-level operations, representing top-down modulation. In this experiment, the cue weights are declared empirically and remain static.

All entries in the Bayesian saliency map are decayed over time, so that a permanent perception of salience is not anchored to previously attended regions. This decay rate ($S_d$) affects how easily the system's attention can be distracted. As with other control parameters, rate $S_d$ can be modulated by higher-level pro-

cesses, depending on the level of concentration required for a particular task. The decay rate also prevents the saliency grid implementation from saturating.

### 8.3.2 Dynamic Inhibition of Return

IOR represents the notion that once we have assessed a particular point or object in a scene, we are less inclined to look there again. For example, a green apple amongst red apples is considered visually salient. If the apple is then moved to a pile of other green apples, it becomes less visually salient. In dealing with dynamics we therefore do **not** propagate saliency. Instead, we deal with dynamics within the IOR map. We initially find the green apple salient when it is amongst red apples, so it is likely to get attended. When the apple moves to its new location, it is still the same object that we previously attended. However, if other interesting events are now occurring in the scene, we might not justify directing foveal attention towards the same green apple again. We might prefer to direct our attention to other as yet unevaluated scene locations. We therefore covertly propagate *suppression* of the saliency of the green apple as it moves, and it remains suppressed when it moves to the pile of green apples. Conversely, a green apple amongst green apples is not visually salient. When the green apple is moved to a pile of red apples, it is still the same green apple that was previously not salient, and was previously not attended or suppressed. When it moves to the pile of red apples, it is considered salient, is not suppressed, and therefore may well win attention. This example helps to express why we covertly *remember* (at least in the short term) and propagate the location of previously attended scene regions in the IOR map only.

The system evaluates IOR every frame. A Gaussian kernel is added to the region around the current fixation point in an IOR accumulation mosaic, every frame (Figure 8.14). The radius of the Gaussian kernel can be modulated according to preference. Expanding upon this for dynamic scenes, accumulated IOR is propagated according to the estimated current optical flow. In this manner, IOR accumulates at attended scene locations, but it remains attached to objects as they move. In propagating IOR, it is spread and reduced according to Gaussian uncertainty in the region's new location.

We decrement the entire IOR mosaic over time according to decay rate $I_d$,

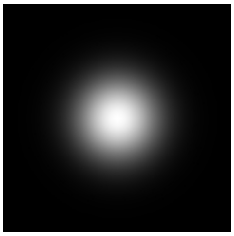


Figure 8.14: Gaussian IOR increment pattern. This kernel is applied at the coordinates of the centre of each view frame in the IOR accumulation mosaic (centre, Figure 8.20).

so that previously inhibited locations eventually become uninhibited. As with saliency decay rate $S_d$, faster $I_d$ decay means more frequent saccades to distractors around the scene. Again, this rate can be modulated by higher-level operations, though we declare it empirically. IOR may suppress an attended object's saliency, but if the object then moves it is not immediately salient (other than by the additional saliency elicited by its motion, but existing IOR usually suppresses this beyond causing a fixation map peak) because it carries its inhibition of salience with it. Its effective saliency continues to be suppressed until the IOR decay rate, or the uncertainty associated with its location, reduces the IOR suppression of its saliency. In this manner, IOR is a retrospective response as it depends upon previous observations. For a given head pose, the mosaic reference frame remains static with respect to the world, and as such, regions of the mosaic not in the current view frame may remain suppressed until inhibition is completely decayed or until that location is next attended and inhibition increases.

Before gaze arbitration, saliency is first modulated by IOR (and then TSB). Figure 8.17 demonstrates the interaction between dynamic IOR and saliency. It shows how inhibition becomes "attached" to the surfaces in the scene, propagating with those surfaces if they move, according to optical flow. Figure 8.18 shows a demonstration movie of this process.

#### 8.3.2.1 Task Dependent Spatial Bias

The prefrontal cortex implements attentional control by amplifying task-relevant information relative to distracting stimuli [Nieuwenhuis & Yeung (2005)]. We introduce a TSB mosaic (Figure 8.19) that can be dynamically tailored according

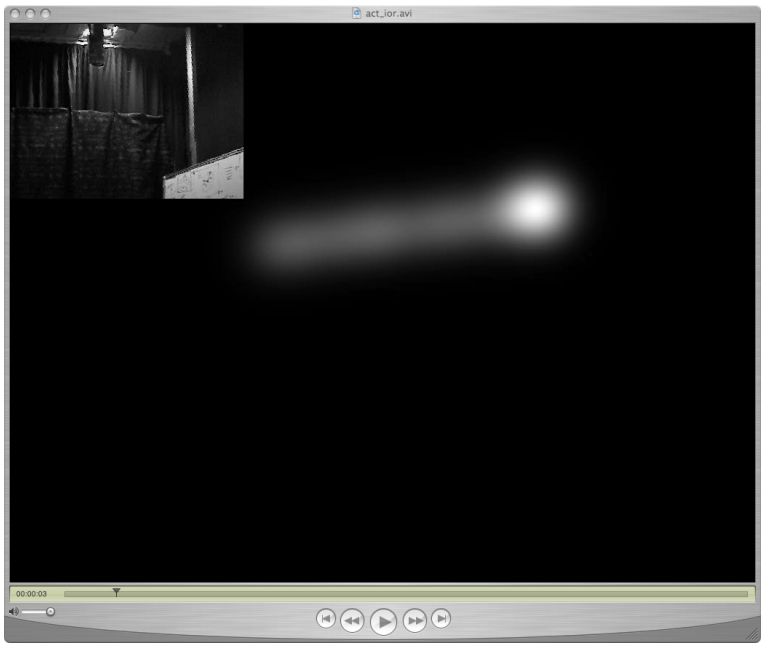



Figure 8.15: Online distribution and accumulation of IOR demonstration (*snapshot - see Appendix C for full video*).

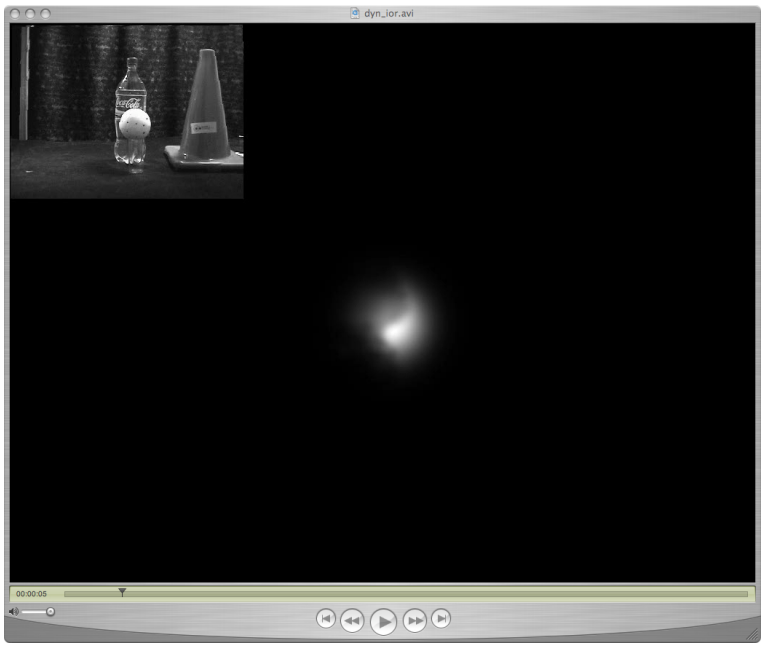



Figure 8.16: Online dynamic accumulation and propagation of IOR according to optical flow demonstration (*snapshot - see Appendix C for full video*).

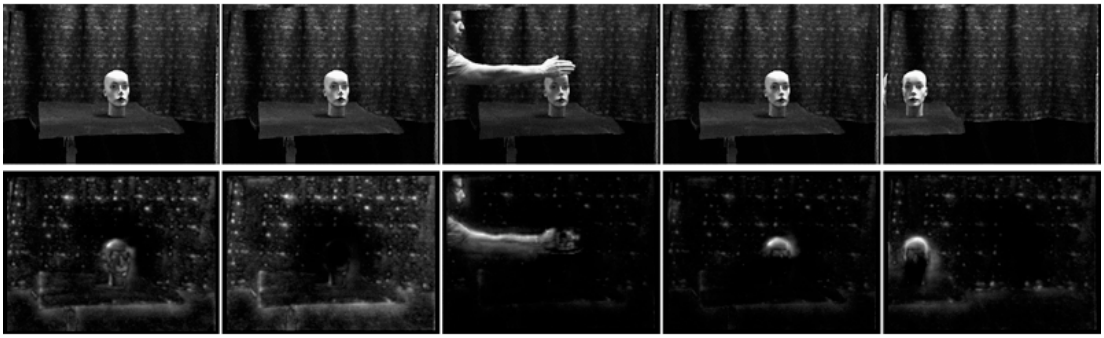


Figure 8.17: Dynamic IOR. From left: *1* the head on trolly moves into fovea, initially uninhibited; *2* after time it becomes inhibited; *3* a salient hand enters fovea; *4* IOR on forehead is reset by occlusion; *5* trolley and head move out of fovea, taking associated IOR pattern.


Figure 8.18: Online demonstration of the effect of dynamic IOR on saliency (*snapshot - see Appendix C for full video*).

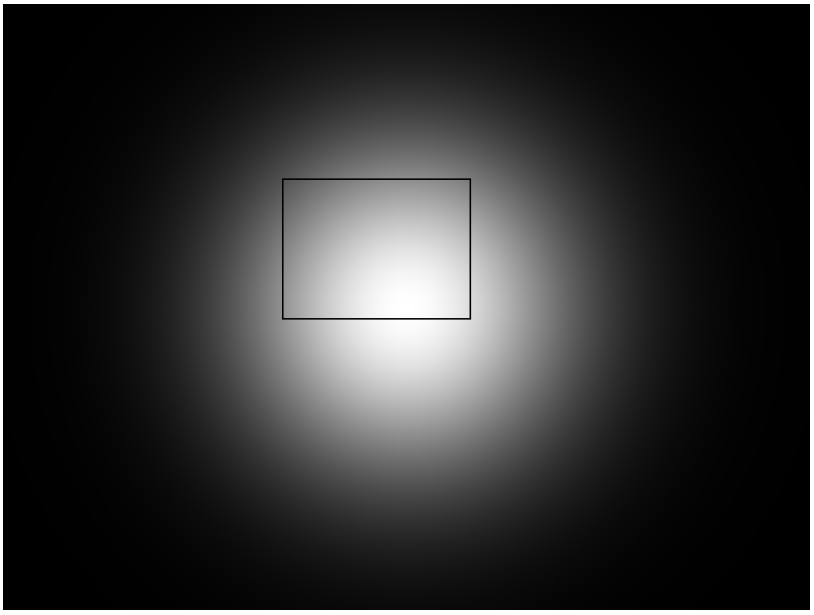


Figure 8.19: Sample TSB mosaic showing current view frame position. This radial TSB could represent a forwards search task – the gradient across the view frame induced by the radial TSB enhances saliency of scene surfaces towards the centre of the mosaic. Like all mosaics, the egocentric TSB mosaic remains static with respect to the scene despite camera motions.

to tasks. For example, if we are driving a car, we know that we should tend to keep our gaze upon the road, and as such we bias the lower half of the mosaic where we would expect to find the road. For a forwards search task, we might like to use a radial TSB, such that the system does not tend to divert its gaze too far away from forwards. The TSB may be dynamically updated as appropriate for the current task. The TSB can be preempted for regions not in the current view frame. Covert attention involves consideration of factors not directly associated with the current target at fixation. It involves consideration of regions towards or beyond the periphery, whether real, expected, or hypothetical. In this manner, TSB is potentially a form of covert attention.

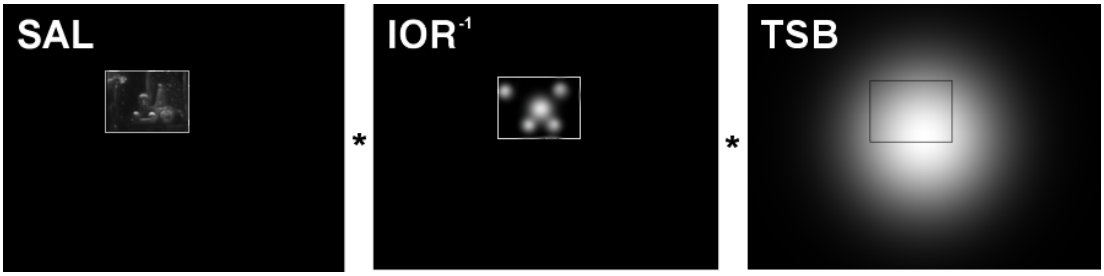



Figure 8.20: The fixation map is the product of Bayesian saliency, dynamic IOR (inverse shown), and TSB.

### 8.3.3 Fixation Map

It is now known that the prefrontal cortex implements attentional control by amplifying task-relevant information rather than inhibiting distracting stimuli [Nieuwenhuis & Yeung (2005)]. To achieve fixation upon salient regions in dynamic scenes with moving cameras, we modulate (multiply) the saliency map by the IOR and TSB maps. This process amplifies relevant visual stimulus. Figure 8.20 shows the components of the fixation map.

### 8.3.4 Gaze Selection and Target Pursuit

In its simplest form, gaze can be directed towards the scene location corresponding to the single maximal peak of both left and right fixation maps. However, as gaze changes so too does cue spatial uniqueness. Ignoring the effect of IOR, attending a new location may immediately render a previously non-salient location salient. This can result in an overly saccadic system. We therefore moderate the winning locations before the winner of fixation is selected. Again, this involves consideration of regions other than that currently receiving overt attention. In this manner, moderation of fixation maxima is potentially a form of covert attention.

We define three modes of moderation:

- **Supersaliency**: a view frame coordinate immediately wins attention if it is $n_s$ times as salient as the next highest peak.

- **Clustered Saliency**: attention is won by the view frame location about

which $n_c$ global peaks occur within $p$ consecutive frames.

- **Timeout**: if neither of the above winners emerge in $t$ seconds, attention is given to the highest peak in the fixation map since the last winner.

Both the left and right fixation maps are scanned to determine the maximal peak. The first peak location that passes moderation is selected as the next gaze fixation point. We call the fixation map and image in which the maximal peak was found, the *primary fixation map* and *primary image* respectively. Using the output of the disparity cue it is possible to find the corresponding image locations of the winning scene location in the *secondary image*. It is noted that the disparity map is simply an estimate of the shift in the pixel location of the projection of the same scene points from one view to the other – it can therefore be used to cross-reference the location of pixels from the left and right views. This process largely eliminates the need to *search* for the corresponding location in the secondary image. The peripheries of the left and right cameras may contain different visual stimulus (the entire visual fields do not usually overlap entirely). In this instance, or if no disparity data is available at the primary image location, a template search is initiated. The template search is conducted over a minimal region in the vicinity of the peak location in the primary image (camera vergence does not usually deviate more than several degrees from parallel). The camera images are parallel epipolar geometry rectified which means that the template search need only be conducted along horizontal scanlines. A small template around the selected fixation point is sought in the secondary image, the starting (and most likely) location for the search is determined by cross referencing the disparity cue. Once the coordinates of the winning location are found in the secondary view, saccade is initiated. If coordinates are not found in the opposing view using the template search, then the disparity map is used to cross reference the location from the winning view, and saccade is nonetheless initiated. In both instances, immediately after saccade, the MRF ZDF node fine-tunes gaze such that the surface that initiated saccade is fixated upon in a coordinated manner (or the nearest surface). It is noted that is is also possible to use the disparity map correspondences to integrate the left and right saliency maps into the stereo saliency map, creating a single saliency map (not only saliency – *any* left and right cue maps may be unified in this manner). However, such a unified map would

require more rich and accurate disparity estimations, so we maintain separate maps.

#### 8.3.4.1 Before and After Attentional Saccades

During an attentional saccade, much motion blur is induced in the contents of the camera images. This blur temporarily reduces image quality and affects cue processing. In particular, the optical flow and disparity cues become excessively noisy. This noise can be misinterpreted in centre-surround processing as saliency. Excessive noise in optical flow calculations can also affect the propagation of dynamic IOR. To overcome this problem, the gaze moderation process broadcasts to the relevant processing nodes that a saccade is about to occur. Then, during the saccade, processing nodes can take appropriate action. For example, the propagation of IOR according to flow does not occur, and Bayesian saliency does not accumulate. Further, the MRF ZDF thread suspends sending tracking commands to the motion axes. Interestingly, there exists a similar mechanism in biology that may serve a similar function: in neural recording studies with monkeys, scientists found that they could predict the occurrence of saccades by monitoring the activity of certain neurons [Sugrue *et al.* (2005), Dorris *et al.* (1997)].

Immediately after attentional saccade completion, tracking control is returned to the MRF ZDF process. This centres gaze upon the target that initiated an attentional saccade, ensuring coordinated fixation upon, and smooth pursuit of, the target.

#### 8.3.4.2 Permitting Top-down Bias

Control of various attentional components can occur on-line. Cue weightings can be sent to any one of the cue processing nodes to increase the contribution of that cue to the saliency map. This may be particularly useful in feature-gate style search. Similarly, the weighting and layout of the TSB may be updated as desired. IOR modulation may also occur in the form of modifying the accumulation and decay rates, as well as the radius of the IOR Gaussian kernel. The fixation moderation parameters may also be changed online.

We can bias the system for specific tasks. For example, by weighting the

colour chrominance distance cue heavily, and selecting a skin-coloured target chrominance, the system could be made to preferentially attend to hands and faces, but is still attentive to other distracting stimuli. Similarly, we experimented with biasing the system to attend to the road, road signs and road lines in the road scene. While preferentially "*keeping its eyes on the road*", the system briefly evaluates other salient events in the road scene.

Similarly, we can affect the system's visual behaviours. For example, we can make the system more saccadic by increasing the IOR decay rate, decreasing the Gaussian kernel accumulation rate and radius, and relaxing moderation strictness of fixation selection.

## 8.4 Integration into Processing Network

We have not required any particular model or structure when distributing processing tasks over nodes in the network. The only requirement is real-time performance. For this reason, we serialise processing such that cues are only determined once in the processing network, despite being used multiple times throughout the processing structure. For example, spatial representation in the occupancy grid node requires depth information from the disparity cue. Disparity is also a cue used for the perception of saliency. Rather than being calculated multiple times, disparity is calculated once only in the DFCS node, and the output is distributed to subsequent nodes that require this map. The structure of the distributed processing system has emerged only from bandwidth minimisation and optimisation of system performance.

### 8.4.1 Functional Structure

We adopt a client-server architecture to allow concurrent serial and parallel processing. At the lowest level, a rectification server distributes rectified images and rectification parameters to dependent nodes. Biological evidence suggests that colour opponents are treated in separate channels in the brain to intensity [Dacey (1996)]. U and V colour chrominance images for both the left and right images are sent to the colour centre-surround ($CCS$) server for processing. Y channels are sent to the orientation ($OCS_L, OCS_R$) servers and the depth and

flow ($DFCS$) server. To minimise network bandwidth, to cope with the processing load of each frame, and to prevent repetition of computations, nodes in the structure are configured simultaneously as clients of processes preceding them in cue serialisation (Figure 8.12), and as servers to nodes following them. Each node is a dual CPU hyper-threaded 3GHz PC with four virtual processors. Trade-offs exist between splitting tasks into sub tasks, passing sub tasks to additional nodes, and minimising network traffic. The best performing solution involves grouping serialised tasks on each server, and performing as many operations on the same image data on the same server as possible, so there is minimal CPU idle time and minimal network traffic. The serial nature of cue computations means there is often no gain possible in distributing the task – in fact further network transfer of data between servers would slow performance. Figure 8.21 is a block diagram summarising data flow occurring between each node in the processing network.

Figure 8.22 shows a broad summary of the major feed-forward interactions in the primate visual brain. Figure 8.23 summarises feed-forward interactions in the synthetic vision system. It is noted that the synthetic structure bears a good resemblance to the broad interactions between visual centres in the primate brain. Loose analogies can be drawn between the image acquisition and transfer functions of the video server, and that of the retina and the pathways through the lateral geniculate nucleus (loose in the sense that centre-surround is calculated within the human retina whereas synthetic DOG maps are currently calculated shortly after rectification). Similarly, the motion control server controls motion in a manner somewhat analogous to the function of superior colliculus. The global representation of space across saccades that occurs in V3, V5 and the inferior parietal lobe performs functions similar to that of the rectification server and occupancy grid representation; the saliency server processes cues in a manner analogous to the inferior parietal lobe (dorsal stream functions). The MRF ZDF server extracts attended objects, potentially for identification, in a fashion analogous to the recognition and identification functions of the infero temporal lobe (ventral stream functions). Both streams rely upon early visual cues in both the synthetic and primate models. The orientation, depth and flow, intensity and colour cue processing functions are loosely analogous to early cue processing occurring in early brain areas V1, V2 and V3. At the highest level, a client process modulates relative cue weightings and updates spatial biasing according

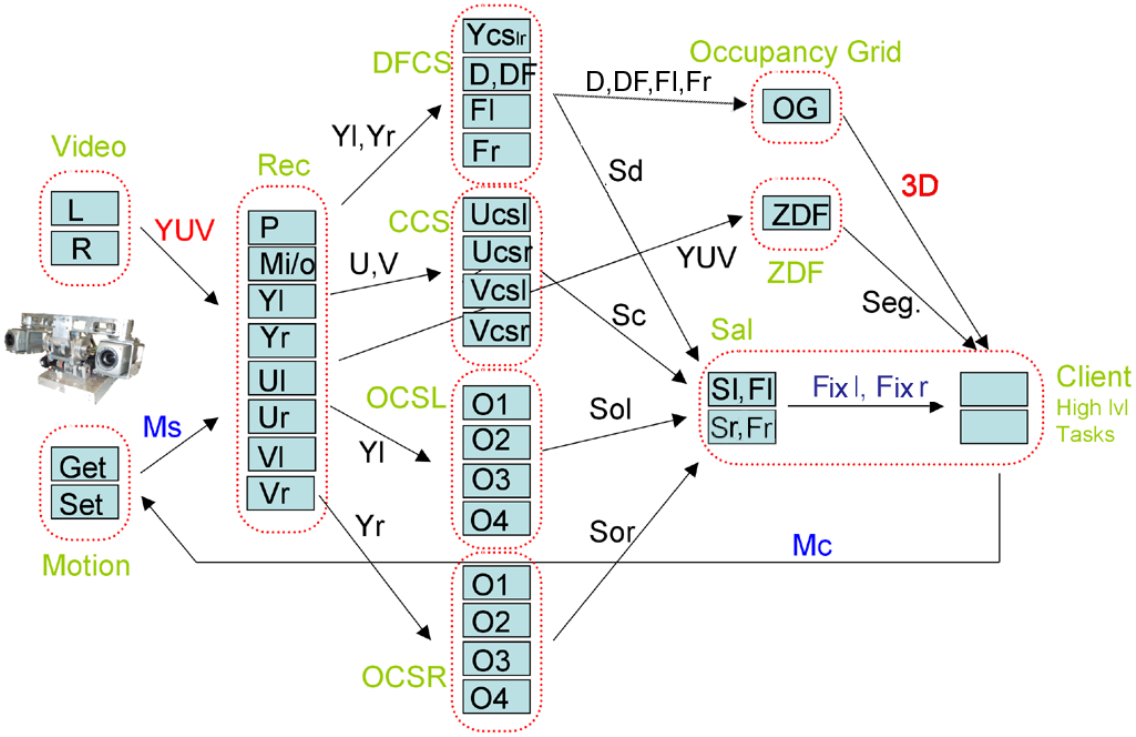



Figure 8.21: Synthetic system block diagram - servers and I/O. The dotted lines surround physical PCs. The boxes show processing threads. The arrows show the major data flows: motion status (Ms), motion commands (Mc), saliency maps (Sd, Sc, Sol, Sor), fixation maps (Fixl, Fixr), target segmentation (Seg.), occupancy grid data (3D), motion cues (D, DF, Fl, Fr), and original camera image channels (Y, U, V).

to the desired task, which are functions generally considered to occur within the prefrontal cortex. Modulation feedback pathways, such as the ability of the prefrontal cortex to modulate neuronal responses in V1 (or the ability for the client process to modulate cue weightings) have been omitted from the diagrams.

## 8.5 Results

The synthetic vision system preferentially directs its attention towards non-suppressed salient objects/regions. Upon saccading to a new target, the MRF ZDF process extracts the object that has won attention, maintaining stereo fixation on that

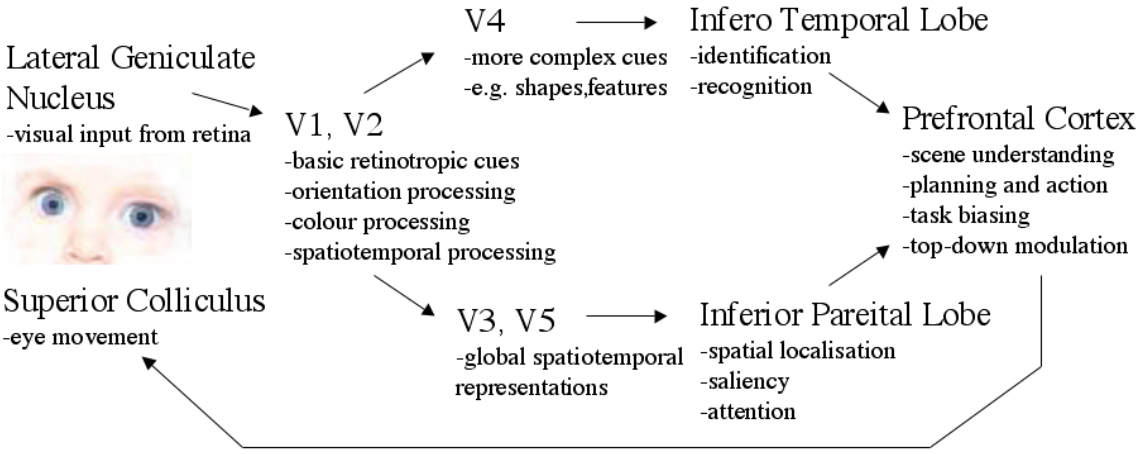



Figure 8.22: Broad interactions in primate visual brain.

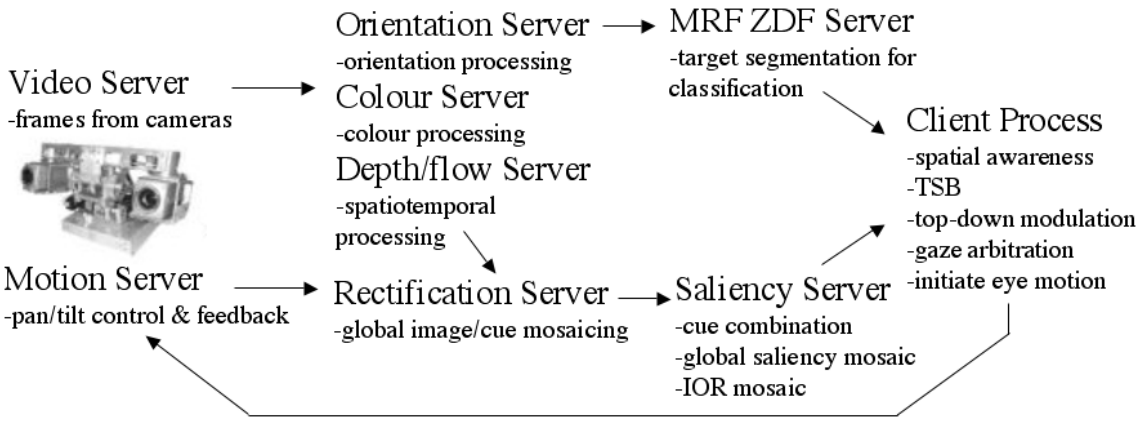



Figure 8.23: Interactions in synthetic vision system.

object (smooth pursuit), regardless of its shape, colour or motion. Attention is maintained until a more salient scene region is encountered, or until IOR allows alternate locations to win fixation (Figure 8.24). If something is comparatively very salient, it is tracked by the MRF ZDF process until its saliency has been reduced by IOR. If several locations have similar saliency (such as when nothing in the scene is being actively moved), attention shifts more frequently amongst those locations. The demonstration movies corresponding to Figures 8.25 and 8.26 show sample functionality of the complete system.

Figure 8.24: Sample system behavior. Each column shows camera image, fixation map and MRF ZDF segmentation. From left: *1* attention shifts to head from inhibited base of cone, forehead is segmented from background in fovea; *2* attention returns from inhibited head to top of cone, cone is segmented; *3* attention shifts from inhibited cone to mug, mug is segmented.

Figure 8.25: Online dynamically updated fixation map demonstration (*snapshot - see Appendix C for full video*). The insets show the rectified camera image (left) and selected target extraction (right).

Figure 8.26: Online system demonstration showing image mosaicing (*snapshot - see Appendix C for full video*). The insets show the fixation map (left); target extraction (centre); and the active head (right).

### 8.5.1 Processing Performance

As the results demonstrate, the system enables gaze arbitration in cluttered scenes. The overall frame rate achieved depends only upon the frame rate of the bottleneck (slowest) system component. System latency depends on network transfer speed, and is usually kept to a few frames.

The frame rate performance of the bottleneck component can be altered by selecting processing settings that increases throughput at the expense of quality. For example, a reduction in processing orientations reduces the resolution of orientation responses, but increases the processing frame rate.

Similarly, other performance parameters (such as the maximum detectable optical flow, maximum MRF ZDF smooth pursuit tracking velocity, maximum discernable depth resolution, etc) have been set according to the trade-off between processing performance and quality. These parameters are non-rigid as it is usually possible to improve the performance of one component of the system. However, this may come at the expense of quality or at the expense of other system components. This is due to the limitation of processing resources. Nevertheless, the processing performance of all system nodes is typically such that the system is able to react to real visual stimulus and novel events in a timeframe commensurate to the rate of change of the environment – that is, in real time.

Implementation of the system provides insight into what capabilities may be achieves on a synthetic processing network. The low-latency real-time performance of the system indicates the feasibility of such a system. This performance, and the flexible nature of the processing network, permits extensive system expansion for future additional processing tasks.

### 8.5.2 Discussion

By specifying system properties similar to those observed in nature, we have developed a synthetic active visual system capable of detecting and reacting to unique and dynamic visual stimuli, and of being tailored to perform basic visual tasks. By implementing biologically plausible early visual cues, system able to actively divert its attention to salient regions of real scenes in real time. The specific processing algorithms may not (and probably do not) reflect what actually happens in the primate brain. Active rectification provides egocentric spatiotem-

poral visual perception. A foveal MRF ZDF algorithm permits attended object tracking and extraction and ensures coordinated stereo fixation upon visual surfaces. Attention and active-dynamic IOR means that a short term memory of previously attended locations can be retained to influence attention retrospectively. Spatial and cue biasing based on observations and prior knowledge allow preemptive top-down modulation of attention towards regions and cues relevant to tasks. Covert consideration of potential saccade destinations before overt attention is deployed provides attentional moderation.These features result in a reactive vision system and the emergence of primate-like attentional behaviors.

In designing the components of the system, we have engineered methods to approximate functions occurring in the primate brain. It is undoubtedly possible to incorporate further biological inspiration in the design of components. The log-Gabor frequency analysis of orientation, for example, synthesises the response of orientation sensitive neurons in the primary visual cortex. The SAD method to determine image disparities does not synthesise neural activity so closely[1].

Although the system presented is perceptive to novel visual events, the addition of further biologically plausible cue competencies (with little extra computational requirements) could benefit the synthetic system in terms of attention, search, and the ability to perform more complex cognitive tasks. For instance, instead of using a Gaussian IOR accumulation kernel, the system may also benefit from using the output of the MRF ZDF target segmentation mask as a target-specific IOR accumulation mask. The target segmentations may well provide the starting stimulus for higher-level target processing. The segmented target may well be considered after extraction from the background. The background may therefore receive little consideration, and perhaps should not be suppressed. This approach would yield a more object-based propagation of dynamic IOR at little computational expense. Indeed, there exists evidence to suggest that feedback in the primate brain is used to bind different visual attributes of an object, such as

[1]It is noted that the log-Gabor frequency/phase analysis of images can however yield other cues useful for visual perception. For instance, Kovesi has shown that image phase congruency can be used to extract image symmetry, corners and edges [Kovesi (2003)]. This process requires little more processing resources beyond the numerous convolutions already taking place in the OCS servers. Similarly, Rougeaux also implemented a method to compute disparity from phase difference [Rougeaux & Kuniyoshi (1997b)] using the output of the same type of complex band-pass filters.

colour or form, into a unitary precept [Trieisman & Gelade (1980); Reynolds *et al.* (2000)]. As well as providing object-based dynamic IOR, the MRF ZDF target extraction may facilitate such object-based binding of cues into a unitary precept for object representation. Using the cue-surface correspondence provided by the occupancy grid spatial representation it is also possible to project mosaic frame cues, such as saliency, into a 3D representation of perception, allowing further investigation into 3D perception.

It is difficult to *prove* that such modifications of the system would further bias the system towards more primate-like visual behaviours. However, because the components of the system have been largely inspired by observations of primate visual perception, we would like to compare the behavioural performance of the system to that of primates. The next two chapters concentrate on comparing the behavioural similarity of the synthetic vision system with that of the human vision system.

## 8.6 Summary

We have presented our approach to synthesising primate active visual attention. We have specified biologically plausible cues and implemented them on a processing network. Cues are combined to create a saliency map. We create a *fixation map* by modulating a saliency map by an IOR bias and a task-dependent spatial bias. Cues contributing to saliency, the accumulation of IOR, and the layout of the task dependent spatial bias can be modulated online. We deal with dynamic scenes by covertly propagating IOR according to the motion of scene objects. peaks in the fixation maps are extracted and covertly moderated before a next target of fixation is selected. Moderation involves covertly accumulating evidence about the strength and spatial consistency of the locations of peaks in the fixation maps. Peaks that pass moderation are selected for overt attention and saccade is initialised. Pre and post-attentional saccade routines exist to ensure stable system performance and attentional target pursuit.

The perception of attention has been integrated with the MRF ZDF coordinated fixation, active rectification and spatial awareness components into a flexible real-time synthetic vision system based on primate vision. The emergent visual behaviours have been shown in demonstration footage. The system is now

ready for further behavioural comparisons with the human vision system.

# Chapter 9

# Human Trials

Figure 9.1: The Punch and Judy show.

In this chapter we conduct psycho-physical trials to benchmark human visual behaviours. Participants are free to gaze as they please while their scanpaths are non-intrusively recorded.

## 9.1 Introduction

Components of a primate-inspired synthetic vision system have been integrated on a processing network. The mechanism, its control, and visual processing components are based on biological inspiration. Ideally, the system would exhibit behaviours that reflect this biological inspiration. It is hoped that similarities in the underlying system models elicit similarities in basic gaze behaviours. We therefore examine human gaze behaviours elicited by 3D visual stimuli moving in a reproducable manner in a controlled scene volume. Participants are given a basic visual task and are free to look wherever they please. A non-intrusive gaze tracker records participant's scanpaths and permits participants to move their heads as they please. Although no two participants' gaze is expected to follow the same scanpath, we do expect to find some statistical similarities in terms of inter-individual gaze behaviours.

### 9.1.1 Aim

We aim to characterise unconstrained human gaze behaviours in a controlled dynamic scene. Such characterisation is to be used as a benchmark for the evaluation of behaviours produced by a primate-inspired synthetic vision system. We aim to identify parameters suitable for numerical characterisation of human gaze behaviours, and to investigate statistical conformity in such parameters across participants. We expect to find that some parameters are largely dependent on the observed scene, some more dependent on the participant observing the scene, and some on the physical capabilities of the eye.

### 9.1.2 Considerations

We first consider previous work in observing human gaze paths produced by presenting visual stimulus to human observers. We then prescribe the experimental method, including apparatus design, trial procedure, participant briefing, and data logging and processing.

We initially conduct and qualitatively analyse two pilot trials such that an empirical examination of gaze characteristics can be obtained. The empirical analysis facilitates the investigation of quantitative parameters that may be ex-

tracted from subsequent trials. We propose suitable behavioural metrics. Trials are then conducted and analysed such that inter-participant behavioural metric statistics are obtained.

In terms of assessing the synthetic vision system, once we have characterised these types of human gaze behaviours, we hope to find that similarities in the underlying synthetic and human system models elicit observable and measurable similarities in basic gaze behaviours.

## 9.2 Background

The first non-intrusive eye trackers were built by George Buswell in 1922. He used beams of light that were reflected off the eye and recorded onto film. Buswell made systematic studies into human scanpaths during reading [Buswell (1922), Buswell (1937)], and picture viewing [Buswell (1935)]. In the 1950s, Alfred L. Yarbus conducted further eye tracking research on the cyclical gaze patterns in the examination of pictures [Yarbus (1967)]. More recently, studies into the distribution of gaze over web pages for advertisement impact assessment have been conducted [Chandon *et al.* (2001)]]. Preferential ordering and distribution of overt attention may be extracted from such studies. Although useful in investigating perceptual saliency of static 2D stimulus, such studies do not examine the temporal or behavioural characteristics of human gaze. Fixation occurs only upon static stimulus, spatiotemporal dynamics are not considered.

Eye gaze tracking on video is reportedly "relatively new and unexplored in the literature" [Djeraba (2006)]. Video investigations are able to include motion, which is known to play a role in visual saliency. Tracking of a participant's gaze is usually conducted to determine which visual stimuli are salient, or which are considered relevant to the participant in the course of executing a task. For example, gaze trackers were used to track a participants gaze during 2D simulations of driving to investigate what types of stimuli attract driver attention, and to assess driver vigilance and alertness [Lappe (2006)].

Such 2D video investigations restrict the visual search space. As we have seen, saliency depends on the entirety of one's view. When humans view a real scene (rather than a photograph or a movie) their deliberate shifts in attention affect the viewable region of a scene, introducing new stimuli, and excluding other

stimuli. In this manner, gaze shifts may affect the perception of visual saliency at all points in the updated view. Additionally, a movie records a subset of the entire scene, as captured by a cameraman. As such, the observer of a movie does not fully arbitrate gaze direction, which may introduce differences between the gaze behaviours of humans during the observation of video, and the observation of a real scene. Other factors, such as pan-induced optical flow, pixel saturation, and camera-operator selected focus may also bias a participant's perception of saliency and their attentional priorities.

Gaze trackers have indeed been used to monitor human gaze in 3D scenes. A driver's gaze was tracked to monitor driver vigilance while driving a real vehicle [Fletcher *et al.* (2005)]. Assessing *vigilance* involved monitoring the frequency at which a driver's gaze was directed towards scene locations where the road is expected to be located. Such existing studies are often designed to investigate *where* humans allocate overt attention during a particular task like driving. They generally ask "what types of things/regions are gazed at?" and "how often?". A task such as driving may be complicated, and significantly dependent upon a participant's previous experience. Experience-dependent and task-specific gaze characteristics may influence gaze in the execution of such complex tasks, for example, the propensity for a driver to overtly attend the focus of expansion of optical flow. Such existing studies seem to focus on assessing task-specific saliency rather than assessing gaze behavioural characteristics.

We are most interested in investigating low-level gaze behaviours involved in the deployment of overt attention during scene perception and novel events. We wish to establish consistent behavioural characteristics in general gaze deployment. We are interested in the spatiotemporal properties of saccade and smooth pursuit during general scene perception, rather than during the execution of a complex task. We therefore conduct trials designed to characterise gaze behaviours during the deployment of overt attention. The trials use real, dynamic, 3D scenes and stimuli, such that the effect of 3D scene structure and motion can be included in the assessment. We aim to leave gaze arbitration entirely up to the participant observing the controlled scene. We are interested in characterising natural gaze behaviours associated with attentional saccade and the smooth pursuit of stimuli.

We create a small scene, similar to a puppet show, in which we can control

3D visual stimulus. We use *FaceLAB*[1], a proprietary gaze tracking package, to record a participant's gaze as they observe the scene. Participants are free to move their heads as they please.

### 9.2.1 FaceLAB v3

FaceLAB by Seeingmachines is commercial software that enables real-time eye and head pose tracking. It is marketed as a high accuracy, high speed, completely vision-based gaze tracking solution. It is reported to operate robustly under various lighting conditions and is not affected by participants wearing glasses. It is non-invasive in that no equipment needs to come in contact with the participant. Calibration for a participant (using FaceLAB v3) takes only a matter of minutes.

The FaceLAB stereo camera rig (Figure 9.5) is placed between the participant and the scene window. We calibrate FaceLAB by taking a snapshot of the participant and selecting face tracking features. A basic *world model* is configured in FaceLAB where a rectangular screen represents the scene window. FaceLAB returns the coordinates of a participant's gaze within the scene window in Cartesian coordinates. Gaze data is accurate to within a few degrees. The blue circle projected onto the screen in Figure 9.2 (left) shows the gaze FaceLAB gaze estimate and approximate error radius. Figure 9.2 (right) shows the concurrent re-projection of head pose and eye gaze used by FaceLAB to determine the intersection of gaze within the screen window.

## 9.3 Method

We first discuss preparations required before commencement of the trials. We then describe the apparatus and experimental procedure.

### 9.3.1 Ethics

The Australian National University has measures in place to ensure that research involving the participation of humans is conducted in a manner that is ethical and responsible. As such, permission and disclosure was required before human trials

[1] A *SeeingMachines* product - http://www.seeingmachines.com.

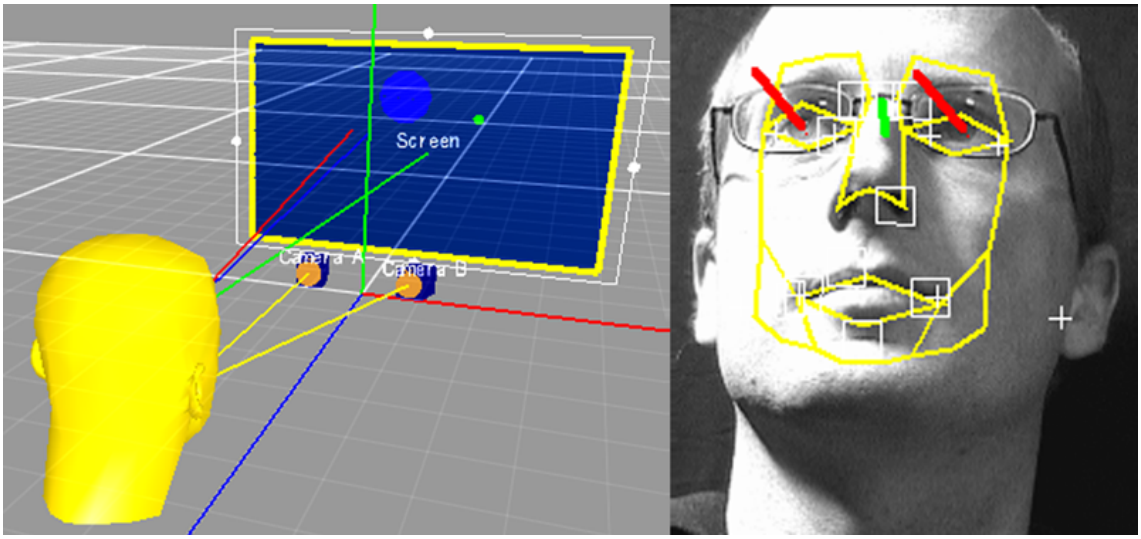



Figure 9.2: FaceLAB output. World model (left), and head pose and eye gaze re-projected onto right FaceLAB camera image (right).

could begin. Official application was registered with the ANU ethics committee, and subsequently granted (see Appendix A).

### 9.3.2 Participants

A cross-section of age, sex and ethnicity was sought. The time-consuming nature of trials and post-processing meant that 20 trials were prescribed. We were interested in participants with reasonable visual acuity, no visual disorders, and who were relaxed and alert.

Participants were sought on a voluntary basis. Every effort was made to protect the interests of participants, and to ensure the comfort of participants. In general, participants indicated the experience was enjoyable. No participants expressed concern or discomfort.

### 9.3.3 Apparatus

Participants were seated in a *viewing booth* (right, Figure 9.4) and were shown controlled 3D stimulus in a *scene booth* (left, Figure 9.4) through a *scene window*. The scene booth measured 1.0m in depth by 2.0m wide and 2.5m high; the viewing

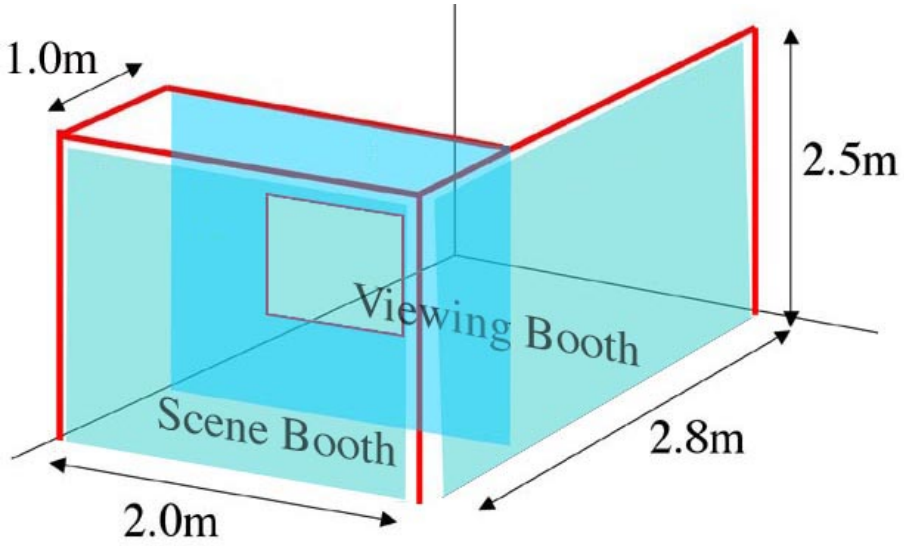



Figure 9.3: Apparatus: scene booth and viewing booth dimensions.

booth measured 1.8m in depth by 2.0m wide and 2.5m high (Figure 9.3). The rectangular scene window measuring 1.0m wide by 0.7m high (left, Figure 9.7) was cut in the centre of the partition separating the scene booth from the viewing booth.

The scene window size was kept small so the potentially different sizes of the participant's visual peripheries would not be a significant factor in gaze selection. Moreover, the synthetic vision system (to be tested later using the same apparatus) has a comparatively small periphery; if a broad window were used, stimulus that may affect a human's gaze selection might not be seen by the synthetic system. The window size was selected such that all manipulated stimuli could usually be seen by both humans and the synthetic system.

Participants were seated 1.5m from the scene window in the viewing booth, such that their eyes were aligned approximately with the centre of the scene window. The booths were constructed from light-blocking fabric supported by a wooden frame. The interior of the scene booth was illuminated from above. The viewing booth and window were designed such that the contents of the scene could be bounded and controlled. Participants were seated in the viewing booth so their surroundings could also be bounded and controlled.

The surfaces surrounding the viewing window and the surfaces within the

Figure 9.4: Apparatus. The *scene booth* and rig (left); and exterior of *viewing booth* (right).

booth that the participant was able to see, were draped with fabric exhibiting a specifically selected pattern (left, Figure 9.7). The fabric was chosen such that it did not incorporate any iconic or geometric patterns; such that it exhibited texture; such that it was considered equally conceptually salient over its entire surface; and such that its visual saliency was considered low in comparison with the objects that were to be introduced into the scene.

Visual stimuli, comprising of various 3D objects, were introduced into the scene booth such that they were visible to the participant through the scene window. The mechanism used to control the motion of the visual stimulus was hidden from the participant's view.

A mechanism was constructed within the scene booth that allowed the manipulation of up to six objects in the scene simultaneously. For simplicity in apparatus construction, the rig allowed objects to be moved in horizontal and vertical directions (left, Figure 9.4), which meant that objects were kept at various constant scene depths relative to the participant. Moveable rods controlled object horizontal positions and fine semi-transparent nylon thread suspended the objects in the scene and controlled their vertical positioning. Discrete positions were marked on the rods and string holders such that repeatability in moving

Figure 9.5: Non-intrusive acquisition of participant's 3D gaze path using FaceLAB. An extra video camera is used to record the participant, placed between FaceLAB cameras.

objects about the scene was improved.

The FaceLAB camera rig was placed between the participant and front curtain, at a distance of 0.75m from the participant's head (Figure 9.5), and sufficiently low such that it did not obstruct the participant's view of the scene window (left, Figure 9.7). The participant was free to move their head in any manner they wish. They were also free to gaze upon any part of their surroundings, including the contents of the viewing booth or scene booth, so long as they did not leave their seat. Two additional video cameras were used to record the contents of each booth. One recorded the participant (centre camera, Figure 9.7), and one recorded the scene they were viewing through the scene window.

A storyboard (Figure 9.6) was created to direct object positions over time such that a puppeteer could recreate the motion paths of objects in the scene for each participant by positioning the rods and strings in a predefined sequence. The scene could not be reproduced exactly for each observing participant, but variables such as the rotation of the objects around string axes, and swinging, were similar in character across all trials.

"How many apples?"

| T | 1 Pear | 2 G aple | 3 Banan | 4 R aple | 5 Peach | 6 Orange | ✓ |
|---|---|---|---|---|---|---|---|
| 0 | R4-S1 | R3-S12 | R3-S12 | R5-S13 | R2-S12 | R0-S12 | |
| 10 | | | | | | S5 | |
| 20 | S8 | | | | | | |
| 30 | | | | | | R5 | |
| 40 | | | | | S9 | | |
| 50 | R1 | | | | | | |
| 60 | | | | | | S10 | |
| 70 | S4 | | | | | | |
| 80 | | | | S5 | | | |
| 90 | | | | | S1 | | |
| 100 | | | | | | S13 | |
| 110 | S12 | | | | | | |
| 120 | | | | R3-S7 | | | |

Figure 9.6: Storyboard used for repeatability of timing and object paths across separate trials. Objects in columns 2 and 3 were not used. "R" represents the motion of a rod (horizontal movements), "S" represents strings (vertical movements); numbers represent rod/string positions.

#### 9.3.3.1 Stimulus

Stimuli that may elicit emotional or significant cognitive responses (the apparent mood on the face of a doll, for example, may affect each participants attention differently) were avoided. Different scene objects were chosen such that they had approximately equal conceptual saliency. Common fruit were selected because it is likely they are perceived as similarly salient, yet come in a variety of different colours, shapes and sizes (right, Figure 9.7). They were deemed to be comparatively salient in front of the selected backdrop fabric. Rather than using real fruit, synthetic replicas were used such that the same objects could be reused in numerous trials.

### 9.3.4 Trial Procedure

Upon presenting themselves for participation, participants were asked to complete a permission slip and a questionnaire (blank copies included in Appendix A). The questionnaire was designed to verify that each of the trials constituted a valid data set, and to register each participant's voluntary participation. The

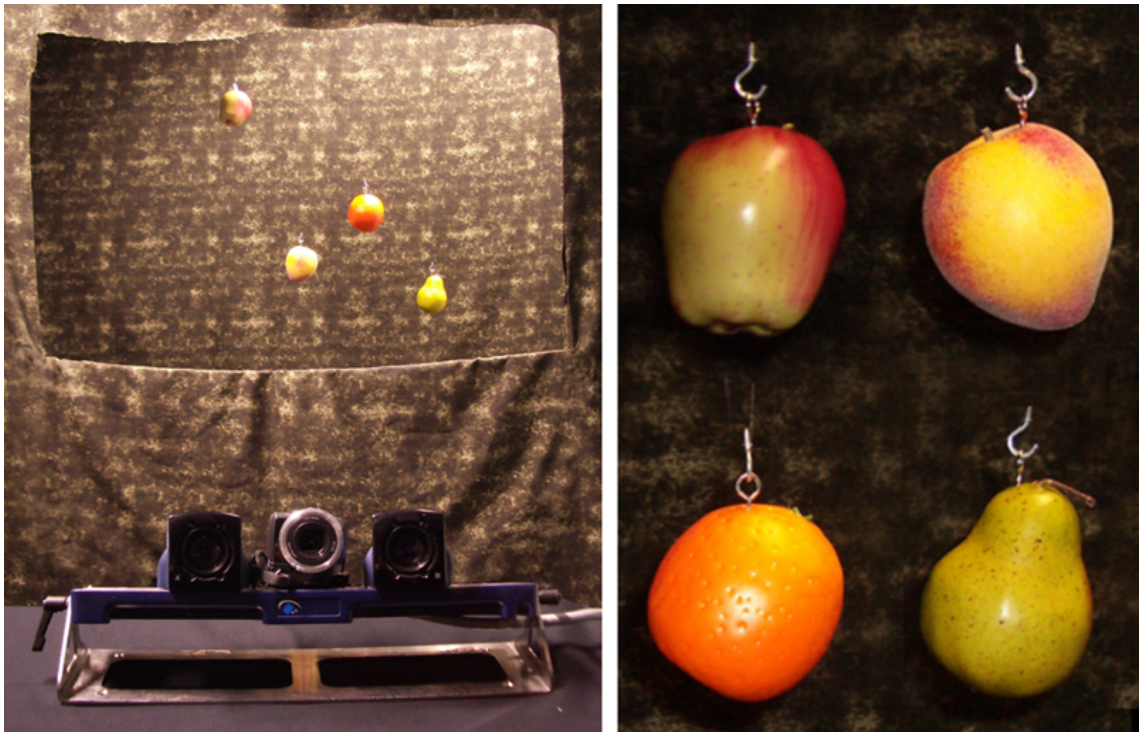


Figure 9.7: Visual stimulus: participant's view (left), and stimuli (right).

questionnaires were designed to help establish that all participants were similarly alert and, in the case of an anomalous trial, to help identify whether the discrepancies corresponded to the participant's registered level of alertness. As well as completing the questionnaire, participants were asked to participate in a basic visual acuity test. They were allowed to wear spectacles during both the acuity test and trial, if they so chose.

Once seated in the booth, and remaining seated, participants were given time to visually explore their surroundings and become comfortable with their surroundings. This time was used to calibrate FaceLAB for each participant. Calibration quality was confirmed online by asking participants to look at the corners of the scene window and confirming that FaceLAB projected their gaze onto the model of the scene window (left, Figure 9.2) accordingly.

When initially seated, no objects were visible in the scene booth. The nylon strings were nearly invisible to the participant, and no participants mentioned noticing strings, nor were recorded as gazing directly at strings that were present before trials commenced. At this time participants were also advised they would shortly be given a visual task to perform while visual stimulus was presented to them. Participants were advised they should voice any queries before commencement of the trial.

Immediately before the trial began, the participants were told the task, which was to count apples they saw in the subsequent trial. They were told there was no need to count aloud, they would be asked after the trial how many were seen, that it was not an assessment of any form, and that there were no right or wrong answers to the question. The question was designed to give the participants a forwards search task, and to predispose participants to consider apples as more salient, based upon their prior knowledge of apples.

The trial began with a blank scene. Various fruit were then moved into and around the scene, one at a time, according to the storyboard. All objects swung naturally to some extent, but deliberate translations were considered likely to increase the visual saliency of the perturbed object. Some differences in the path of stimuli in different trials were considered acceptable. In fact, exact repeatability would involve more rigid connections to the manipulated objects. Mechanisms to ensure more exact repeatability would be more difficult to conceal, and rigid motions could be considered visually salient in their own right, or make the partic-

ipant ponder the mechanical setup, which could distract some participants more than others. It was expected that all participants were likely to be affected in the same manner, and minimally, using the selected apparatus design.

After each trial, participants were asked informally how they thought the scene was manipulated; all responded that it involved dangling fruit on strings, though most admitted they could not see the strings until after the trials began. Some did not see the strings at all, though they were aware they existed. No participants were recorded to have directed their attention to locations occupied by strings for significant periods of time - they were far more inclined to track the fruit, as predicted. Also, during the trials, no participants looked at the FaceLAB cameras. They were not told not to look at the cameras, or anything else in the booth; they were merely asked to concentrate on the task. Participants were sufficiently comfortable with the cameras to supress/ignore them. They were evidently far more interested in the novelty of the task.

Each participant participated in only one short trial because familiarity with the trial setup might affect a participant's gaze behaviour in subsequent trials.

### 9.3.5 Trial Logging

The following sequence of events was required in order to capture trial data:

- prepare start positions for all objects
- invite participant into viewing booth
- calibrate FaceLAB for participant (5-10 mins during which they could become comfortable in the booth)
- start video cameras recording and FaceLAB logging
- give participant the task
- perform trial according to storyboard
- ask participant question related to task
- stop recording

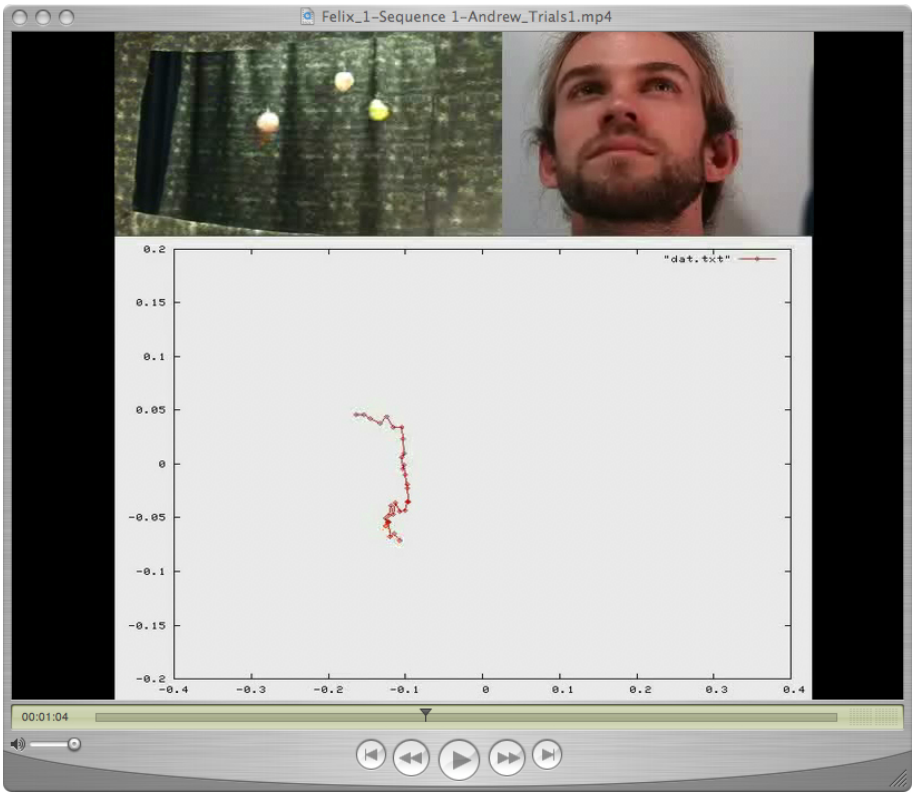



Figure 9.8: Video log summary, Pilot 1 (*snapshot - see Appendix C for full video*).

### 9.3.6 Data Processing

The data obtained from the trials is time-stamped eye gaze x,y-coordinates within a 2D projection of the scenario window. A video of the gaze path was created and synchronised with video recordings of the scene and participant. An example video log is shown in Figure 9.8. This movie was used to hand mark the data according to periods when objects were being actively moved, and when they were not (residual swinging may exist). Once fully labelled, the data was ready for analysis so that human fixation behaviours in each trial could be analysed.

## 9.4 Pilot Trials

Two pilot trials with two different participants were initially conducted. These pilot trials were used to observe the types of visual behaviours that emerged during the trials, and how such behaviours could be characterised statistically.

The two pilot trials were conducted according to the procedure prescribed above. Gaze data was successfully logged and marked according to perturbation periods. After markup, plots showing various aspects of the pilot trial data were constructed. This empirical analysis of the trial data was first conducted to establish an approach to developing metrics for characterising the data statistically.

We now present an empirical assessment of the pilot trials. We present plots that help to establish metrics of gaze behaviours. We define parameters useful for a statistical characterisation of the gaze behaviours that we may extract from subsequent trial data.

### 9.4.1 Empirical Examination of Pilot Trials

Figure B.2.1.4 shows plots of recorded FaceLAB gaze points over the entire duration of the pilot trials. Points are projected into a 2D representation of the scene window. Gaze velocities between points can be extracted because position data is time-stamped at 60Hz.

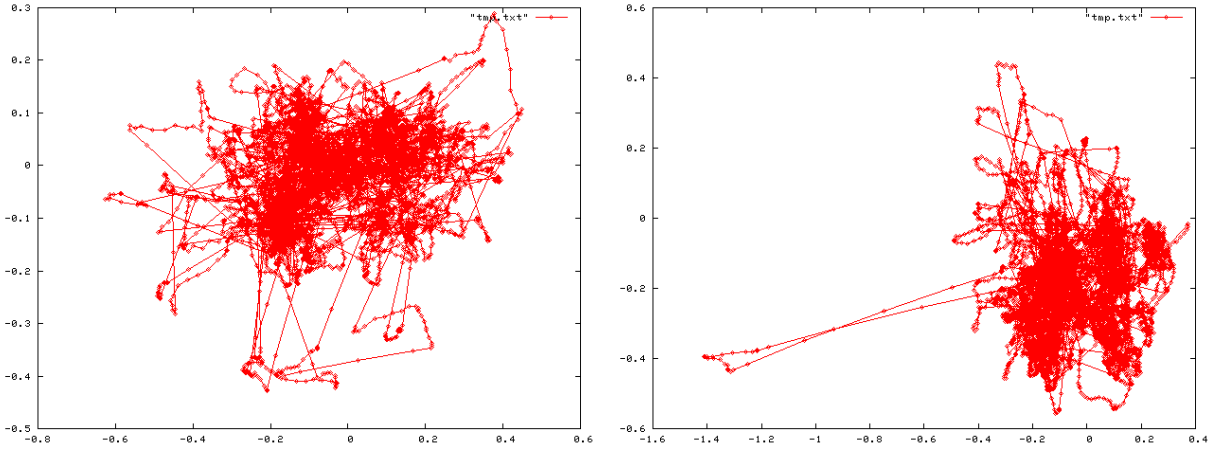



Figure 9.9: Complete trial scanpaths (not to scale). Pilot 1 (left), and Pilot 2 (right).

We may then construct a histogram of gaze velocities over the entire trial. Figure B.2.1.4 shows velocity histograms over the duration of the trials for both pilot participants. High velocities tend to get saturated in this representation by the quantity of low velocity frames. Consequently, the velocity histogram can be represented over the gaze path (distance), rather than per image frame (time) by multiplying the velocity histograms by the distance traversed between frames, to obtain a distance-weighted velocity histogram (Figure 9.11).

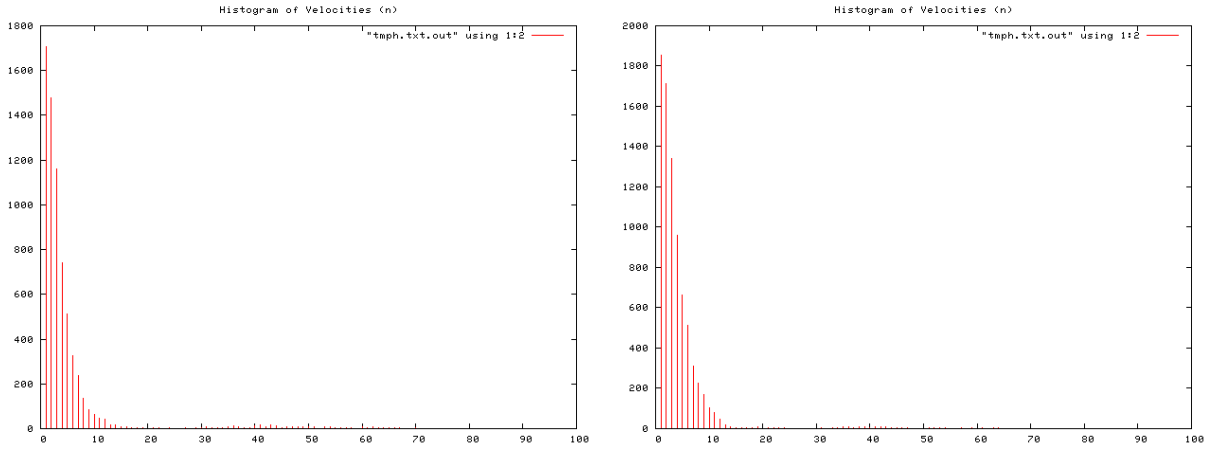



Figure 9.10: Histogram of Velocity Magnitudes. Pilot 1(left), and Pilot 2 (right).

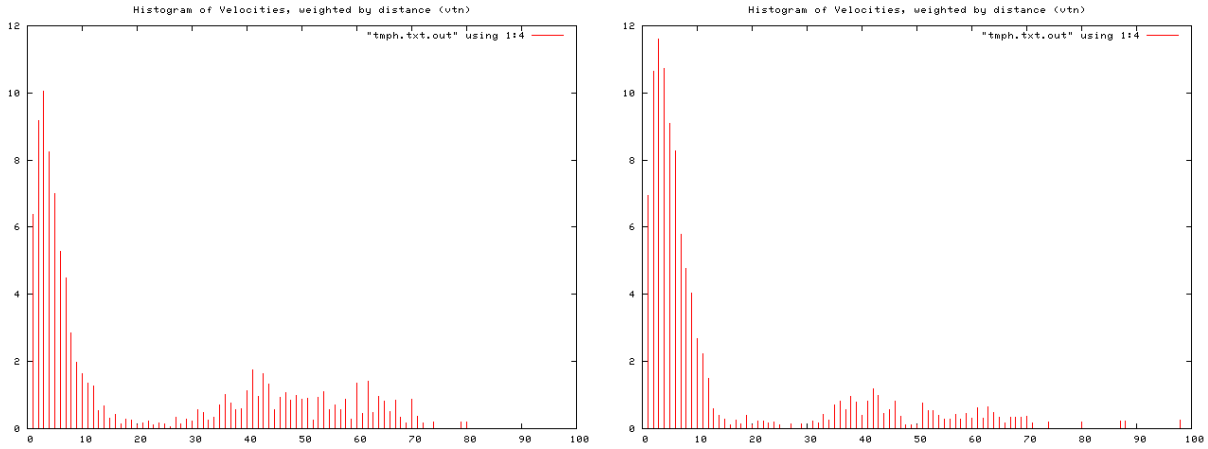



Figure 9.11: Histogram of distance weighted velocities. Pilot 1(left), and Pilot 2 (right).

It is then evident that much of the gaze path is attended at either low (near-zero) velocities, or high velocities, with few frames exhibiting velocities in-between. The velocities near zero are likely to correspond to frames during which a target is fixated upon and/or smoothly pursued. The high velocities correspond to saccades. For each trial, a threshold is selected that separates the group of high velocities from low velocities in the distance-weighted velocity histograms. Frames exhibiting velocities above the threshold are labelled as saccades; those below the threshold are labelled as smooth pursuits (Figure 9.12).

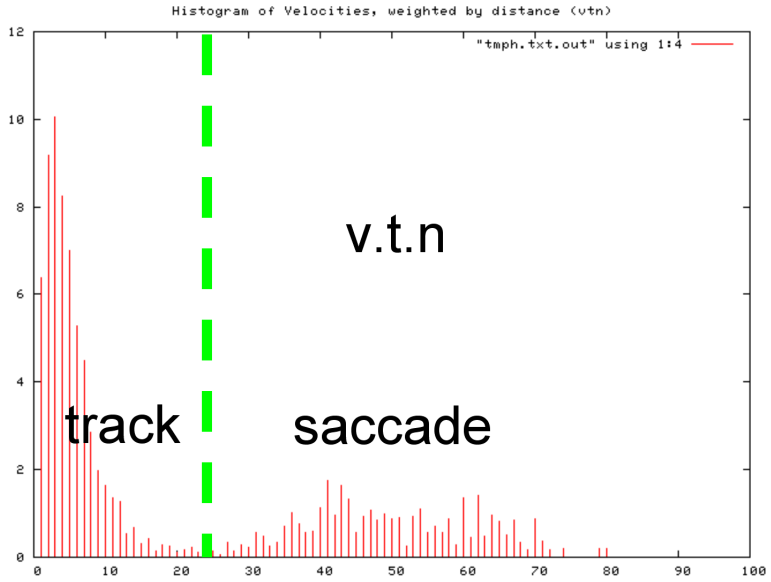



Figure 9.12: Choosing the saccade velocity threshold, Pilot 1.

We can then label frames in the 60Hz gaze data as saccade or smooth pursuit. Figure 9.13 shows the velocity magnitudes of eye gaze at all frames. The blue lines mark frames where the threshold velocity has been exceeded (saccades) and verifies the selection of the saccade velocity threshold. Figure 9.14 shows a zoomed view of the velocity magnitude profile. The green steps mark periods of time where an object was being actively perturbed. The data has been marked according to periods when on object is being actively perturbed or not. We can therefore plot histograms of velocities during these classes for comparision. Figure 9.15 shows velocity histograms separated into these two classes.

The histograms show a reduction in the number of frames above the saccade threshold velocity (saccades) when objects are being perturbed. Conversely,

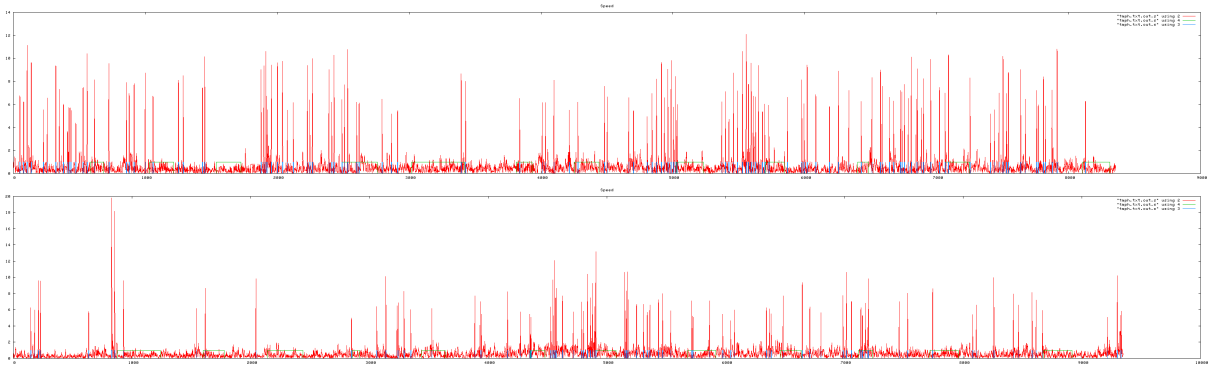


Figure 9.13: Velocity profile. Velocity magnitudes per frame. Pilot 1 (top), and Pilot 2 (bottom).

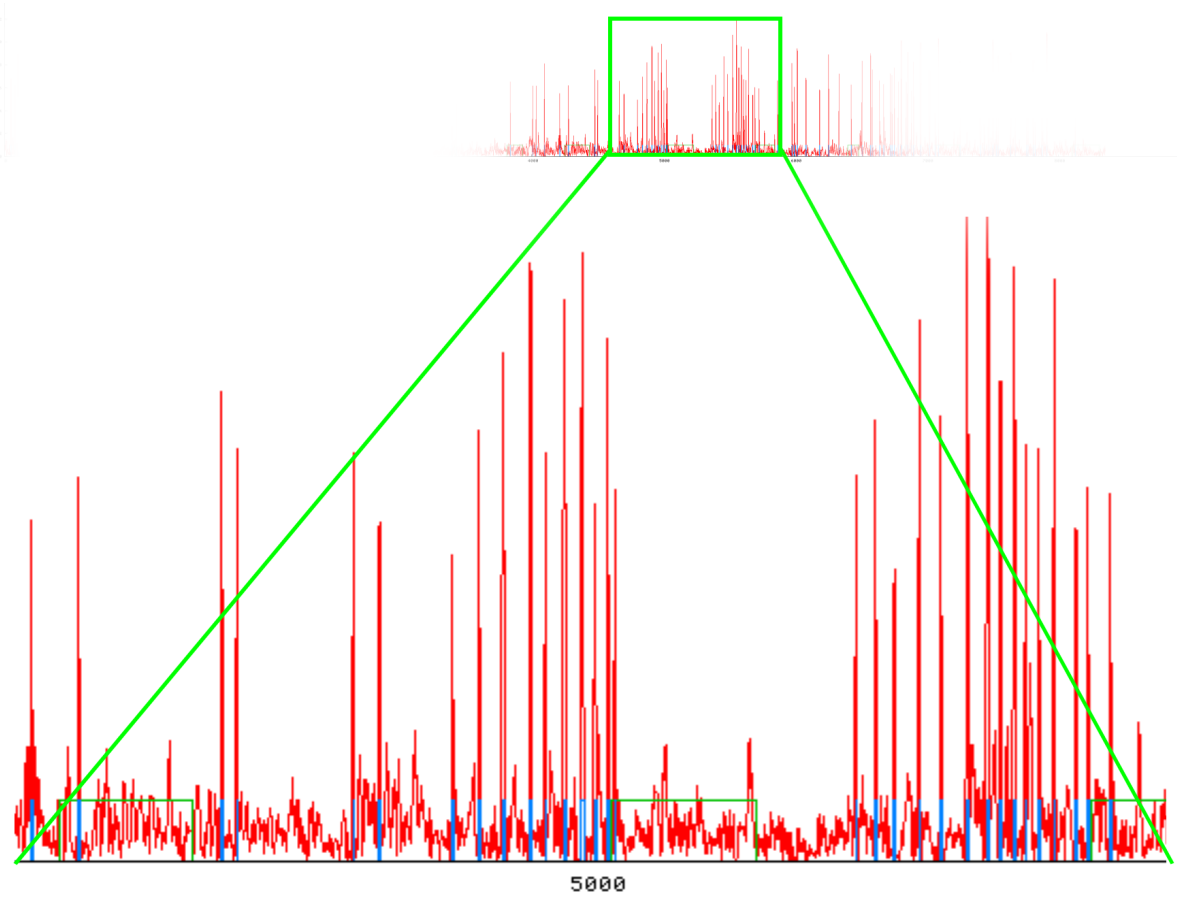



Figure 9.14: Zoomed velocity profile.

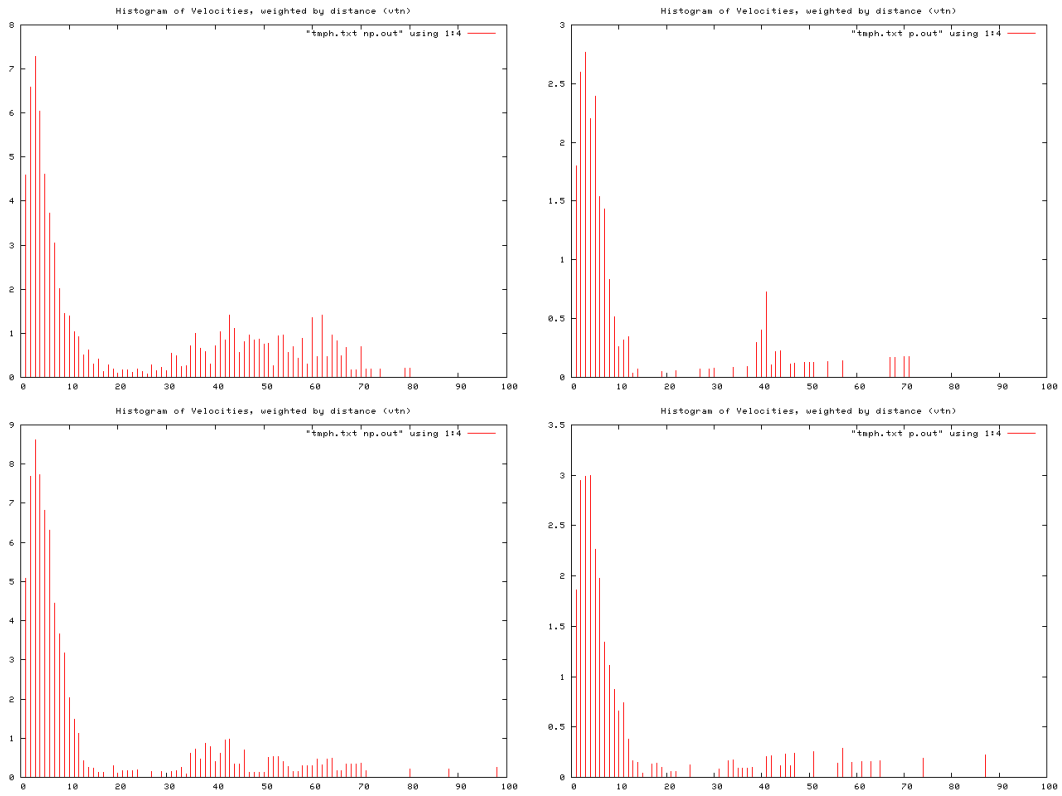



Figure 9.15: Histogram of velocities during non-perturbation (left), and during perturbation (right). Pilot 1 (top) and Pilot 2 (bottom).

when objects were not being actively perturbed, gaze was more saccadic. This suggests that the participant preferentially diverted their attention to the moving object and tracked it, resulting in a reduction in the saccade rate. During the non-perturbed periods, the scene does not change, so the increase in saccades is likely to be caused by other factors. Nevertheless, the discrepancy in velocity histograms indicates that different behaviours are present during the *non-perturbed* and *perturbed* periods. This distinction suggests an examination of the data:

- as a whole
- during the non-perturbed periods
- during the actively perturbed periods.

We first examine the spatial distribution of gaze locations according to these categories. We plot the recorded saccade and smooth pursuit locations according to non-perturbed, perturbed and both periods. Figure 9.16 shows such plots for smooth pursuit points, Figure 9.18 shows saccade points.

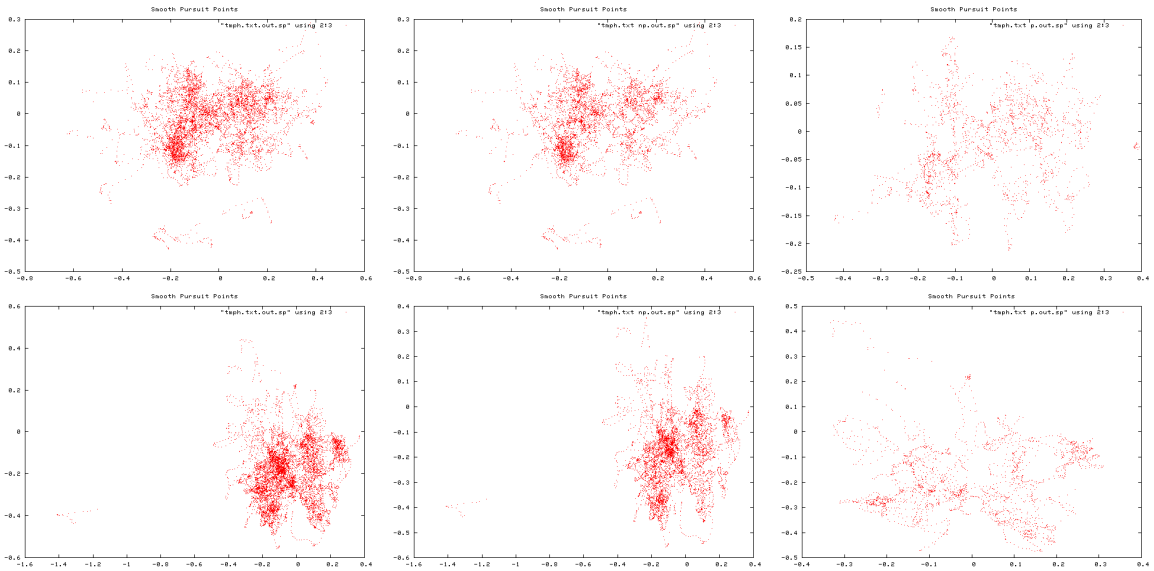



Figure 9.16: Smooth pursuit gaze locations. Pilot 1 (top), and Pilot 2 (bottom). Entire trial (left), during periods of non-perturbation (middle), and during perturbation (right).

As should be expected, the distribution of all saccade points is sparse and covers the entire scene somewhat evenly. We would expect the smooth pursuit gaze location points to correspond with the real motion paths of objects in the scene. During perturbed periods, this should mostly correspond to moving objects; during non-perturbed periods it should correspond to the storyboard locations of static objects. We therefore draw approximate motion paths of perturbed objects and superimpose these on the distribution of smooth pursuit points during perturbed periods (Figure 9.21). The coloured lines show the approximate storyboard motion paths we aim to replicate consistently over all trials. The crosses show where objects were stationary. Crosses are more likely to correspond to smooth pursuit gaze locations for non-perturbed periods and saccades.

Figure 9.21 shows a likeness between the density of smooth pursuit locations during perturbation, and the storyboard motion paths of objects. The objects

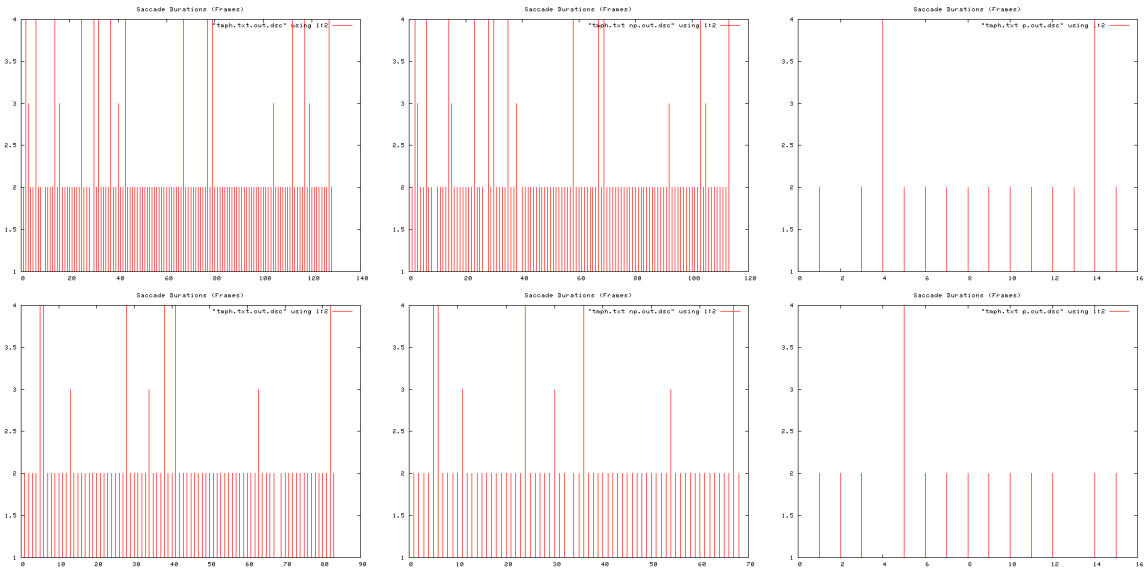



Figure 9.17: Saccade durations. Pilot 1 (top), and Pilot 2 (bottom). Entire trial (left), during periods of non-perturbation (middle), and during perturbation (right).

may swing, so high correlation is not expected, but the density of points is low where storyboard paths are not present, and the density of points increases near the object storyboard paths. This is a good indicator that the participants tended to track perturbed objects. As expected, the plot of smooth pursuit locations during non-perturbation periods, and non-perturbation saccade locations show a likeness to the storyboard location of stationary stimulus.

We may also consider smooth pursuit velocities in these three instances. Figure B.2.1.4 shows histograms of smooth pursuit velocities according to perturbed, non-perturbed and both periods. It was expected that smooth pursuit velocities would be slightly higher during perturbation periods, corresponding to the tracking of moving objects. However, no significant variation in smooth pursuit velocities is evident, probably due to error in the gaze estimation. Although gaze may have been stable, error in gaze estimation means that recorded gaze oscillates around a location where fixation was actually stable, inducing measured velocities similar to smooth pursuit velocities, where no smooth pursuit existed. Filtering may be necessary for more accurate comparison.

Next, we look at saccade speed histograms for these three instances (Fig-

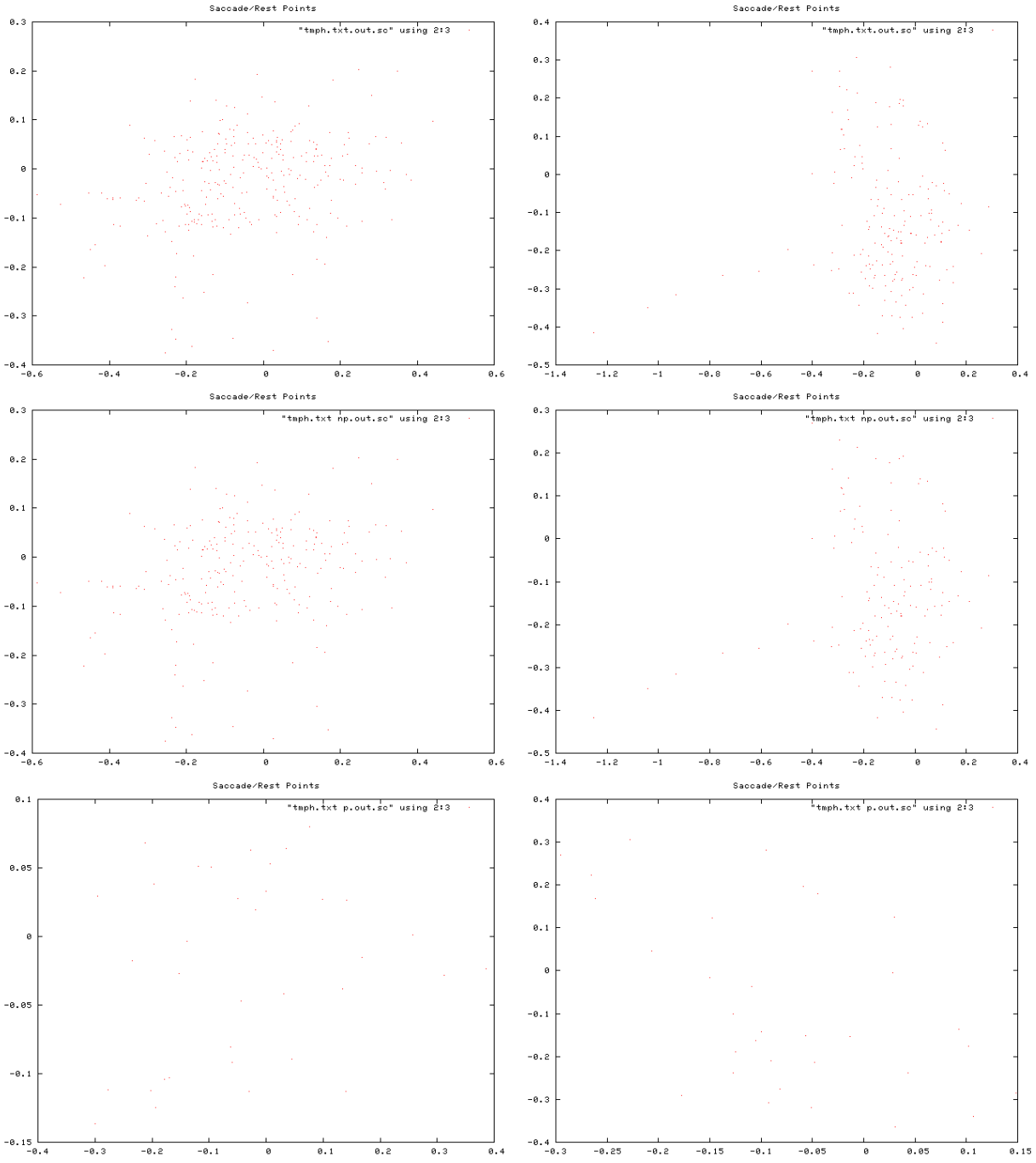



Figure 9.18: Saccade gaze locations. Pilot 1 (left), and Pilot 2 (right). Entire trial (top), during periods of non-perturbation (middle), and during perturbation (bottom).

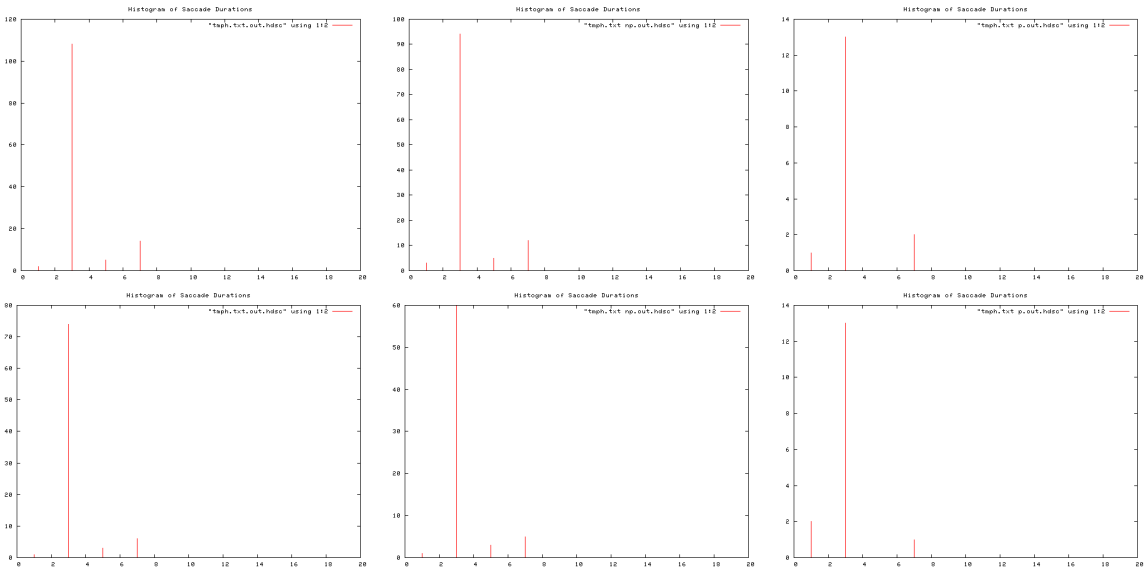



Figure 9.19: Histogram of saccade durations. Pilot 1 (top), and Pilot 2 (bottom). Entire trial (left), during periods of non-perturbation (middle), and during perturbation (right).

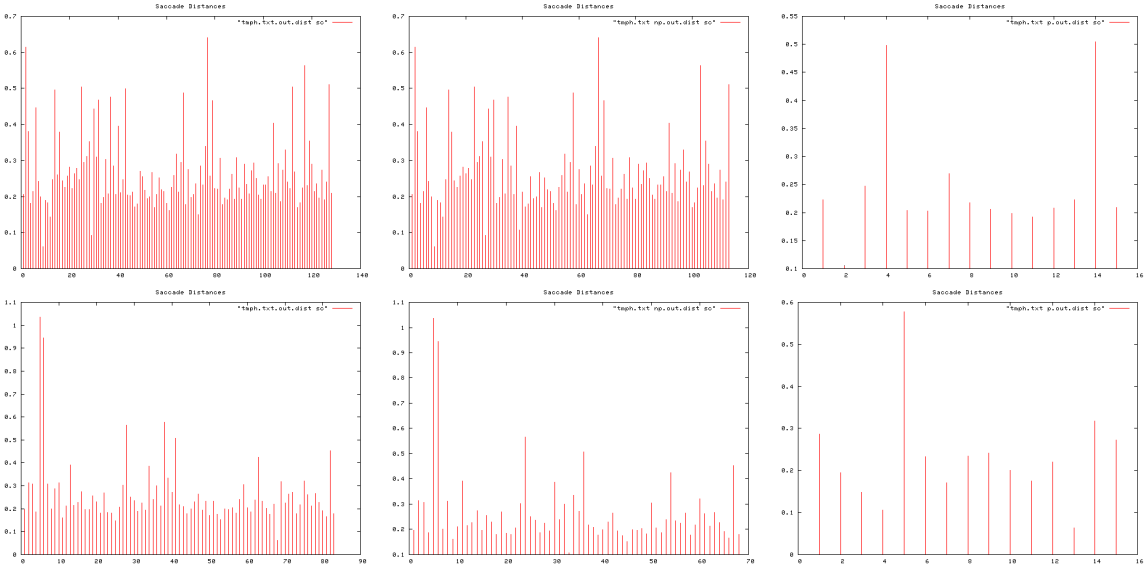



Figure 9.20: Saccade distances. Pilot 1 (top), and Pilot 2 (bottom). Entire trial (left), during periods of non-perturbation (middle), and during perturbation (right).

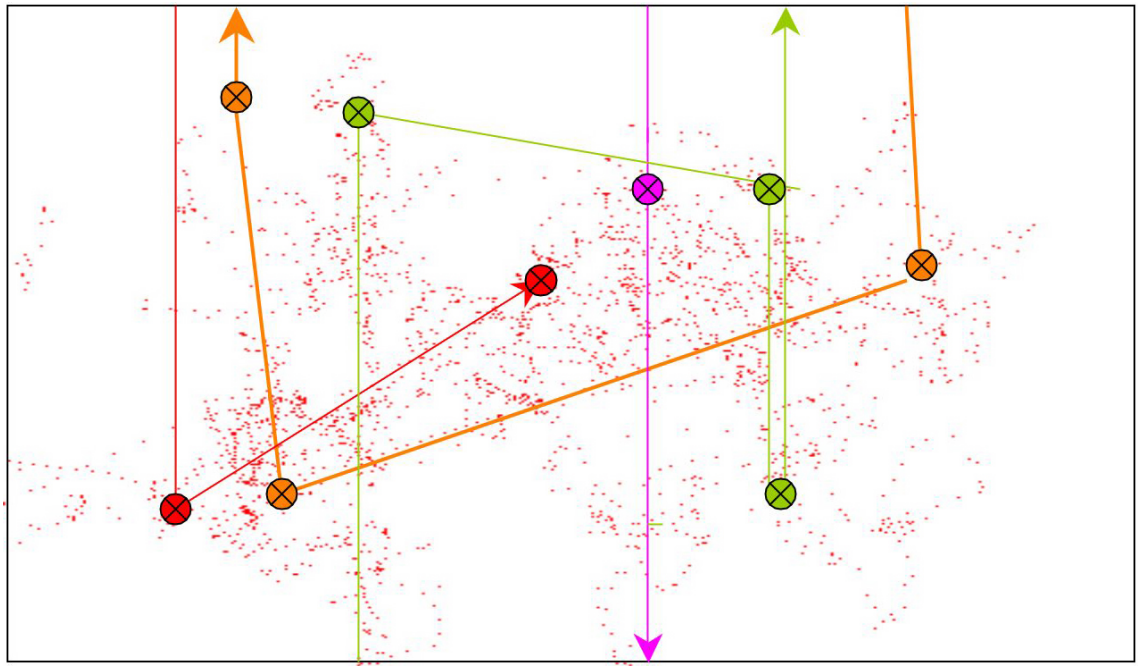


Figure 9.21: Approximate storyboard motion paths superimposed over gaze locations during smooth pursuit (red dots), Pilot 1(orange – orange, red – apple, green – pear, pink – peach).

ure B.2.1.4). It is evident that the number of saccades during perturbation periods is significantly less than during non-perturbation, but that velocities are centred around the same value.

We can also compare how smooth pursuit durations and histograms vary across different participants. Smooth pursuit durations are periods of low velocity separated by saccades. Figure 9.24 shows smooth pursuit durations, Figure 9.25 shows corresponding histograms. Figure 9.25 shows that a lower proportion of short durations exist during periods of perturbation than non-perturbation. As with a lowering in the saccade rate, this corresponds to preferential tracking of the perturbed object.

Smooth pursuit tracking distances were extracted (Figure B.2.1.4 and associated histograms were created (Figure B.2.1.4). Distance was measured as the vector subtraction of the smooth pursuit trajectory start and end points. Figure B.2.1.4 shows a significant shift towards a larger proportion of longer tracking distances occurs during periods where objects are perturbed.

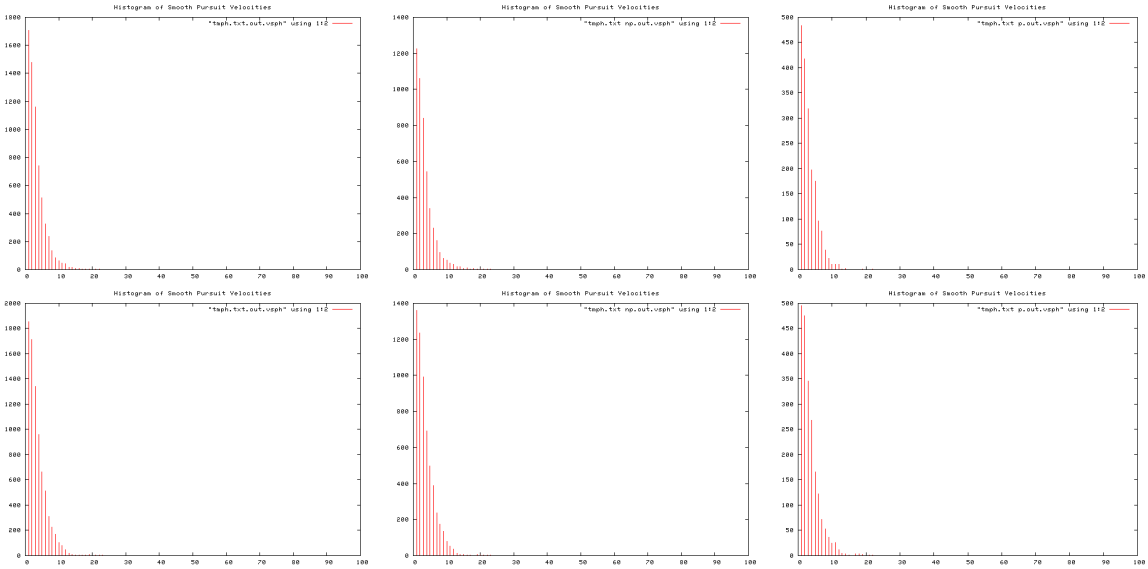



Figure 9.22: Histogram of smooth pursuit velocities. Pilot 1 (top), and Pilot 2 (bottom). Entire trial (left), during periods of non-perturbation (middle), and during perturbation (right).

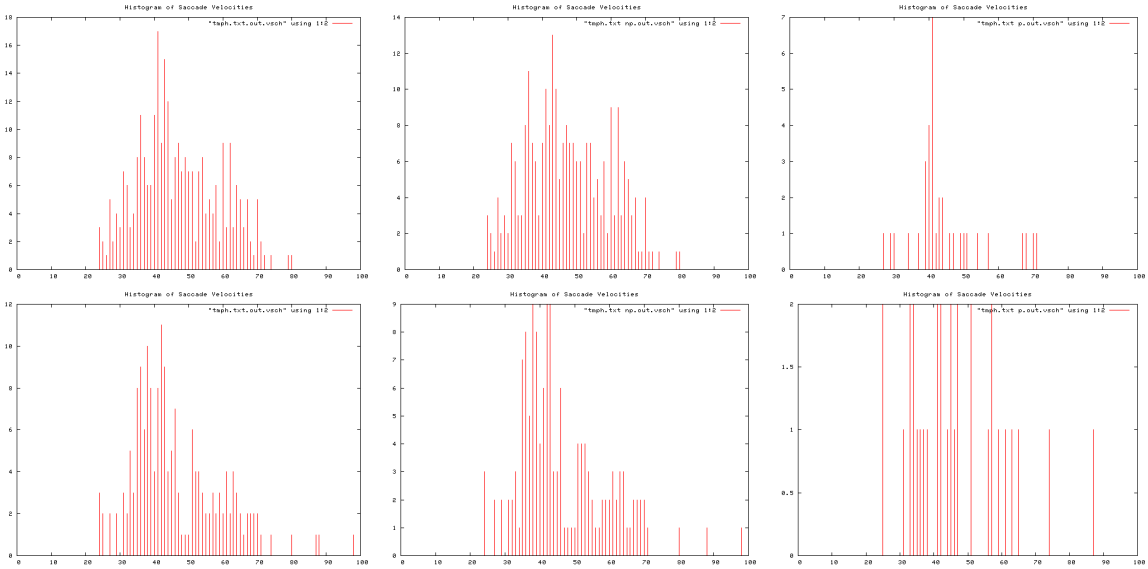



Figure 9.23: Histogram of saccade velocities. Pilot 1 (top), and Pilot 2 (bottom). Entire trial (left), during periods of non-perturbation (middle), and during perturbation (right).

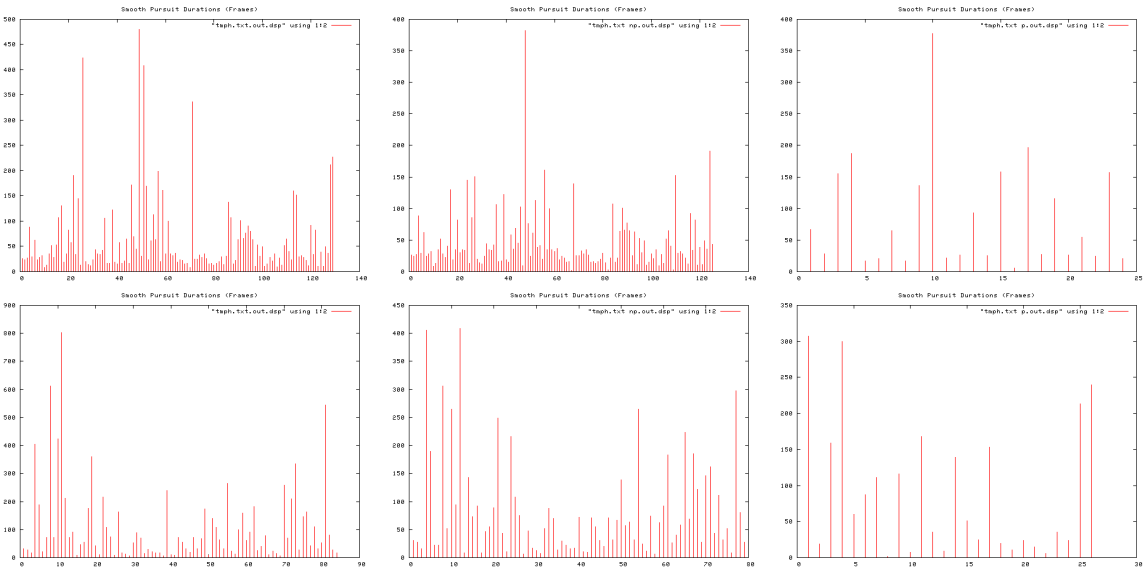



Figure 9.24: Smooth pursuit durations. Pilot 1 (top), and Pilot 2 (bottom). Entire trial (left), during periods of non-perturbation (middle), and during perturbation (right).

We next extracted saccade durations (Figure 9.28 and Figure 9.29), and distances (Figure B.2.1.4 and Figure B.2.1.4). No significant difference was noted between the saccade durations or distances during perturbation and non-perturbation periods. Nearly all saccades took place within the duration of two frames, and all within four, for both pilot trials. This is likely to be dependent on the similar speed capability of the eye across participants. The 60Hz frame rate of FaceLAB limits the resolution of estimates of the saccade duration.

Finally, we can consider the frequency at which each individual object in the scenario is attended. As discussed, during periods of perturbation attention was nearly exclusively given to the perturbed object. It therefore remains to consider the object re-attention frequency/period during periods of no perturbation. It was noted during the pilot trials that all objects were attended with an approximately similar frequency during periods of no perturbation. We therefore examine re-attention periods during periods of no perturbation for comparison across trials. The re-attention period is simply defined as the total time the object is present during periods of no perturbation divided by the number of occasions an object is attended during periods of no perturbation. The standard deviation

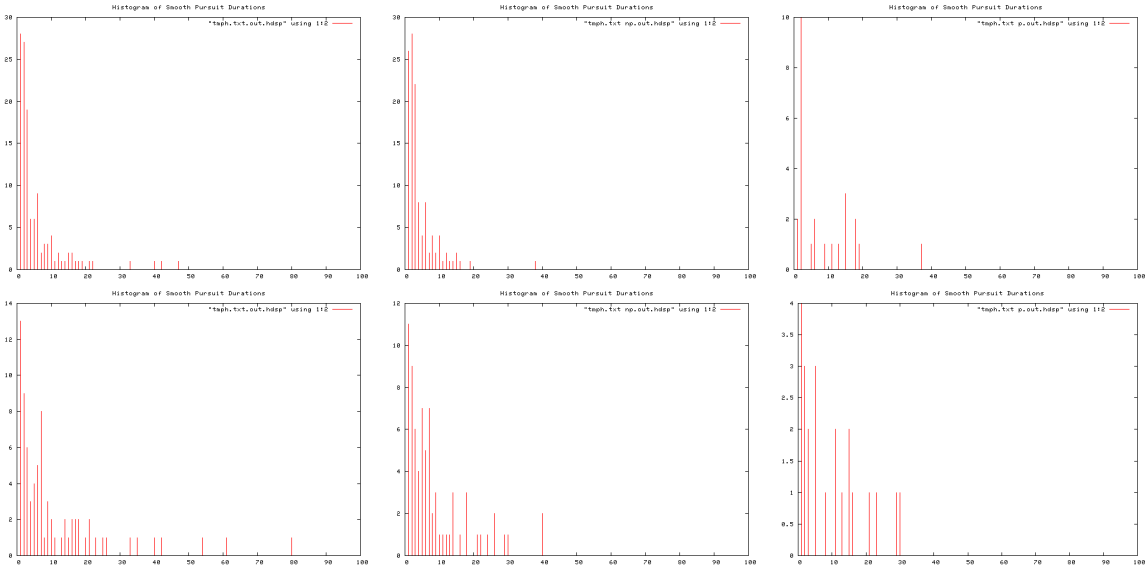



Figure 9.25: Histogram of smooth pursuit durations. Pilot 1 (top), and Pilot 2 (bottom). Entire trial (left), during periods of non-perturbation (middle), and during perturbation (right).

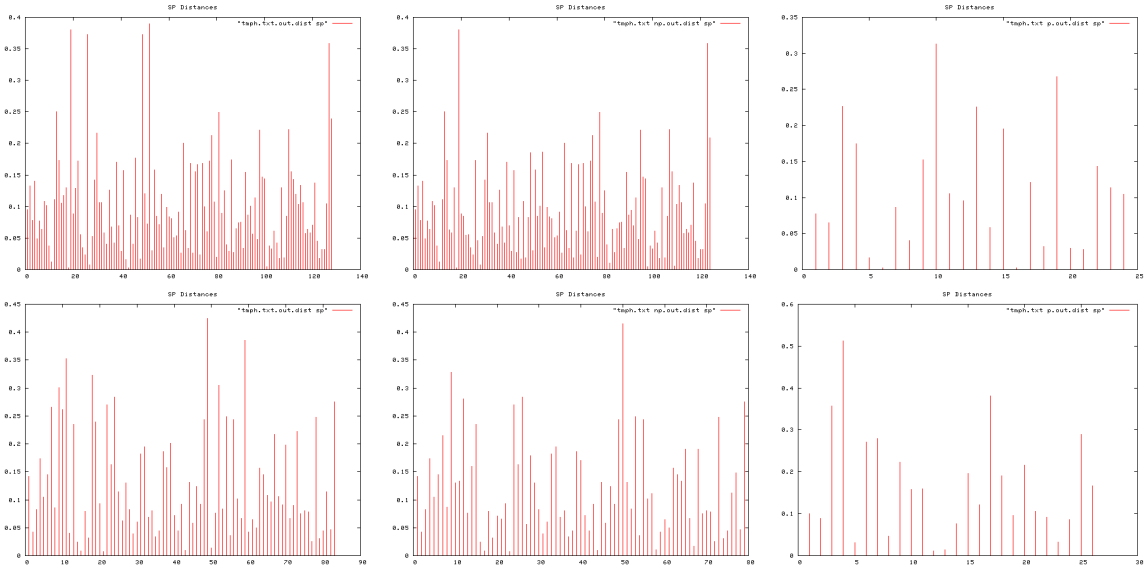



Figure 9.26: Smooth pursuit distances. Pilot 1 (top), and Pilot 2 (bottom). Entire trial (left), during periods of non-perturbation (middle), and during perturbation (right).

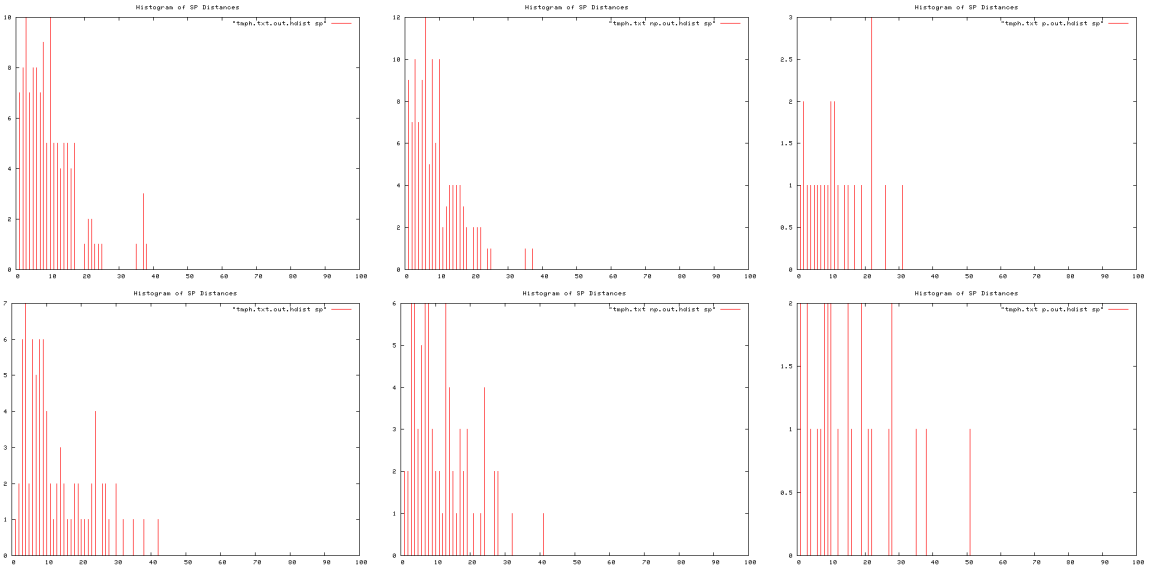



Figure 9.27: Histogram of smooth pursuit distances. Pilot 1 (top), and Pilot 2 (bottom). Entire trial (left), during periods of non-perturbation (middle), and during perturbation (right).

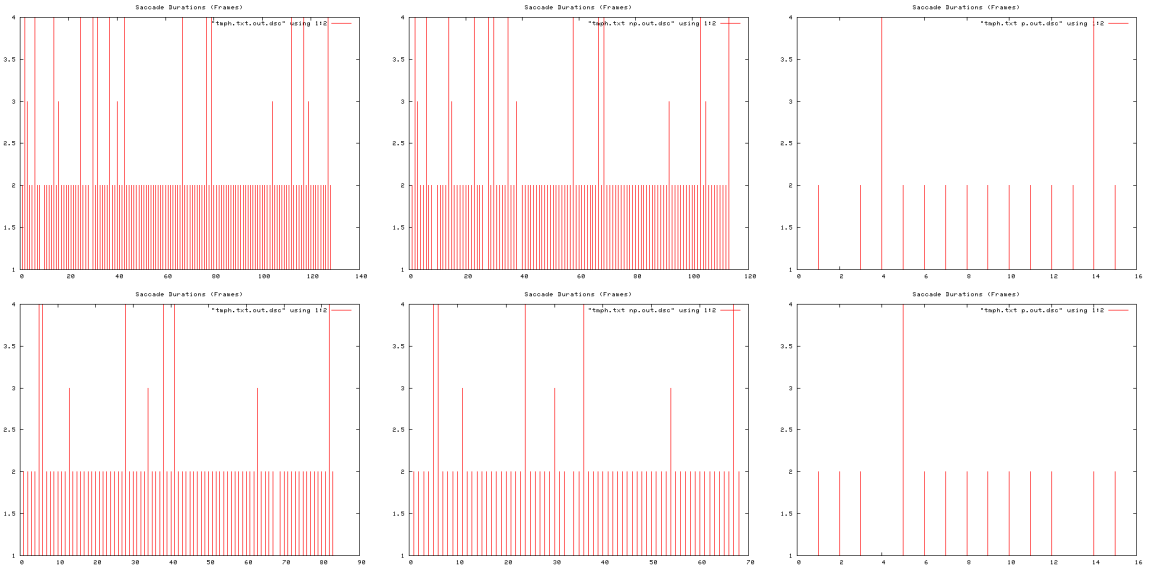



Figure 9.28: Saccade durations. Pilot 1 (top), and Pilot 2 (bottom). Entire trial (left), during periods of non-perturbation (middle), and during perturbation (right).

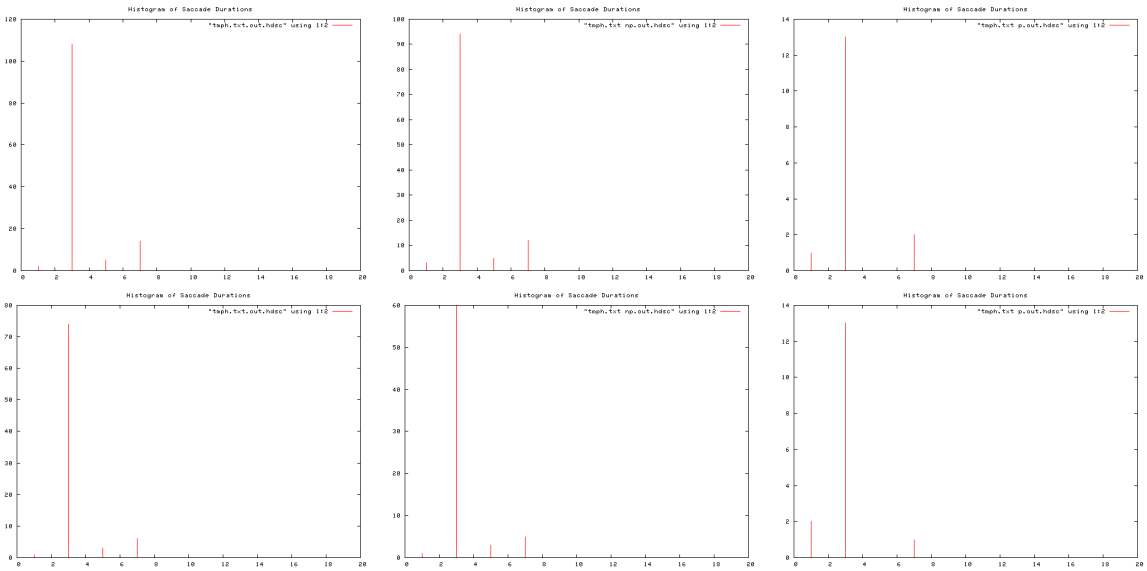



Figure 9.29: Histogram of saccade durations. Pilot 1 (top), and Pilot 2 (bottom). Entire trial (left), during periods of non-perturbation (middle), and during perturbation (right).

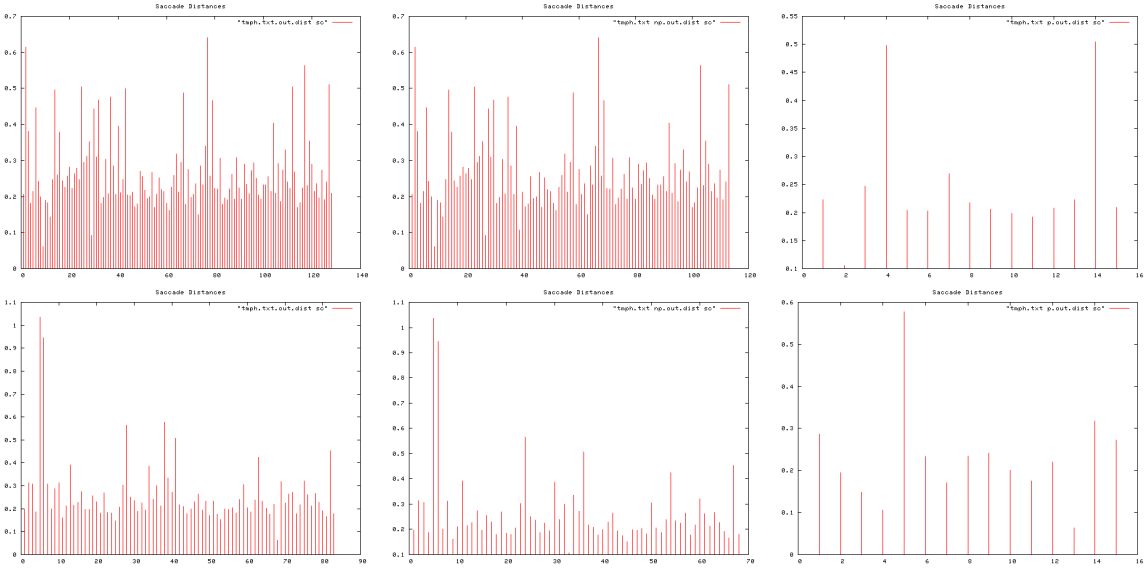



Figure 9.30: Saccade distances. Pilot 1 (top), and Pilot 2 (bottom). Entire trial (left), during periods of non-perturbation (middle), and during perturbation (right).

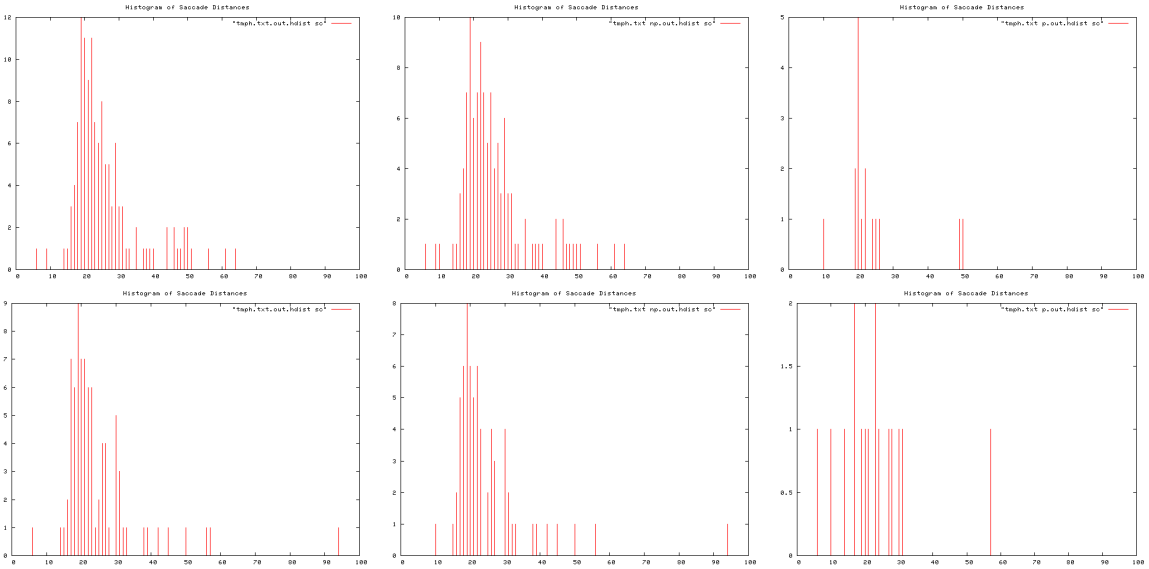



Figure 9.31: Histograms of saccade distances. Pilot 1 (top), and Pilot 2 (bottom). Entire trial (left), during periods of non-perturbation (middle), and during perturbation (right).

of the re-attention period across all four objects in a trial was used as a measure of re-attention period consistency for that trial. Figure 9.32 shows return-to-object incidences during non-perturbed periods over the course of a trial, and the use of this data to determine the re-attention period and the enumeration of consistency in the re-attention period across objects in that trial.

| Ti | To | T | I/O Event | Pear | Apple | Peach | Orange |
|---|---|---|---|---|---|---|---|
| 12 | 17 | 5 | Orange In | | | | 1 |
| 20 | 23 | 3 | Pear In | 1 | | | 1 |
| 29 | 43 | 4 | | 6 | | | 5 |
| 46 | 49 | 3 | Peach In | 1 | | 1 | 1 |
| 57 | 1:03 | 6 | | 2 | | 1 | 1 |
| 1:07 | 1:12 | 5 | | 1 | | 1 | 1 |
| 1:16 | 1:23 | 7 | | 2 | | 3 | 2 |
| 1:29 | 1:35 | 6 | Apple In, Peach Out | 1 | 2 | 1 | 1 |
| 1:39 | 1:47 | 8 | Orange Out | 2 | 2 | | 3 |
| 1:49 | 1:58 | 9 | Pear Out | 3 | 5 | | |
| 2:02 | 2:15 | 13 | Apple Out | | 3 | | |
| | | | TOTAL Rets | 19 | 12 | 7 | 16 |
| | | | TOTAL T | 51 | 36 | 27 | 47 |
| | | | R/T | 0.37 | 0.33 | 0.26 | 0.34 |
| | | | **Av. Re-attention Period** | **2.7** | **3** | **3.8** | **2.9** |
| | | | **SD(Av. Pr)** | **0.48** | | | |

Figure 9.32: Object re-attention during non-perturbed periods, Pilot 1.

### 9.4.2 Extracting Behavioural Parameters

Having considered the pilot trials empirically, we now stipulate parameters suitable in characterising human gaze behaviours as observed during the trials. It is likely that large inter-trial differences in absolute parameters may be present. For example, if object speeds are generally higher in one particular trial, it may influence the average smooth pursuit duration and average smooth pursuit velocity for that trial. Such absolute parameters may also vary across trials according to a participant's mood and alertness, and the precision of trial calibration. However, the trend of how these rates change between periods of perturbation and non-perturbation are more likely to show consistency. We therefore rely on parameter ratios, rather than absolute levels. We extract absolute parameters and compare the ratio change in parameters between periods of perturbation and non-perturbation. Absolute parameters extracted in each trial are now listed.

Duration parameters:

- $Spt_p$, $Spt_{np}$: av. smooth pursuit duration during times when things are being perturbed and not being perturbed respectively.
- $Sct_p$, $Sct_{np}$: av. saccade duration during times when things are being perturbed and not being perturbed respectively.

Distance parameters:

- $Spl_p$, $Spl_{np}$: av. smooth pursuit distance during times when things are being perturbed and not being perturbed respectively.
- $Scl_p$, $Scl_{np}$: av. saccade distance during times when things are being perturbed and not being perturbed respectively.

Velocity parameters:

- $Spv_p$, $Spv_{np}$: av. smooth pursuit velocity during times when things are being perturbed and not being perturbed respectively.
- $Scv_p$, $Scv_{np}$: av. saccade velocity during times when things are being perturbed and not being perturbed respectively.

Saccade proportion parameters:

- $Scp_p$, $Scp_{np}$: proportion of frames that are above the saccade velocity threshold during a trial when things are being perturbed and not being perturbed respectively.

Re-attention period parameter:

- $Pr$ = av. re-attention period for all objects during non-perturbation periods.

Once these absolute parameters have been extracted from a trial data log, we obtain the ratios for comparison across trials:

- $Spt_r = Spt_{np}/Spt_p$: rate change in average smooth pursuit duration time from when things are perturbed (P) to when they are not perturbed (NP).
- $Sct_r = Sct_{np}/Sct_p$: rate change in average saccade execution time from P to NP.
- $Spl_r = Spl_{np}/Spl_p$: rate change in average smooth pursuit distance from P to NP.
- $Scl_r = Scl_{np}/Scl_p$: rate change in average saccade distance from P to NP.
- $Spv_r = Spv_{np}/Spv_p$: rate change in average smooth pursuit velocity from P to NP.
- $Scv_r = Scv_{np}/Scv_p$: rate change in average saccade velocity from P to NP.
- $Scp_r = Scp_{np}/Scp_p$: rate change in average saccade proportion from P to NP.

For the re-attention period, we are more interested in measuring coherence/ consistency to this parameter during a trial. We use the standard deviation of object re-attention periods within each trial as a metric to estimate coherence to a constant re-attention period:

- $Pr_{sd}$ = STD_DEV($Pr_n$), (where n = 0...4 corresponding to each of the four different objects in a trial): re-attention consistency measure.

These eight rate parameters form the basis of subsequent numeric behavioural comparisons. A script has been written to automatically extract these parameters from the data for each trial. The script also automates plotting of all plots described in the empirical analysis. Sample output produced by the automatic script is shown in Table B.4. All output and plots are included in Appendix B. If we have selected meaningful parameters, we would expect to see consistency in trends in the ratio values in the last column of Table B.4 across trials in the broader study.

Table 9.1: Sample parameter extraction output, Pilot 1 (units omitted).

| P Param | Val | NP Param | Val | Ratio Param | Val |
|---|---|---|---|---|---|
| $Spt_p$ | 86.038462 | $Spt_{np}$ | 46.269841 | $Spt_r$ | 0.537781 |
| $Sct_p$ | 2.187500 | $Sct_{np}$ | 2.228070 | $Sct_r$ | 1.018546 |
| $Spl_p$ | 0.115915 | $Spl_{np}$ | 0.092618 | $Spl_r$ | 0.799019 |
| $Scl_p$ | 0.247815 | $Scl_{np}$ | 0.266183 | $Scl_r$ | 1.074120 |
| $Spv_p$ | 0.476394 | $Spv_{np}$ | 0.484677 | $Spv_r$ | 1.017385 |
| $Scv_p$ | 6.797215 | $Scv_{np}$ | 7.168081 | $Scv_r$ | 1.054561 |
| $Scp_p$ | 1.564595 | $Scp_{np}$ | 4.356028 | $Scp_r$ | 2.784124 |
| | | | | $Pr_{sd}$ | 0.483045892 |

## 9.5 Results

In addition to the two pilots, 20 trials were conducted. All participants reported seeing one apple during the trial. Questionnaire responses and trial data logs were obtained for each trial.

### 9.5.1 Questionnaire Responses

The responses indicate that all participants were considered to constitute valid data sets. According to the questionnaires, the mean age of participants was 31 years. Ages ranged from 23 to 53 years with a standard deviation of 7.7 years.

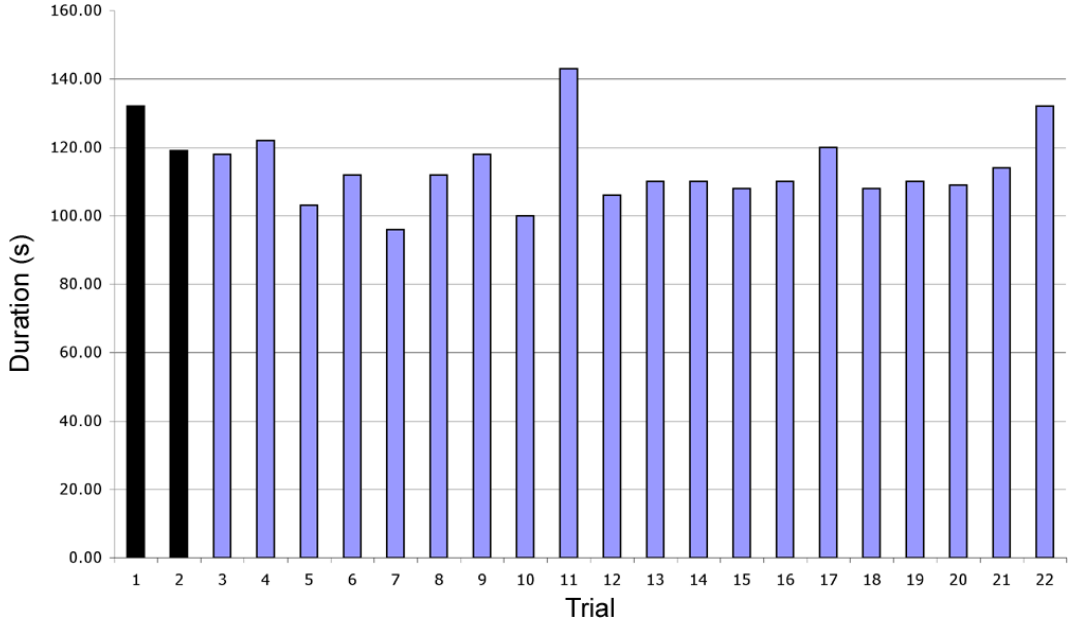



Figure 9.33: Trial execution durations. Columns 1 and 2 correspond to pilot trials.

Ethical confidentiality means we cannot reproduce questionnaire responses in this thesis.

### 9.5.2 Trial Logs

All trials ran according to the storyboard. Figure 9.33 shows storyboard execution duration for each trial. Trial duration was measured from the time the first object entered the scene to the time the last object reaches its final resting place. Average trial duration was 114.18s; the standard deviation in trial durations was 10.95s. Trial duration consistency suggests that timing and velocities of storyborad object motions is likely to have been consistent over all trials.

Figure 9.8 shows a sample screenshot of a trial's video log. Appendix C shows video logs of each trial synchronised with the 2D projection display of FaceLAB gaze data. No significant changes in participant behaviour was observed over the course of each trial. Appendix B shows the output of the automatic processing script for each trial, including plots and parameter extraction.

## 9.6 Analysis

We now consider the data obtained in the course of all trials. We first consider the results empirically, then numerically.

### 9.6.1 Empirical Observations

As was expected, the data histograms show consistency across trials. The main empirical observations include:

- Movie logs show that gaze consistently saccades to the perturbed object.
- Histograms of gaze velocities in the human trials are strongly bimodal. That is, there is consistently a group of low velocities corresponding to smooth pursuit of the slower moving targets or stationary locations; and a grouping of high velocities corresponding to saccades.
- Velocity histograms show the proportion of frames exhibiting above saccade threshold velocities consistently increases during periods of no object perturbation.
- Average saccade velocities, distances and durations are similar for periods of perturbation and non-perturbation.
- Histograms of smooth pursuit distances show that a lower proportion of short distances exist during periods of perturbation than non-purturbation.
- The distribution of smooth pursuit gaze points during perturbation periods correspond well to the paths of perturbed objects.
- Re-attention periods were largely constant for all objects within an individual trial.

In general, a significant increase in the rate of saccades during periods of non-perturbation was observed. Other saccade characteristics were *not* observed to vary significantly between periods of non-perturbation and perturbation. Smooth pursuit characteristics *were* observed to vary significantly between periods of non-perturbation and perturbation.

### 9.6.2 Numerical Characterisation

We now qualify the empirical observations made in the previous section. We characterise general gaze behaviours during trials via inter-participant statistics. Data extracted from each individual trial is provided in Appendix B.

We can compile histograms for each of the extracted parameters from all trials. For example, Figure 9.34 (left) shows a histogram of the rate of increase in the proportion of saccade frames from periods of perturbation to non-perturbation (parameter $Scp_r$). The ratio parameters were selected because they are less susceptible to being influenced by inter-participant inconsistencies (such as mood) and so extracted parameters may be treated as independent samples from the same underlying PDFs (PDFs that are more likely to be similar). The small sample size (20 trials) makes it difficult to confirm that the underlying PDFs associated with extracted rate parameters adhere to normal distributions. For example, the histogram of extracted $Scp_r$ parameters across trials (left, Figure 9.34) is inconclusive. We therefore conduct standard JB-tests and KS-tests [Mitchell (1997)] for PDF normality. Both JB-test and a KS-test failed unless less restrictive thresholds are chosen than standard (both fail for most rate parameters with usual significance level = 0.05). Again, this is probably due to the small sample size of 20 participants. A normality plot for each rate parameter across all trials (for example, for $Scp_r$ – right, Figure 9.34) is non-linear, also suggesting that the underlying PDFs are non-normal (normality plots for all rate parameters are presented in Appendix B). For all of these reasons, it is not safe to assume the underlying rate parameter PDFs conform to normal distributions, and normal theory is not suitable for characterising numerical observations.

Therefore, we "bootstrap" [Efron & Tibshirani (1993)] the mean and variance on each parameter. Bootstrapping does not rely upon normally distributed PDFs. The expected distribution of the mean and associated variances for all rate parameters were subsequently calculated using the bootstrapping technique. 95% Confidence intervals (CIs) in the mean and standard deviation statistics were calculated using Matlab bootstrap functions.

Table 9.2 summarises the bootstrapped 95% CI on the mean for each rate parameter, and the bootstrapped 95% CI on the standard deviation for each parameter; calculated over all data from all human trials. The last column is based

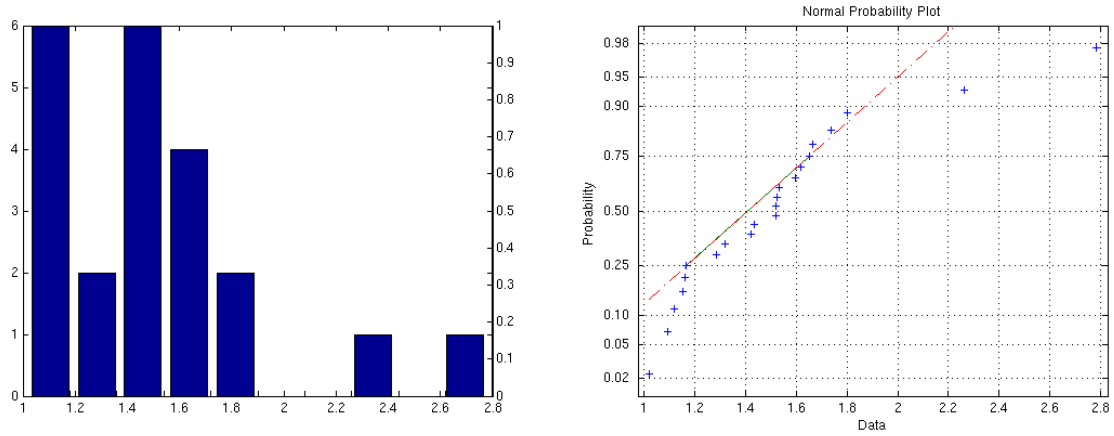



Figure 9.34: $Sc_r$ parameter. Histogram (left); and normality plot (right).

on the previous columns and indicates whether the parameter rate characteristically is expected to increase (+) or decrease (-) when transitioning from perturbed (P) to non-perturbed (NP) scene. For example, if one sees a '+' at $Scp_r$, one would expect to see that parameter rise (on average) when things are not being perturbed, compared to when they are. The last parameter, the re-attention period coherence parameter $Pr_{sd}$ is an absolute measure across the entire trial; it is not a ratio of P to NP but has nonetheless been added to the summary table as a relevant statistic.

A plot of the distribution of the density of expected means for parameter $Scp_r$ is shown in blue in Figure 9.35. The bootstrapped 95% CI for the bootstrapped mean density of $Scp_r$ ranges between a lower bound of $CI_lb = 1.3709$ and an upper bound of $CI_ub = 1.6841$, as highlighted in blue. The lower and upper bounds of the 95% CI on two standard deviations is highlighted in red around this interval. For comparison, we may also calculate the mean and standard deviations according to normal theory. These parameters should not be trusted for statistical analysis. For example, for $Sc_r$ parameter, the normal theory mean and standard deviation are 1.5169 and 0.4040 respectively. We have plotted this normal distribution on the same plot (green curve, Figure 9.35), and we have highlighted the interval associated with the two normal theory standard deviations (green in Figure 9.35). The bootstrapped standard deviation statistics suggest a broader distribution than normal theory.

Table 9.2: Parameter changes when going from P to NP.

| Parameter | $Mean\ CI_l$ | $Mean\ CI_u$ | $SD\ CI_l$ | $SD\ CI_u$ | $+/-$ |
|---|---|---|---|---|---|
| $Spt_r$ | 0.8273 | 1.4045 | 0.1024 | 0.4297 | = |
| $Sct_r$ | 1.0360 | 1.1559 | 0.0907 | 0.1881 | = |
| $Spl_r$ | 0.7691 | 0.9077 | 0.1187 | 0.2034 | – |
| $Scl_r$ | 1.1477 | 1.6520 | 0.1903 | 0.8471 | + |
| $Spv_r$ | 0.9621 | 1.0125 | 0.0414 | 0.0783 | – |
| $Scv_r$ | 1.0906 | 1.3583 | 0.1113 | 0.4904 | = |
| $Scp_r$ | 1.3738 | 1.6944 | 0.2027 | 0.5708 | + |
| $Pr_{sd}$ | 0.3502 | 0.5189 | 0.1162 | 0.2922 | n/a |

Tendencies - '+': increase, '=': unchanged, '–': decrease.

The bootstrapped distribution of the density of expected means for parameter $Scp_r$ appears somewhat Gaussian (blue curve, Figure 9.35). However, we can plot a Gaussian with standard deviation $sd_m = sd/\sqrt{n_t}$ ($n_t$ = number of trials) for comparision. The red dashed plot in Figure 9.35 shows this Gaussian. It can then be seen that the bootstrapped density of means exhibits skew towards lower values.

The same bootstrapping procedure was executed for each rate parameter. Associated plots can be found in Appendix B. Table 9.2 summarises bootstrapped statistics as indicated earlier.

## 9.7 Discussion

The numerical characterisation conducted over all trials confirmed the general empirical observations. We briefly discuss behavioural characteristics based on both empirical and numerical observations. Bootstrapped mean ratios can be used to numerically establish expected behaviours. Where possible, we relate behavioural expectations to the system model described in Chapter 3. The trials are not designed to test what the components of the underlying model are; we may merely relate observed behaviours to the predictions the model may provide. The bootstrapped standard deviations may indicate which rate parameters are

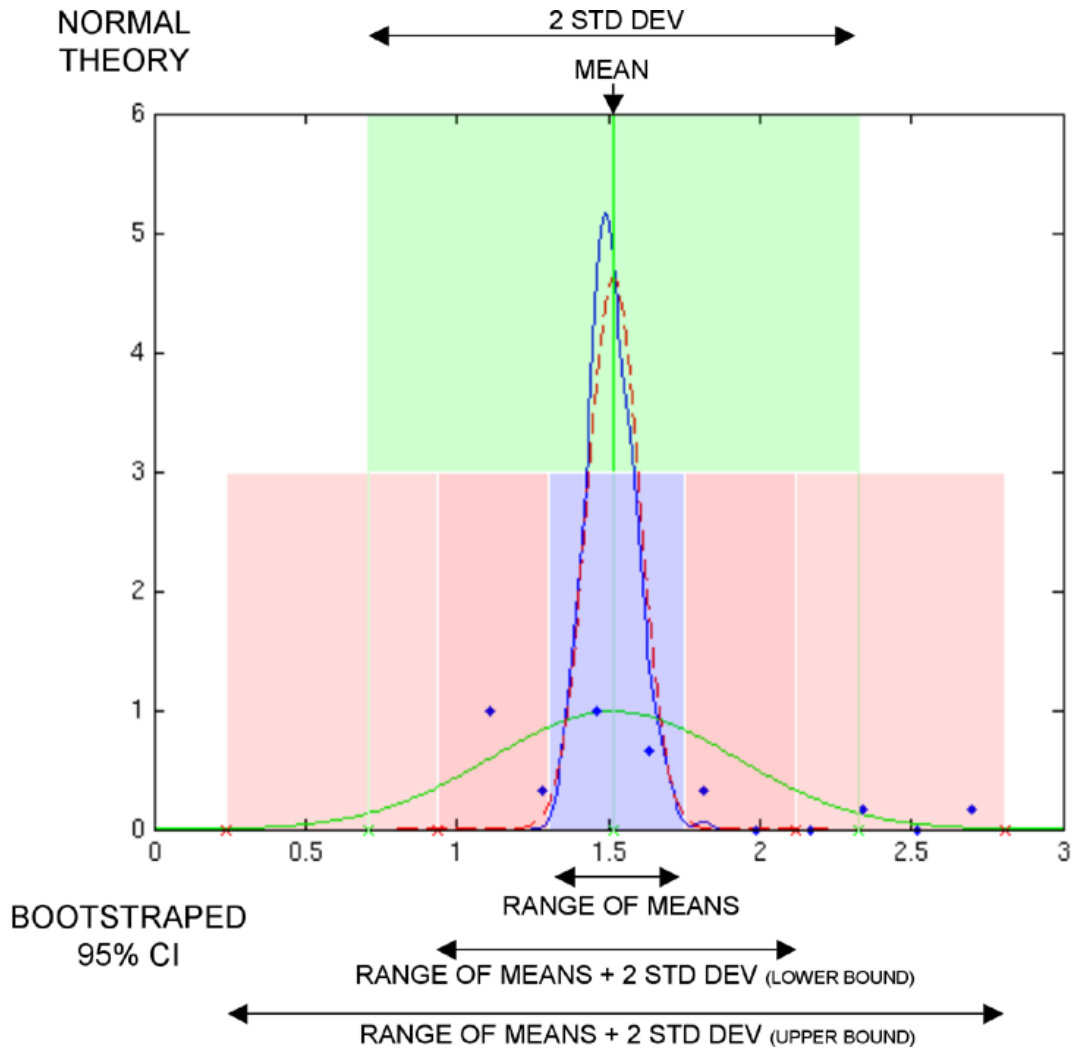



Figure 9.35: Interpreting bootstrap results, $Sr$ parameter. The green lines and shading represent normal theory approximations. The blue and red components show bootstrap results. The blue dots are the superposition of rate parameter histograms. Appendix B shows all additional trial parameters.

more consistent across participants (hardware or scene-dependent), and which are more variable (dependent on the participant). We first discuss the observed saccade proportion ratio, then saccade characteristics, smooth pursuit characteristics, and the re-attention period. No significant age-related trends or variations were observed.

### 9.7.1 Saccade Rate Characteristics

When an object in the scene is being actively perturbed, the saccade proportion rate has been observed to decrease in general. When no perturbations are present, the saccade proportion rate tend to increase. This is evident in the histograms, and has been shown numerically ($Scp_r > 1.0$). Some variance across participants is evident in the comparatively medial bootstrapped standard deviation in this rate parameter. The tendency is for the parameter to increase, the amount of increase is dependent on the participant.

The trials conducted cannot determine the underlying *cause* for this observation. Nevertheless, we described a model of saliency in the previous chapter where fixation is selected by modulating IOR with view-frame saliency, and we may relate the observations to this model. According to this model, a likely explanation for the increase in saccade proportion rate is that participants inhibit previously attended salient regions of the scene such that the product of inhibition and saliency equalises over the entire scene. When scene saliency is relatively static (periods of no perturbation), the model may predict more saccades between similarly salient regions. This is an effect of dynamic IOR: after a short while all objects in the scene are attended and largely suppressed, so the product of saliency and IOR evens out over the entire scene and there is no distinct salient peak for the system to track; instead, the participant's saccade rate increases because most locations exhibit a similar response. When an object is being actively moved, its saliency is initially increased due to that motion, beyond suppression, and it wins attention for the period that it begins to move. Then, although it is moving, it begins to be suppressed - but it continues to be tracked in the fovea (despite its motion) until it is fully suppressed, or another more salient object is detected, which may not be until after the moving object is no longer being perturbed (perhaps some swinging remains). Also, the default action is to continue

to fixate upon the current target, so even after it is no longer a fixation peak, it can continue to be tracked until an alternate peak passes fixation moderation.

### 9.7.2 Saccade Characteristics

We would expect that saccade characteristics (such as average saccade velocity, distance and duration) should remain similar throughout the duration of a trial, because saccades involve fixation shifts in a minimal amount of time, at a maximal velocity (hardware dependent). Histograms of these parameters support this expectation.

We have observed a relatively even distribution of saccade destinations over the entire scene window (Figure 9.18). As expected, the location of smooth pursuit locations (static fixations - middle, Figure 9.16) also correlates well with the location of saccade destinations. According to the system model, we may expect to see a slight increase in parameters $Sct_r$ and $Scl_r$ because when a perturbed object enters the scene or moves, it will not always be a large distance from the current fixation point (beyond the inhibited region surrounding the current fixation point), and it may even be fixated upon before it moves; whereas saccades are likely be to regions where inhibition is not accumulating. We may therefore expect parameter $Scl_r$ to be approximately 1.0, or perhaps slightly larger. This expectation was confirmed by the trials. Saccade durations were consistent across periods of perturbation and non-perturbation (parameter $Sct_r$ approximately 1.0), indicating it is not significantly dependent on variations in the scene. The low bootstrapped standard deviation in parameter $Sct_r$ across participants may indicate that this parameter is somewhat hardware dependent (more dependent on ocular muscle performance). Parameter $Scl_r$ was in general observed to increase during periods of no perturbation, suggesting some scene dependancy. The comparatively large bootstrapped standard deviation in parameter $Scl_r$ also suggests the amount of increase is largely dependent on the participant.

If the saccade velocity is hardware-related (dependent on the performance of eye muscles), we would expect saccade velocity parameter $Scv_r$ to be approximately 1.0, and saccade duration parameter $Sct_r$ to be correlated with saccade distance parameter $Scl_r$ (also approximately 1.0). Indeed, the numerical analysis confirms these expectations. Parameter $Scv_r$ was reasonably constant across

periods of perturbation and no perturbation, its density of means centred about 1.0. This indicates the parameter is not significantly scene-dependent.

### 9.7.3 Smooth Pursuit Characteristics

We would of course expect smooth pursuit distances to reduce during non-perturbation periods because no objects are translating. Figure B.2.1.4 shows that a lower proportion of short smooth pursuit distances exist during periods of perturbation than non-purturbation. Inspection of the 2D distribution of gaze points over the scene window (Figure 9.16) indicates the storyboard motion paths correlates best with smooth pursuit locations during perturbation. These observations indicate that participants have a strong tendency to smoothly pursue actively perturbed objects. Indeed, after reviewing the video logs, actively perturbed objects were consistently and rapidly attended, then tracked.

Calibration differences mean we cannot compare absolute smooth pursuit distances directly across trials, but similar smooth pursuit distance histograms during perturbation periods may indicate the stimulus were equally as compelling for all participants to track. Indeed, all such histograms exhibit similar appearance. We would expect them to be somewhat similar because all candidates were briefed and prepared for the trials in the same manner. If some participants had shown consistently short smooth pursuit distances during perturbation periods, and others had shown long ones, it may have suggested that the participants were not responding to the stimulus similarly. This was not the case. Numerically, the low bootstrapped standard deviation on parameter $Spl_r$ (relative to the bootstrapped standard deviation of other parameters) also indicates that smooth pursuit characteristics were similar across participants.

It was expected that the smooth pursuit durations would differ in periods of perturbation and periods of non-perturbation. Participants preferentially attended moving objects so that smooth pursuit durations during perturbation are more dependent on the time objects are translating. During non-perturbation periods they are correlated more with the saccade rate, vis a vis, the time a participant tended to linger gaze at the same location. Participants whose gaze behaviours were more saccadic may therefore be expected to show an increase in smooth pursuit duration from periods of non-perturbation to perturbation

($Spt_r > 1.0$). Conversely, less saccadic participants may be expected to show a reduction because the translation of a perturbed object may last less than their typical non-perturbation linger period ($Spt_r < 1.0$). Therefore, rate parameter $Spt_r$ was expected to vary across participants. The ratio of smooth pursuit durations from periods of perturbation to periods of non-perturbation ($Spt_r$) varied significantly across participants, as characterised by the comparatively large bootstrapped standard deviation. There was generally no distinct tendency for the parameter to increase or decrease, but the bootstrapped mean was centred about 1.0. This parameter is therefore largely dependent on the participant.

Smooth pursuit velocities were expected to go down during periods of no perturbation because no translating objects were present to track. Error in gaze locations provided by FaceaLAB may mean that zero velocity frames were recorded as non-zero (when gaze was stationary FaceLAB tended to return a gaze path that oscillated around the stationary point; see trial videos). Nevertheless, a slight tendency for the smooth pursuit velocities to decrease was indicated by the numerical analysis.

As expected, smooth pursuit characteristics are largely dependent on the motion of objects in the scene. The smooth pursuit distance and velocity ratios from periods of perturbation to periods of non-perturbation ($Spl_r$ and $Spv_r$) both consistently decreased ($< 1.0$), commensurate with the tendency for participants to track translating stimuli. This statistic, and the comparatively small bootstrapped standard deviations on these parameters characterise a generally similar amount of decrease across all participants, and demonstrate that these parameters are largely scene-dependent.

### 9.7.4 Re-attention Period Characteristics

The average re-attention period loosely correlated with the saccade rate for a given trial. It enumerates the number of occasions an object is attended divided by the total time the object is present during non-perturbation periods. This parameter has been demonstrated to be remarkably constant for all objects in a single trial, and consistently coherent over all trials. One participant may re-attend objects approximately once every three seconds, for example, and another participant may be less saccadic and only attend each object approximately once

every six seconds, but for each participant the number was observed to be consistent across all objects in a trial. This may suggest that all participants found all the stimuli objects (fruit) similarly salient when no active perturbation was present.

Re-attention period consistency may be predicted by the system model, largely as an effect of dynamic IOR, according to our model, as an object becomes inhibited, gaze is transferred to the next peak of IOR modulated saliency that passes moderation. Because IOR decays, previously attended objects may again be fixated upon. Thus, the model may cause cycling through the dominantly salient scene locations. If the objects are similarly salient, and the scene is suitably static, each object should then pass fixation moderation periodically .

The low mean of coherence parameter $Pr_{sd}$ shows the general tendency for participants to attend all four objects within their trial with similar frequency. In particular, the low bootstrapped standard deviation for the coherence parameter across all trials shows this behaviour is consistent across participants. This further demonstrates that, regardless of the average re-attention period for each trial, all objects were attended with similar frequency within each trial.

The average standard deviation of all the re-attention periods is 0.43s. The standard deviation of the re-attention period for all objects in all trials is 1.92s, much higher than the average standard deviation of individual trials. This is also evidence that re-attention period varies greatly over all objects in all trials, but is relatively constant within each trial.

The trials exhibiting the two largest standard deviations in re-attention period are the two trials with longest re-attention periods. This may suggest that participants exhibiting longer re-attention periods also had slightly more variation in re-attention period rate. Hence there may be a relationship between re-attention period and re-attention coherence. This is intuitively plausible: a longer average attention period is likely to yield more variance in the re-attention period across objects. If this is the case, we could then normalise the re-attention period standard deviations for each trial according to the trial's average re-attention period, which would likely show further conformity in all re-attention period consistency metrics. However, a graph plotting the re-attention standard deviation versus average re-attention period for each trial is not conclusive (no clear relationship, Figure 9.36). More trials would be needed to confirm such a relationship.

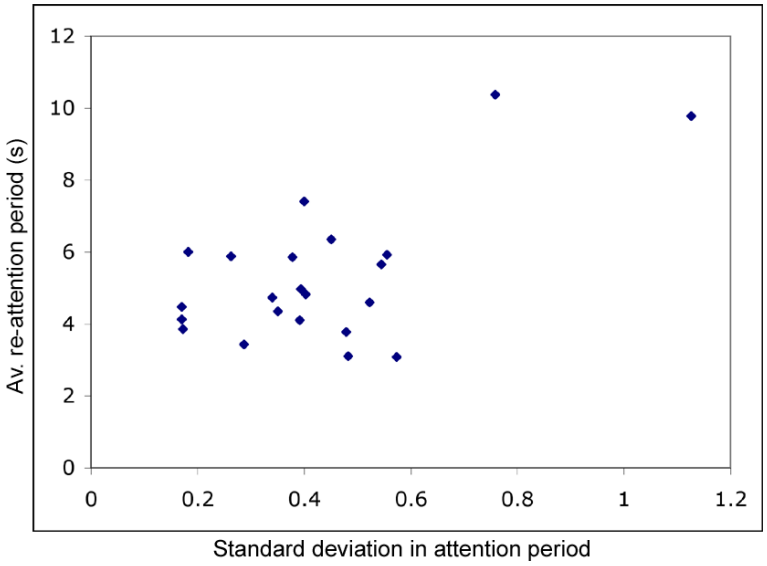



Figure 9.36: Av. re-attention period vs re-attention period standard deviation for each trial.

## 9.8 Summary

Psycho-physical trials were conducted to establish and extract metrics that characterise unconstrained human gaze behaviours when viewing a simple, repeatable, controlled 3D scene. The scene incorporated stationary and translating stimuli that were considered approximately equally salient and non-iconic. All humans were observed to react to the stimulus in a quantifiably similar manner. All participants' distance-weighted velocity magnitude histograms were distinctly bimodal, exhibiting a group of low velocities (corresponding to smooth pursuit motions after inspecting the video logs) and a group of high velocities (corresponding to saccades) separated by a range of sparsely occupied medial velocities.

During periods of perturbation, participants preferentially smoothly pursued translating stimulus. Accordingly, saccade frequency was observed to decline during periods of object perturbation, and increase during non-perturbed periods. During non-perturbation periods, gaze also frequented the locations corresponding to objects more than the background. Smooth pursuit locations and saccade destination locations during non-perturbation corresponded well with the location of objects.

Due to inter-participant differences (mood, alertness, etc), absolute parameter values (such as extracting average smooth pursuit velocities in a *single* trial to determine an average smooth pursuit velocity over *all* trials) were not considered to yield good metrics for characterising gaze *behaviours*. We therefore examined the *ratio* of such parameters across extractable modes of stimuli presentation. We compared parameter ratios from periods of perturbation (translating stimulus present) to periods of non-perturbation (no translating stimulus present). Analysis of extracted inter-participant behavioural rate parameters shows the following characteristic trends:

- The ratio of saccades from periods of perturbation to periods of non-perturbation ($Scp_r$) constantly increased ($> 1.0$), characterising the consistent tendency for the saccade rate to increase during periods of non-perturbation across all participants. Reasonable variance in this parameter across participants is shown statistically by the comparatively medial range in the parameter's bootstrapped standard deviation. Therefore, the amount of increase in saccade rate was somewhat dependent on the participant.

- The smooth pursuit distance and velocity ratios from periods of perturbation to periods of non-perturbation ($Spl_r$ and $Spv_r$) both consistently decreased ($< 1.0$), commensurate with the tendency for participants to track translating stimuli. This statistic, and the comparatively small bootstrapped standard deviations on these parameters characterise a generally similar amount of decrease across all participants, and demonstrate that these parameters are largely scene-dependent.

- The ratio of smooth pursuit durations from periods of perturbation to periods of non-perturbation ($Spt_r$) varied significantly across participants, as characterised by the comparatively large bootstrapped standard deviation. There was generally no distinct tendency for the parameter to increase or decrease, but the bootstrapped mean was centred about 1.0. This parameter is therefore largely dependent on the participant.

- No significant change in saccade durations was detected across periods of perturbation and non-perturbation (bootstrapped distribution of means on

parameter $Scv_r$ was centred approximately at 1.0), suggesting that this parameter is not significantly dependent on the scene. The low bootstrapped standard deviation in the parameter across participants suggests that it is largely dependent on hardware (ocular muscle agility).

- No significant change in saccade velocities was detected across periods of perturbation and non-perturbation (bootstrapped distribution of means on parameter $Sct_r$ was centred approximately at 1.0), suggesting that this parameter is not significantly dependent on the scene. The comparatively large bootstrapped standard deviation in this parameter across participants suggests that it is largely dependent on the participant.

- The saccade length tended to increase from periods of perturbation to periods of non-perturbation ($Scl_r > 1.0$). The bootstrapped standard deviation in the parameter was also comparatively large. The tendency to increase suggests some general scene dependency, but the amount of increase depends largely on the participant.

- The average re-attention period for each participant varied significantly. Object re-attention periods were approximately constant during periods where no object was being actively perturbed for each participant.

The observed parameter trends and associated statistics serve to benchmark the inter-individual consistencies in the human gaze behaviours elicited during the pshyco-physical trials.

# Chapter 10

# Synthetic Trials

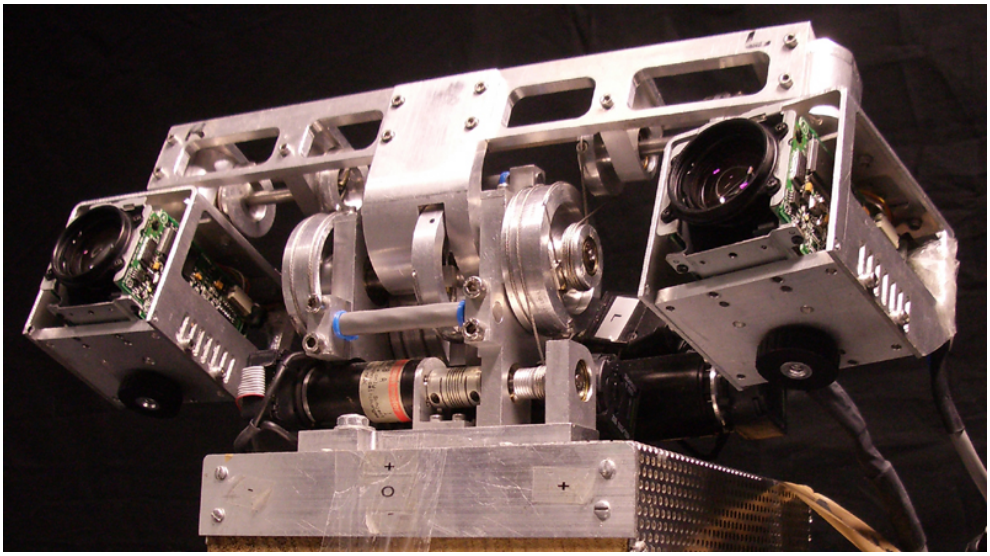


Figure 10.1: CeDAR participating in synthetic trials.

In this chapter we conduct trials with the synthetic vision system. We compare behavioural characteristics of the synthetic trials to the benchmarks obtained from the human trials.

## 10.1 Introduction

We conduct the same analysis of gaze behaviour using the synthetic system in place of human participants. Rather than conducting trials with different indi-

viduals, we conduct trials using the synthetic system with different configuration settings. Metrics extracted from the synthetic trials are used to compare system behaviours with those established from the human trials. We also use the synthetic trials to determine if predictions based upon the adopted model conform to experimental observations.

We analyse the synthetic trials in the same manner as the human trials. The same absolute and ratio parameters are extracted for comparison. Similarities in the underlying human and synthetic system models should elicit similar basic gaze behaviours. Similarities can be confirmed if rate parameters extracted from the individual synthetic trials fall within the variance of the benchmark human parameters (system tuning may also yield rate parameters that better conform to the human benchmarks). In this manner, we may select system configurations that produce the trial whose behaviours best match the human benchmarks.

We may also expect that system behaviours depend largely on the system itself, and its biological inspiration, not just on the specific configuration settings selected for a particular trial. We therefore consider inter-trial consistency of the synthetic trials for comparison to the human benchmarks.

## 10.2 Synthetic Trials

Synthetic trials were conducted using the same apparatus and stimuli as the human trials in the previous chapter. The same storyboard was used to reproduce the translation of stimuli. CeDAR was placed in the viewing booth such that its cameras were positioned in the location where participants' eyes had been during the human trials. Figure 10.1 shows CeDAR positioned accordingly. CeDAR (and the participants in the human trials) was situated in a stationary manner, such that involuntary vestibulo-ocular reflexes were not elicited, and the vestibulo-ocular reflex was not incorporated into synthetic system operation.

Delivering the visual task to the synthetic system is not as brief as asking a human to "count apples" because as yet the system has no means to accumulate and incorporate *a priori* knowledge autonomously. Instead, we take colour chrominance samples from pixels on the apple in the camera images. These chrominance levels were used to set the desired search colours in the colour processing server node. We may then bias the contribution of this cue more heavily than others

in the construction of the saliency map. We may also bias the response of multiple orientations on the orientation processing server node. In this manner we manually predispose the system to look for small, round objects coloured like the target apple.

### 10.2.1 Configuration Settings

Four synthetic trials were conducted, each with different configuration settings. Before the first trial was conducted, the configuration settings were set by hand to mid-range values. The first trial was then conducted, during which system performance was assessed empirically. The settings were then adjusted such that the system was likely to perform in a more human-like manner. The second trial was then conducted with adjusted settings. This process was iterated until four trials had been conducted. Figure 10.2 summarises configuration settings for each trial. Configuration settings that were left static in the course of all trials (those that did not affect gaze behaviour) are not shown.

The first trial was noticeably more saccadic than the human trials. Predictions based on the system model were used to adjust the configuration settings to reduce the saccade rate. For example, increasing the rate of accumulation of IOR over the fixation point, reducing the IOR decay rate of the entire dynamic IOR mosaic, and adopting more strict fixation map peak moderation settings were likely to lower the saccade rate.

| | Config | T1 | T2 | T3 | T4 |
|---|---|---|---|---|---|
| IOR | G_gain | 0.2 | 0.1 | 0.1 | 0.01 |
| | G_sigma | 12 | 13 | 14 | 15 |
| | IOR_dec | 0.95 | 0.96 | 0.96 | 0.98 |
| Saliency | Gain CCS | 3 | 4 | 4 | 4 |
| | Gain OCS | 1 | 1 | 1.5 | 1.5 |
| | Gain DFCS | 1 | 1 | 1.5 | 2 |
| Saccade Moderation | N_Clust | 3 | 4 | 5 | 2 |
| | N_Supersal | 1 | 1.4 | 1.3 | 1.3 |
| | Timed_shft | 2 | 3 | 3 | 5 |
| | Fix_dec | 0.95 | 0.95 | 0.98 | 0.98 |
| Tracking | ZDF T_gain | 1 | 1 | 1.1 | 1 |

Figure 10.2: System configuration variations across trials (units omitted).

### 10.2.2 Data Logging and Processing

Time-stamped angles were obtained from the CeDAR axis encoders and converted to scene-window Cartesian coordinates. As with the human trials, both CeDAR and the scene were filmed. Logged data was then processed as per the human trials.

## 10.3 Results

Four trials were conducted. Trial durations were 155, 218, 227 and 203 seconds respectively. Figure 10.3 shows online cue maps synchronised with video footage of CeDAR. The video shows (clockwise from top left) CeDAR, the view-frame saliency map, the MRF ZDF extracted object upon which fixation occurs, the dynamic IOR mosaic, and the rectified scene mosaic. Footage for each of the four synthetic trials is available in Appendix C. The synchronised footage was used to hand-mark logged data according to periods of perturbation and non-perturbation. The data was subsequently analysed as per the human data.

## 10.4 Analysis

As for the analysis of the human trials, we consider the output both empirically and numerically. We first compare the generated plots to the human trial plots. We then extract parameters from the synthetic trials to determine how coherently they compare to those extracted in the human trials.

### 10.4.1 Empirical Observations

We walk through plots generated from synthetic trial four. Plots corresponding to all trials can be found in Appendix 2.

Figure 10.4 shows the complete gaze path during synthetic trial four. The effect of the swinging motion of the stimuli can be clearly seen because of the more accurate gaze data obtained from encoders. Figure 10.5 shows histograms of the velocity magnitudes. As per the human trials, we have provided a distance weighted velocity histogram (right, Figure 10.5) so that any bimodal separation

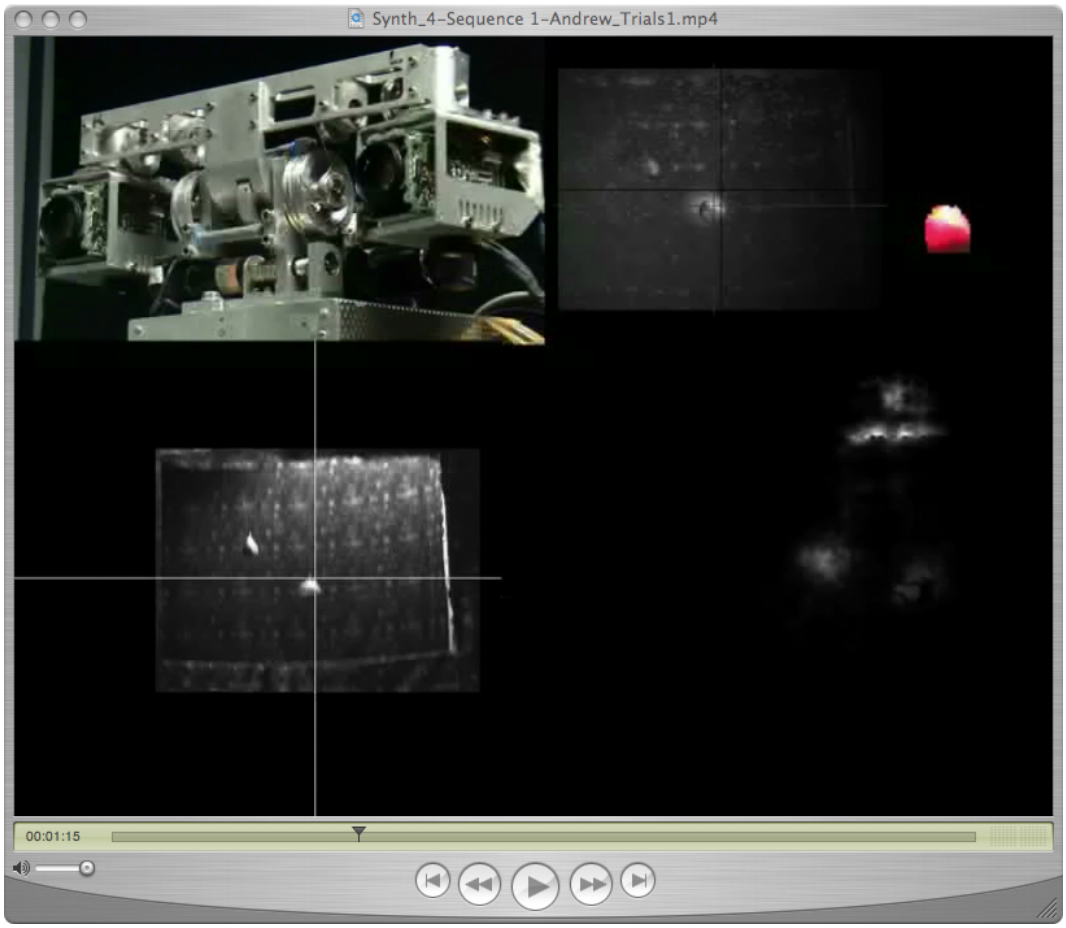



Figure 10.3: Online operation of the synthetic vision system demonstration, synthetic Trial 4 (*snapshot - see Appendix C for full video*).

of velocities can be seen more clearly. Indeed, a significant number of low and high velocities exist, separated by a region of sparse velocities, as was present in the human version (Figure 9.11). As with the human analysis, this region of sparse population allows us to select a threshold velocity above which a frame's velocity is considered saccade, and below which it is considered smooth pursuit.

It is noted that fewer "low" velocities are present in the synthetic velocity histogram than the human histogram. FaceLAB samples gaze data at 60Hz. The synthetic gaze data was recorded at below 20Hz. Velocity is calculated as the distance covered between samples divided by the time between samples. It may therefore not accurately reflect the actual gaze velocity. The lower sampling rate

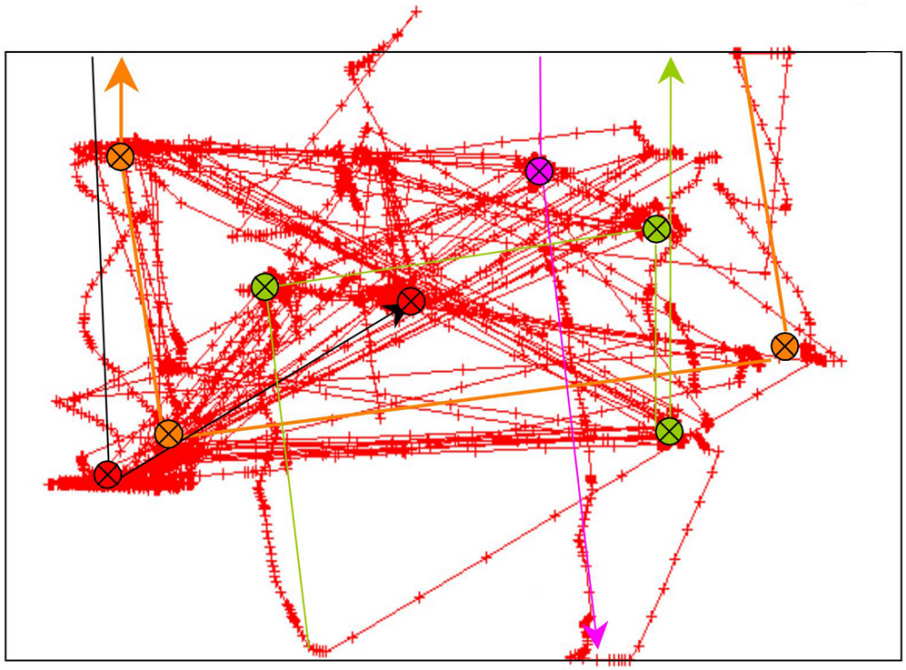


Figure 10.4: Complete synthetic scan paths with approximate object storyboard translations superimposed, synthetic Trial 4 (not to scale).

of the synthetic trials means fewer low velocity samples are accumulated but all saccades were still detected in large position shifts as the sample period was always higher than the duration between saccades. The appearance of the human histogram showing numerous low velocities may have been amplified by FaceLAB filtering. Velocity filtering may smooth (broaden) the group of saccade velocities, reducing the altitude of the high velocity peak, making the low velocity peak appear relatively higher. The more "peaky" appearance of saccade velocities in the synthetic histogram than in the human histogram may also have been due to a more repeatable maximum saccade velocity induced by the mechanical actuators. The "flatter" human saccade velocities may be elicited via reduced repeatability in the maximal velocity of muscular actuators. This would again increase the height of the low velocity peak in the human velocity histogram compared to that of the synthetic histogram. Nonetheless, the histograms were both characteristically bimodal.

Figure 10.6 shows the velocity magnitudes over the course of synthetic trial four. Saccades (blue) and periods of active perturbation (green) have been anno-

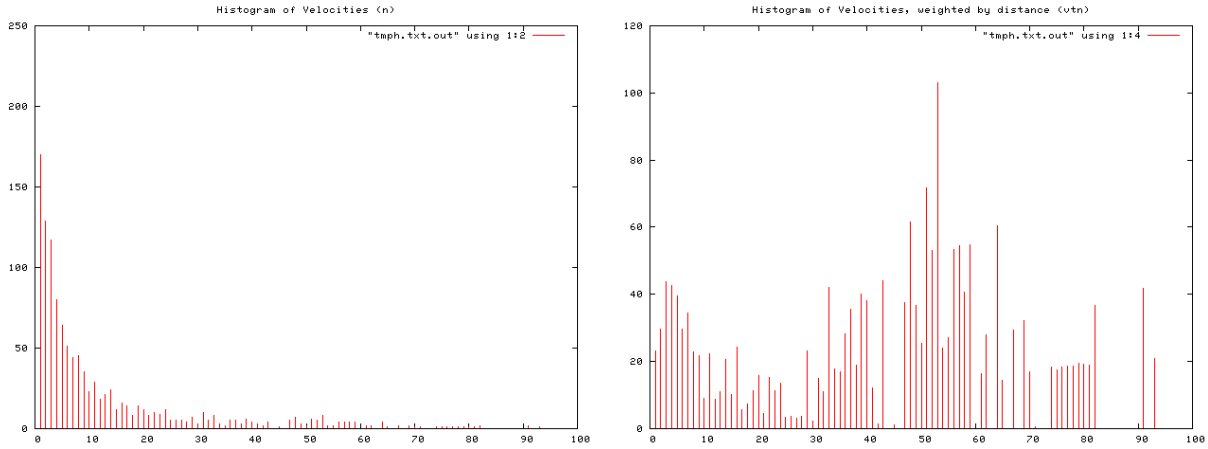



Figure 10.5: Histogram of velocity magnitudes (left). Histogram of distance weighted velocities (right).

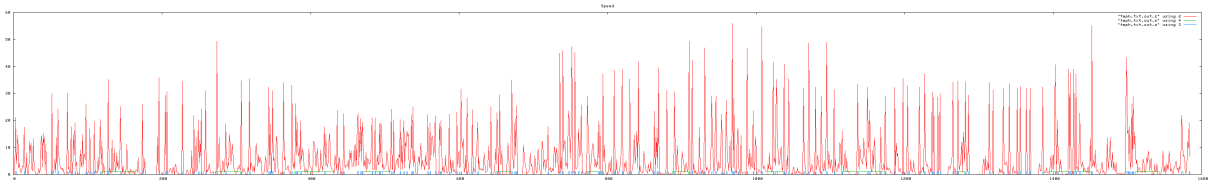


Figure 10.6: Velocity profile. Velocity magnitudes per frame, synthetic Trial 4.

tated on the profile. We may look at velocity histograms during non-perturbed (left, Figure 10.7) and perturbed (right, Figure 10.7) periods. As with the human trials, it is evident that a greater portion of velocities are below the sparse region around the saccade threshold (smooth pursuit) during periods of perturbation, and vice versa.

We can plot the coordinates of gaze locations during smooth pursuit (Figure 10.8) and saccade (Figure 10.9). As with the human trials, we see that the plot of smooth pursuit locations during perturbation shows best likeness to the prescribed storyboard motions (lines, Figure 10.4). As expected, the plot of smooth pursuit locations during non-perturbation periods, and non-perturbation saccade locations show a likeness to the stationary storyboard locations (crosses, Figure 10.4) of stimulus.

We now consider smooth pursuit velocities histograms. Figure 10.10 shows histograms of smooth pursuit velocities according to perturbed, non-perturbed

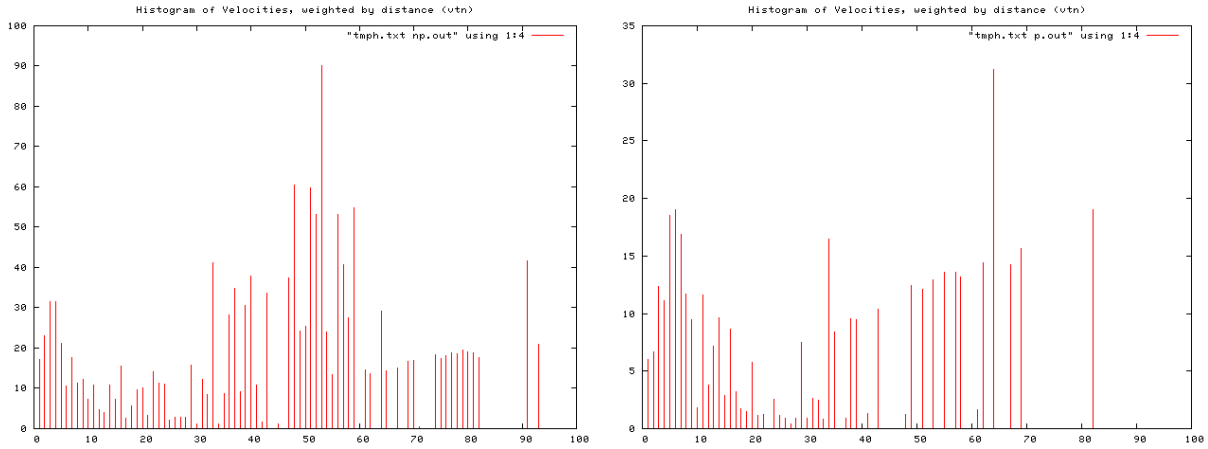



Figure 10.7: Histogram of velocities during non-perturbation (left), and during perturbation (right).

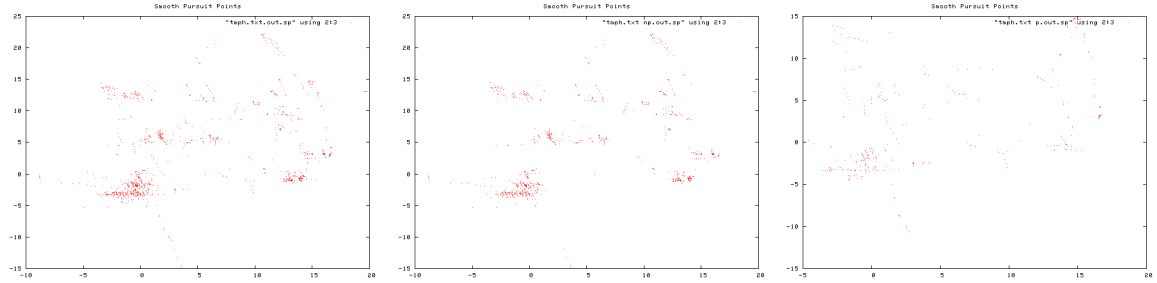



Figure 10.8: Smooth pursuit gaze locations, synthetic Trial 4. Entire trial (left), during periods of non-perturbation (middle), and during perturbation (right).

and both periods. It was expected that smooth pursuit velocities would be slightly higher during perturbation periods, corresponding to the tracking of moving objects. However, no significant variation in smooth pursuit velocities was evident during the human trials, due to error in the gaze estimation (although gaze may have been stable, error in gaze estimation means that recorded gaze oscillates around a location where fixation was actually stable, inducing smooth pursuit velocity where none existed). However, the data provided by the synthetic trials is more accurate, and a "thickening" of low velocities in the perturbation period histogram (bins 5-10) can be seen, in comparison to the non-perturbation plot. This corresponds to a small shift towards a higher average velocity of smooth pursuits during perturbation periods, as would be expected.

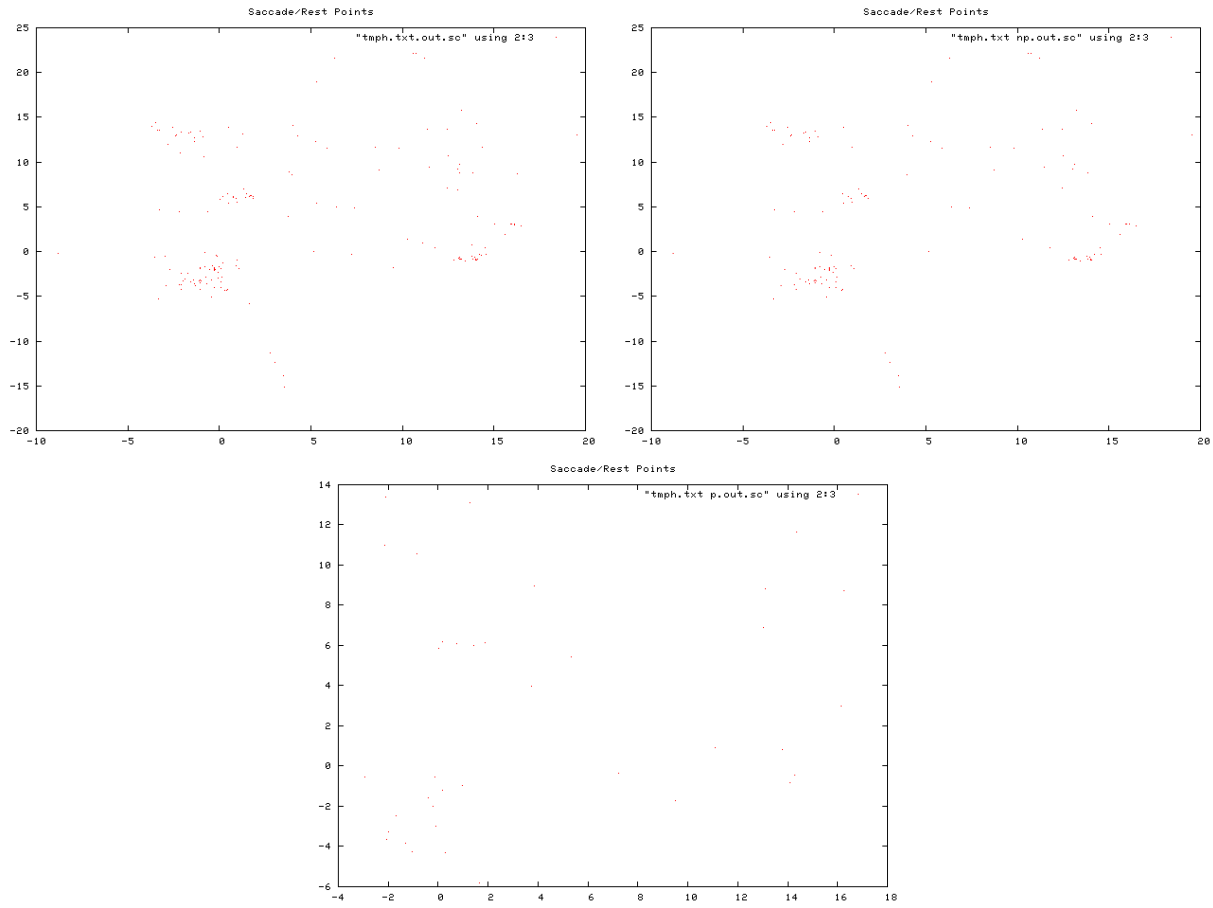



Figure 10.9: Saccade gaze locations, synthetic Trial 4. Entire trial (top left), during periods of non-perturbation (top right), and during perturbation (bottom).

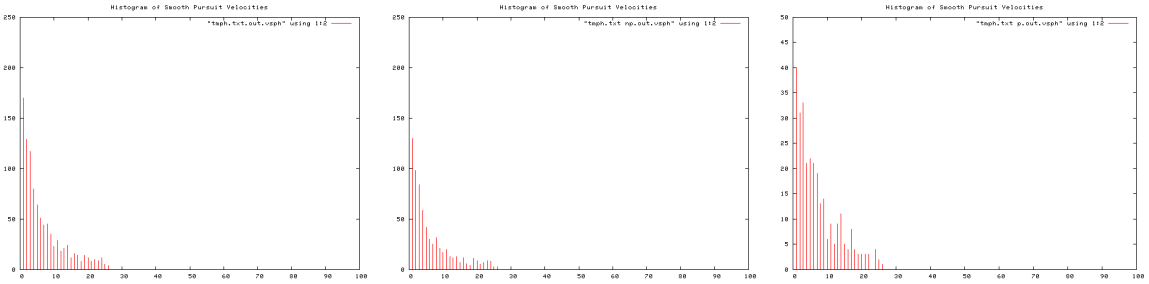



Figure 10.10: Histogram of smooth pursuit velocities, synthetic Trial 4. Entire trial (left), during periods of non-perturbation (middle), and during perturbation (right).

Figure 10.11 shows histograms of saccade velocities according to perturbed, non-perturbed and both periods. As for the human trials, it is evident the number of saccades during perturbation periods is significantly less than during non-perturbation periods, but that velocities are spread over the same values. The spread is quite narrow, induced by the axis motors reaching the same saccade ceiling velocity in the trapezoidal profile motion axis control.

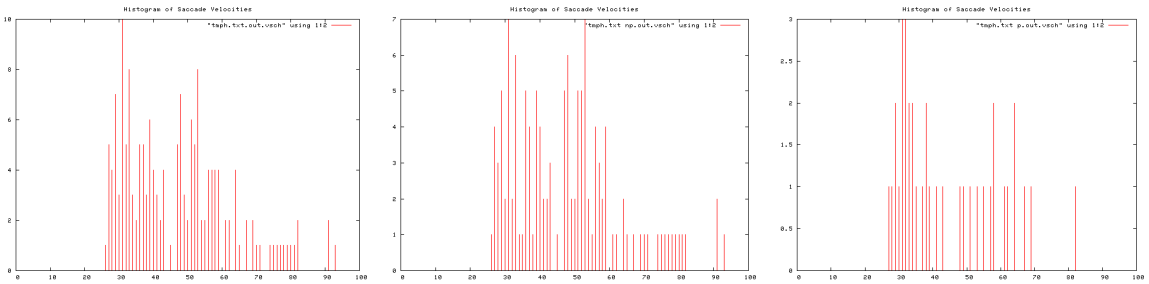



Figure 10.11: Histogram of saccade velocities, synthetic Trial 4. Entire trial (left), during periods of non-perturbation (middle), and during perturbation (right).

Figure 10.12 shows smooth pursuit durations; Figure 10.13 shows corresponding histograms. The histograms show a significant shift in smooth pursuit durations was present during periods of non-perturbation. This was seen in the human trials, and corresponds to preferential sustained tracking of perturbed object and shorter fixations upon objects during non-perturbation periods.

Next, smooth pursuit tracking distances were extracted (Figure 10.14) and associated histograms were created (Figure 10.15). Figure 10.15 shows a significant

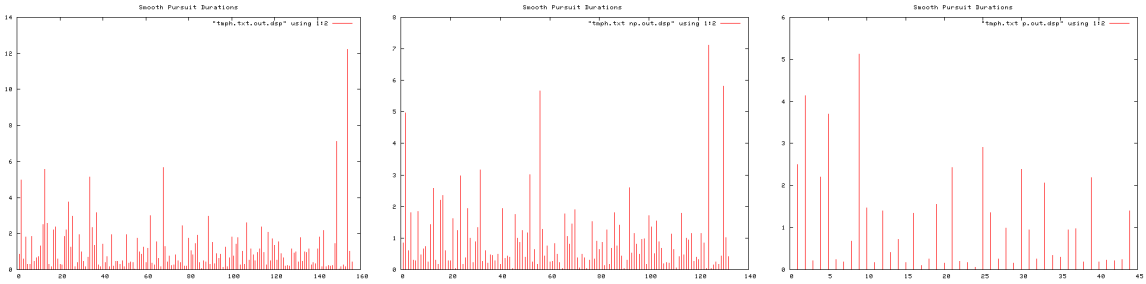



Figure 10.12: Smooth pursuit durations, synthetic Trial 4. Entire trial (left), during periods of non-perturbation (middle), and during perturbation (right).

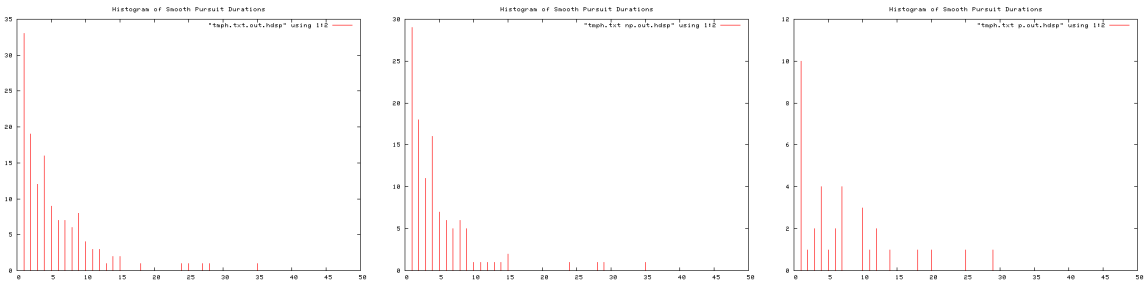



Figure 10.13: Histogram of smooth pursuit durations, synthetic Trial 4. Entire trial (left), during periods of non-perturbation (middle), and during perturbation (right).

shift towards shorter tracking distances occurs during periods of no perturbation. Again, this reflects what was observed in the human trials.

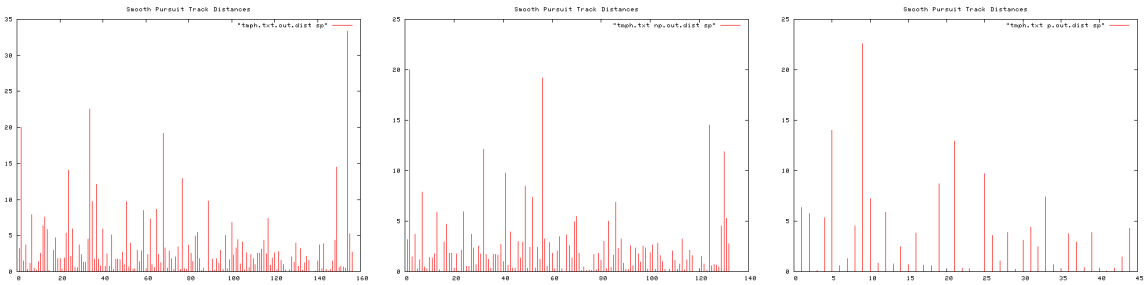



Figure 10.14: Smooth pursuit distances, synthetic Trial 4. Entire trial (left), during periods of non-perturbation (middle), and during perturbation (right).

As with the human trials, saccade durations are calculated as the number of

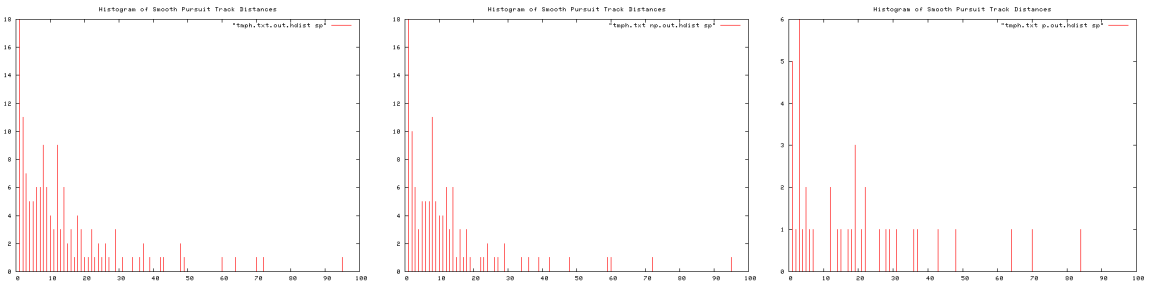



Figure 10.15: Histogram of smooth pursuit distances, synthetic Trial 4. Entire trial (left), during periods of non-perturbation (middle), and during perturbation (right).

consecutive samples above the saccade velocity, where velocity is calculated as the distance covered between samples divided by the time between samples. The sampling rate varied somewhat throughout the trial. Saccade durations are of the order of the arbitrarily varying sampling rate. The recorded saccade durations may therefore be coupled to the sample rate, and may not reflect the actual saccade durations. Saccade durations were therefore not used as behavioural metrics.

Saccade distances were extracted (Figure 10.16), and saccade distance histograms constructed (Figure 10.17). The appearance was as per the human trials, except for a second cluster of short saccade distances. After inspecting the video logs, it is evident the group of short distance saccades is induced during the MRF ZDF tracking of objects. The MRF ZDF server node segments the object upon which fixation exists, and centres gaze upon its centre of gravity. If the segmentation is not exact, or no segmentation is returned for a frame and the object is moving, the new location of the centre of gravity may be more than a few pixels from the previous location. If the distance is more than a few pixels, a rapid "catch-up" motion is initiated, and may reach velocities considered saccade.

This group of short distance saccades was not observed in the human trials. However, based on the human trials, we cannot conclude that humans do not make similar corrective motions during the tracking of objects. This is because the FaceLAB data is not resolute enough to detect such small motions and velocity fluctuations. It is likely that humans actually make less of these catch-up corrections than the synthetic system. Moreover, FaceLAB output data appears

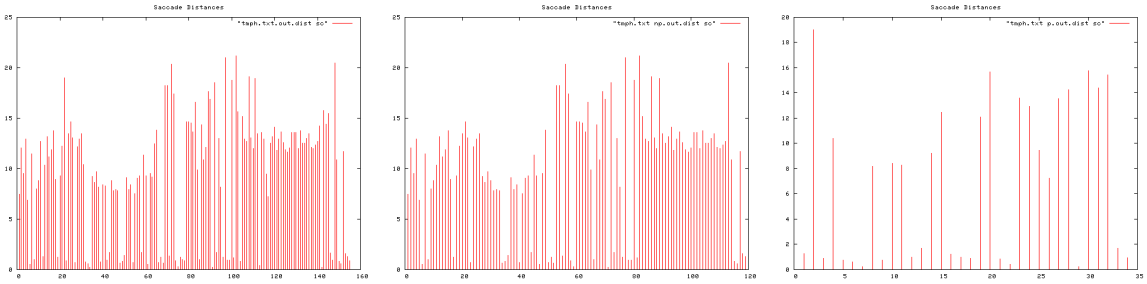



Figure 10.16: Saccade distances, synthetic Trial 4. Entire trial (left), during periods of non-perturbation (middle), and during perturbation (right).

to be filtered to remove such small fluctuations.

The reduced peripheral vision of the synthetic system in comparison to human vision means saccades in the synthetic trials were likely to be shorter distances than that of humans. It is noted that when the synthetic system gazed towards the top of the scene, the very bottom was sometimes out of view. Smaller distances for the synthetic system might also mean lower saccade durations in the synthetic system in comparison to humans.

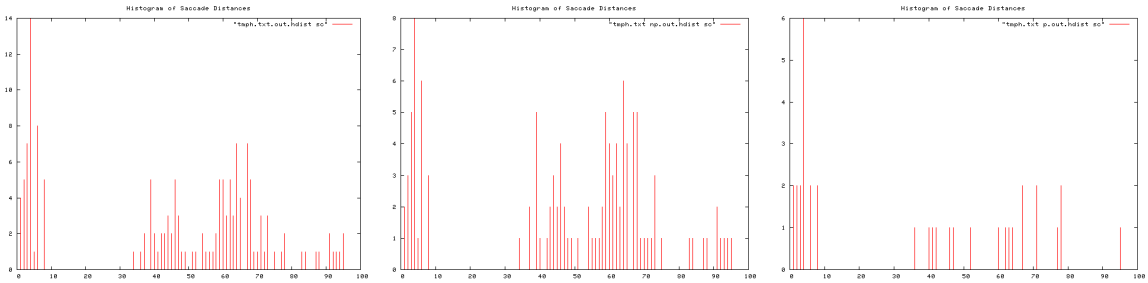



Figure 10.17: Histogram of saccade distances, synthetic Trial 4. Entire trial (left), during periods of non-perturbation (middle), and during perturbation (right).

Re-attention periods during non-perturbation were determined as per the human trials. Re-attention period data for synthetic trial 4 is shown in Figure 10.18. Please refer to Appendix B for all synthetic trial re-attention data.

| Ti | To | T | I/O Event | Pear | Apple | Peach | Orange |
|---|---|---|---|---|---|---|---|
| 21 | 38 | 17 | Orange In | | | | 3 |
| 45 | 0:55 | 10 | Pear In | 4 | | | 3 |
| 1:01 | 1:12 | 11 | | 2 | | | 3 |
| 1:18 | 1:37 | 9 | Peach In | 5 | | 1 | 6 |
| 1:42 | 1:53 | 11 | | 2 | | 2 | 3 |
| 1:57 | 2:08 | 11 | | 2 | | 4 | 5 |
| 2:11 | 2:26 | 15 | | 4 | | 3 | 4 |
| 2:30 | 2:40 | 10 | Apple In, Peach Out | 2 | 3 | 2 | 2 |
| 2:48 | 3:01 | 13 | Orange Out | 6 | 5 | | 0 |
| 3:02 | 3:17 | 15 | Pear Out | 5 | 4 | | |
| 3:25 | 3:33 | 8 | Apple Out | | 2 | | |
| | | | TOTAL Rets | 32 | 14 | 12 | 29 |
| | | | TOTAL T | 105 | 46 | 56 | 107 |
| | | | **Av. Re-attention Period** | **3.3** | **3.3** | **4.7** | **3.7** |
| | | | **SD(Av. Pr)** | **0.66** | | | |

Figure 10.18: Object re-attention during non-perturbation periods, synthetic Trial 4.

### 10.4.2 Numerical Characterisation

It is often possible to compare the performance of a system to a theoretical model by monitoring output and performing model-based residual analyses. However, human gaze behaviours are the product of an intricately complex biological system. As such, it is notoriously difficult to consider all the likely input influences and internal factors involved in determining human scanpaths. There is no general theory of human gaze behaviour that would permit such a systematic comparison.

Higher order relations are exceedingly difficult to capture. Hidden Markov models (HMMs), and various other Markov models have been attempted, but they have always been unsatisfactory. In attempts to show that a particular random number generator is "sufficiently" random, Don Knuth developed a large set of tests, and shows that "good" generators pass all the tests, regardless of their underlying generation method [Knuth (1997)]. Similarly, it is possible to compare the gaze behaviours of humans and machines by comparing the statistics and PDFs associated with specific parameters derived from gaze behaviour. In this regard, cluster overlap and KL divergence methods [Mitchell (1997)] to compare gaze parameters may not be appropriate due to small sample sizes in the human (20 samples) and synthetic (four non-independent samples) trials. We therefore select parameters that we may extract from both human and synthetic trials for comparison. The bootstrapped human statistics are used as a benchmark to which average parameters extracted from each synthetic trial are compared.

We examine the compliance of each of the average parameters extracted from each of the four synthetic trials to the statistics associated with the parameters obtained from the human trials. In so doing, we may establish which synthetic trial configuration elicited gaze behaviours that best resembled human behaviour in terms of extracted parameters. In this comparison we treat the trials individually, rather than as independent samples from the same underlying behavioural PDFs.

#### 10.4.2.1 Individual Trials

Ratio parameters, and the re-attention consistency parameter, were extracted from each of the synthetic trials individually. All absolute parameters extracted from all trials are provided in Appendix B. Table 10.1 summarises extracted

ratio parameters for each trial. For direct comparison, the four sample points associated with each of the synthetic trials have been plotted on the respective histograms of human parameter data (Figure 10.19 shows this for saccade rate parameter $Sr$, see Appendix B for all other parameters).

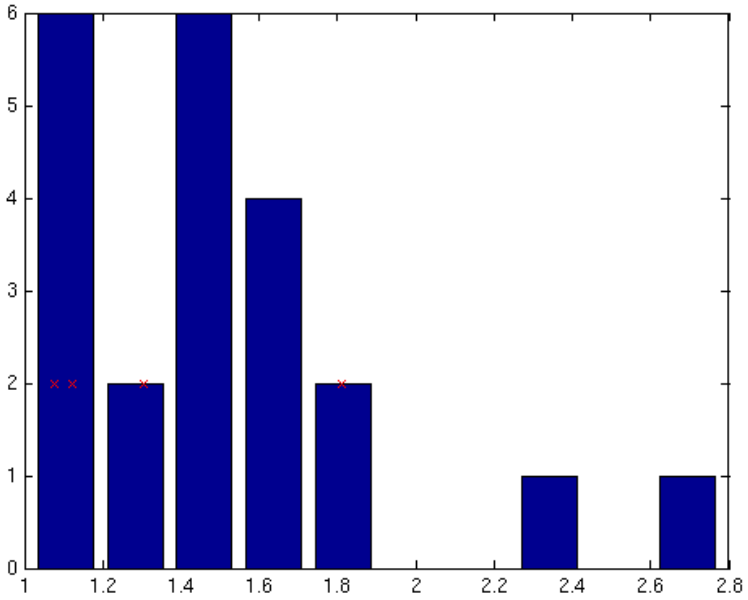



Figure 10.19: Histogram of saccade rate parameter $Sr$ from human benchmarks with synthetic trial samples superimposed (red crosses).

Table 10.1 shows that the trials exhibited parameter values that conformed to the expected trends set by the human trials (+/=/- column, Table 9.2). More specifically, we checked each value to determine if they fell within one bootstrapped standard deviation of the bootstrapped human means. First, we check if each parameter fell withing one bootstrapped standard deviation of the 95% mean confidence interval using the *lower* bound of the bootstrapped standard deviation 95% confidence interval (left, Table 10.2). We subsequently compared each parameter to see if it fell within one bootstrapped standard deviation of the 95% mean confidence interval using the *upper* bound of the bootstrapped standard deviation 95% confidence interval (right, Table 10.2).

The majority of extracted synthetic parameters fell within one maximal stan-

Table 10.1: Extracted average rate parameters for each trial.

| Parameter | T1 | T2 | T3 | T4 |
|---|---|---|---|---|
| $Spt_r$ | 1.069097 | 1.624368 | 1.541978 | 0.790567 |
| $Spl_r$ | 0.587497 | 0.743268 | 0.647662 | 0.613530 |
| $Scl_r$ | 0.988786 | 1.787190 | 1.877772 | 1.399369 |
| $Spv_r$ | 0.474087 | 0.492012 | 0.558833 | 0.756725 |
| $Scv_r$ | 0.903545 | 1.132790 | 1.030711 | 1.081753 |
| $Scp_r$ | 1.123602 | 1.076486 | 1.305637 | 1.809144 |
| $Pr$ | 0.505799697 | 0.602079729 | 0.469041576 | 0.660807587 |

Table 10.2: Comparing individual trial parameters with human benchmarks. Within one *minimal* bootstrapped standard deviation of 95% mean CI (left). Within one *maximal* bootstrapped standard deviation of 95% mean CI (right).

| Parameter (lb) | T1 | T2 | T3 | T4 | Parameter (ub) | T1 | T2 | T3 | T4 |
|---|---|---|---|---|---|---|---|---|---|
| $Spt_r$ | y | n | y | y | $Spt_r$ | y | y | y | y |
| $Spl_r$ | n | y | n | n | $Spl_r$ | y | y | y | y |
| $Scl_r$ | y | y | n | y | $Scl_r$ | y | y | y | y |
| $Spv_r$ | n | n | n | n | $Spv_r$ | n | n | n | n |
| $Scv_r$ | n | y | y | y | $Scv_r$ | y | y | y | y |
| $Scp_r$ | n | n | y | y | $Sc_r$ | y | n | y | y |
| $Pr$ | y | y | y | y | $Pr$ | y | y | y | y |
| Total votes: | 3 | 4 | 4 | 5 | Total votes: | 6 | 5 | 6 | 6 |

lb – lower bound, ub – upper bound.

dard deviation. One standard deviation is a rather tight range to test conformity; we should not expect all values to conform to this range. Therefore, the same comparison was conducted using a broader range of two standard deviations (Table 10.3).

The majority of parameters, and certainly those compared to two upper bound standard deviations, conform to benchmarks set from the human trials. The main discrepancy exists for parameter $Spv_r$, the ratio of smooth pursuit velocities in

Table 10.3: Comparing individual trial parameters with human benchmarks (2 SD). Within two *minimal* bootstrapped standard deviations of 95% mean CI (left). Within two *maximal* bootstrapped standard deviations of 95% mean CI (right).

| Parameter (lb) | T1 | T2 | T3 | T4 | Parameter (ub) | T1 | T2 | T3 | T4 |
|---|---|---|---|---|---|---|---|---|---|
| $Spt_r$ | y | y | y | y | $Spt_r$ | y | y | y | y |
| $Spl_r$ | n | y | n | n | $Spl_r$ | y | y | y | y |
| $Scl_r$ | y | y | n | y | $Scl_r$ | y | y | y | y |
| $Spv_r$ | n | n | n | n | $Spv_r$ | n | n | n | n |
| $Scv_r$ | n | y | y | y | $Scv_r$ | y | y | y | y |
| $Scp_r$ | y | y | y | y | $Sc_r$ | y | y | y | y |
| $Pr$ | y | y | y | y | $Pr$ | y | y | y | y |
| Total votes: | 4 | 6 | 4 | 5 | Total votes: | 6 | 6 | 6 | 6 |

lb – lower bound, ub – upper bound.

perturbed to non-perturbed periods. This value decreased ($< 1.0$) as per the human trends, but decreased even more than measured in human trials. This discrepancy is likely due to the low accuracy (low signal to noise ratio) involved in detecting low velocity motions with FaceLAB.

Conformity to human benchmarks was tallied for each trial. Trial 4 performed the best in terms of extracted averages conforming to human parameters in most instances.

#### 10.4.2.2 Group Parameters

The trials were iteratively configured to produce gaze behaviours that appeared more similar to human behaviours. Analysis of parameter conformity demonstrates that all trials exhibited reasonable resemblance to human trends. Moreover, the system was observed to produce human-like behaviours in all trials, regardless of a wide variance in configuration settings. This suggests the behaviours elicited are highly dependent on the implemented system model, and not just on the configuration settings selected for a particular trial. As a case in point, if considered as a set of four independent synthetic samples, we can

bootstrap group statistics for comparison to bootstrapped human group statistics. We find that the bootstrapped synthetic mean rates consistently change in the same direction as the bootstrapped human rates: where human rates tended to increase in going from periods of perturbation to periods of non-perturbation, so did the synthetic rates. Of course, the trials were *not* conducted completely independently with completely random configuration settings.

Despite the fact that the selection of configuration settings was not entirely random, for the purpose of comparison we assume the variance in configuration settings is significant enough to test the hypothesis that the synthetic behaviours are largely a product of the model, and not just configuration settings. It has been shown that the human parameters do not necessarily conform to normal distributions, motivating a bootstrap analysis to capture parameter statistics. As was done for the human rate parameters, we bootstrapped the same statistics for the synthetic trials, treating the four synthetic trials as four independent sample points (Table 10.4).

Table 10.4: Group statistics. Parameter changes when going from perturbation to non-perturbation periods.

| Parameter | $MeanCI_l$ | $MeanCI_u$ | $STDCI_l$ | $STDCI_u$ | $< +/- >$ |
|---|---|---|---|---|---|
| $Spt_r$ | 0.9298 | 1.5832 | 0.0412 | 0.4588 | = |
| $Spl_r$ | 0.5293 | 0.6518 | 0.0042 | 0.0917 | – |
| $Scl_r$ | 1.1941 | 1.8325 | 0.0453 | 0.4885 | + |
| $Spv_r$ | 0.4830 | 0.6905 | 0.0090 | 0.1582 | – |
| $Scv_r$ | 0.9609 | 1.1073 | 0.0255 | 0.1324 | = |
| $Scp_r$ | 1.0883 | 1.6378 | 0.0236 | 0.4099 | + |
| $Pr_{sd}$ | 0.4874 | 0.6461 | 0.0184 | 0.1107 | n/a |

Table 10.4 shows that when considering all trials as separate individuals contributing to group statistics, all parameters exhibited the same trends as observed in the benchmark human trials. Figure 10.20 shows the bootstrapped statistics for parameter $Scp_r$, the rate of increase in the proportion of saccade frames from periods of perturbation compared to periods of perturbation (left, synthetic; right, human). The synthetic sample parameter values are marked with blue crosses;

two 95% CI bootstrapped standard deviations (upper and lower bounds) are marked with red crosses, as per the explanation provided in Section 9.6.2, Figure 9.35. Normal theory statistics are shown in green, including markers of two standard deviation (green crosses). As per the human trials, the blue curve shows the density of synthetic means, the dashed red curve is provided for comparison to a Gaussian. The synthetic and human plots are shown side-by-side for comparison. Bootstrapped group statistic plots for all synthetic parameters are provided in Appendix B.

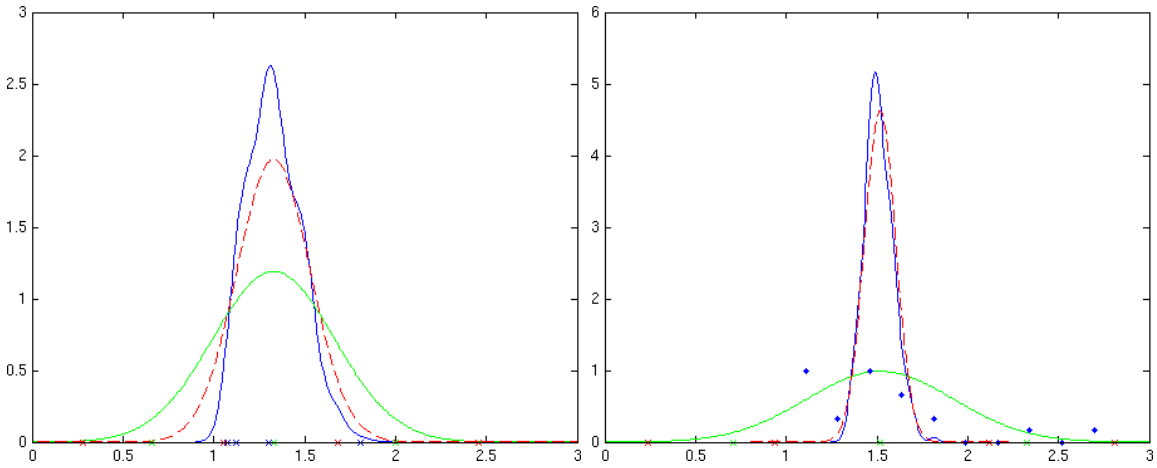



Figure 10.20: Bootstrapped parameter $Sr$. Synthetic system (left), and human benchmarks (right).

It is noted that there is considerable overlap between the bootstrapped human and synthetic group parameters. With input from only four synthetic trials and the poor assumption of trial independence, this is only a weak claim. Nonetheless, it suggests the synthetic gaze behaviours are largely a product of the system model, not just the configuration settings, and that the synthetic behaviour trends are similar to the benchmark trends compiled from the human trials.

### 10.4.3 Sensitivity

In the human trials the largest variation in extracted rate parameter ranges occurred in the saccade distance rate $Scl_r$ (upper bound on standard deviation 0.85), saccade proportion rate $Scp_r$ (0.57), saccade velocity rate $Scv_r$ (0.49), and smooth pursuit time $Spt_r$ (0.42). Other rate parameters exhibited low variance.

In the synthetic trials, the largest variation in extracted rate parameter ranges occurred in the saccade distance rate $Scl_r$ (0.49), the smooth pursuit time $Spt_r$ (0.46), saccade proportion rate $Scp_r$ (0.41). $Scv_r$ did not exhibit a variance as large as the human trials, but this is likely to be due to higher accuracy in velocity measurements using the synthetic system's encoders. Other than this instance, parameter sensitivity was similar for both the synthetic and human trials. Parameters that exhibited greatest variance ($Scl_r$, $Spt_r$, $Scp_r$) suggest that these are more sensitive to configuration setting changes. More synthetic trials with stronger independence (more randomly selected settings) would be required to confirm this hypothesis. The ability to infer which configuration settings cause most variation in extracted rate parameters would require numerous additional trial sets where only one configuration setting is varied over a wide range in each set of trials.

The re-attention coherence parameter was not significantly sensitive to parameter variations. Re-attention was consistently coherent in all trials. The variation in average re-attention period for each trial, however, was very sensitive to configuration variations, as expected. The average re-attention period varied from 3.75s to 6.43s across trials.

## 10.5 Discussion

As would be expected in selecting configuration settings to iterate towards more human-like behaviours, Trial 4 exhibited behavioural rate parameters that best conformed to the majority of human parameter distributions. Trial 4 involved configuration settings with a slower rate of accumulation of IOR, a larger radius of inhibition, and a slower IOR decay rate than other trials. It also involved slower decay of the fixation map (the product of IOR and saliency), longer inactivity timed shift period, and medium restrictiveness on selecting fixation map peak locations for attentional saccades, in comparison to the other synthetic trials.

The plots and initial processing of absolute parameters show that Trial 1 exhibited little discrimination between smooth pursuit and saccade velocities. Trial 2 exhibited good bimodal discrimination, but was rather saccadic. Trial 3 was less saccadic than Trial 2, but exhibited less smooth pursuit velocities than did the human benchmarks. Trial 4 exhibited a better balance of saccade to

smooth pursuit velocities and a good discrimination between the two modes, as well as more smooth pursuit velocities, and was slightly less saccadic than the previous trials.

When the weak assumptions made earlier are assumed valid and the trials are considered as independent, the group trends and range variances cohered to the human trial statistical benchmarks. This suggests the system elicits human-like gaze behaviours (somewhat) regardless of selected configuration settings. Indeed, footage of synthetic behaviour produced human-like reactions to visual stimulus regardless of settings. The model also shows flexibility in that it can be tuned to better replicate specific human behaviours.

Consideration of parameter sensitivity shows that changing synthetic configuration settings elicited variances in output parameter ranges similar to the ranges observed across human participants. Parameters predominantly dependent on hardware (muscles/actuators) showed low variance across trials, and parameters more dependent on the underlying system model showed greater variance, for both human and synthetic trials.

The trials were not tailored to determine the correct object re-attention period, IOR radius, or decay rates, or tracking periods (for example). These parameters are likely to differ greatly across individuals, and even over time for a particular individual. The trials serve to determine what particular *combinations* of synthetic configuration settings result in human-like behaviours. Even though the system components take biological inspiration, the trials do not provide information about the structural similarity of the system to the primate visual brain. The synthetic system incorporates various engineering components, but the trials were not tailored to test the functionality of individual components for conformity to components of the primate visual system. They may only be used to comment on emergent attentional gaze behaviours observed in the synthetic trials for comparison with benchmarks obtained from the human trials.

The trials allow comparison of the characteristics of saccade and smooth pursuit between the synthetic system and human vision system. Aspects of other human gaze motions, such as the vestibulo-occular reflex, microsaccades, and the optokinetic reflex, despite being observable gaze-affecting phenomenon, were not considered in the trials, largely due to the inability to detect such small, rapid motions using unobtrusive methods such as FaceLAB.

Re-attention periods were slightly less coherent in the synthetic trials (average standard deviation was 0.56 for synthetic trials, 0.43 for humans), but standard deviations remained consistently low nonetheless. The standard deviation of the re-attention period within a trial was far less than standard deviation of all objects over all trials. The standard deviation across all objects in all trials was 1.19 for the synthetic trials, 1.92 for humans. This figure, and the magnitude of the increase over the average coherence for individual trials shows good correspondence. The small numerical discrepancies are likely to be due in part to the low number of synthetic trials. Overall the same trends were observed as for the human trials, demonstrating that the re-attention period varied across all synthetic trials, but that there was coherence in the re-attention period for each individual trial.

Horowitz and Wolfe proposed that visual search is memoryless [Horowitz & Wolfe (1998)] - when elements of a search array randomly re-organised while subjects searched for a specific target, search efficiency was not degraded. Performance gains for searches on a stable array would indicate memory use. However, this may just preclude perfect memorisation and does not necessarily preclude the possibility that the last few attended locations are remembered, in accordance with the limited lifespan of IOR. Other psycho-physical experimentation with static stimulus [Irwin & Zelinsky (2002)] suggested that a short-term attentional memory maintains information about salient visual features and their locations ("object files") across saccades, and that up to three or four object files may be retained.

When four objects were present and pseudo-static in our scene, they were cyclically attended by the system at an approximately even rate, and in the same cyclical order. The re-attention behaviour elicited by our synthetic system is therefore consistent with both of the above psycho-physical observations.

## 10.6 Summary

Trials identical in character to the human trials of the previous chapter were conducted to record the gaze behaviours elicited by a primate-inspired synthetic vision system. Behavioural metrics were extracted for comparison with those established during the human trials.

With the exception of the smooth pursuit velocity rate parameter, all rate parameters extracted from individual synthetic trials fell within two bootstrapped standard deviations of the bootstrapped density of means of the benchmark human trials (upper bound 95% confidence intervals). The smooth pursuit velocity parameter discrepancy is likely due to the low accuracy (low signal-to-noise ratio) involved in detecting low velocity gaze motions with FaceLAB. This demonstrates that all trials, in terms of the extracted rate parameters, fitted the norms of primate gaze behavior. In particular, the configuration settings associated with Trial 4 elicited behaviours that best matched human performance, in terms of the extracted rate parameters.

The fact that all trials, all with different configuration settings, exhibited a majority of rate parameters that fell within the bootstrapped standard deviations of human benchmark parameters suggests similar performance does not rely upon the selection of configuration settings. Rather, gaze behaviours of the synthetic system are largely a product of the underlying model. Though the assumption that all trials may be treated as individual sample points is weak, when treated as such, the group statistics thus formed also conform well to the human benchmarks. The density of bootstrapped synthetic means matched human benchmark means, for all parameters. That is, where human parameters produced mean densities above, below, or centred at 1.0, so too were the densities of synthetic means, for all parameters. The fact that only four synthetic trial samples were used for this trend comparison further weakens this finding. Furthermore, the ranges of bootstrapped standard deviations for rate parameters were similar for the human trials and the synthetic trials, though this is also a weak claim as standard deviation estimations are based upon only four samples. In any case, the grouped bootstrap analysis is supportive, but not necessary because the strong conformity of individual synthetic trials to the benchmark human data is sufficient to prove good consistency between human and synthetic gaze behaviours.

The strong conformity of individual synthetic trials to the human benchmarks indicates that, in terms of these trials, the primate-inspired synthetic system elicits primate-like gaze behaviours.

# Chapter 11

# Conclusion

## 11.1 Summary

An investigation of machine vision for seeing in the real world has been conducted. Based upon the real-world success of biological vision, we have considered components of the primate vision system. We have reviewed components of primate vision useful for making a synthetic primate vision system. Primates benefit from active vision in several ways. It enables continual foveal alignment of objects in the scene. It permits correction of retinal shifts induced by head perturbations within reflexive, rather than cognitive, timespans. It permits coordinated fixation and smooth pursuit of targets such that target motion blur is reduced. Active foveal perception and attention allows data reduction and high equivalent resolutions in observing a scene. An egocentric spatial perception provides primates with an awareness of the location of visual surfaces in a scene, and their motion. We have also considered existing models of primate vision and justified components using biological inspiration. We have developed a real-time synthetic active vision system that incorporates such components. The system has been implemented on a real-time processing network based around a biologically-inspired active vision mechanism able to perform the behavioural eye movements of primates.

By specifying system properties similar to those of primate vision, we have developed a synthetic active visual system capable of detecting and reacting to unique and dynamic visual stimuli, and of being tailored to perform basic vi-

sual tasks. The specific processing algorithms may not (and probably do not) reflect what actually happens in the primate brain. Active rectification provides egocentric spatiotemporal visual perception. We have presented a method to augment active vision disparity data into an egocentric, unified, space-variant occupancy grid representation. The occupancy grid has been explicitly designed for integrating data from active vision, and for providing low-bandwidth and useful representations useful for perception in real-time. We have shown how the occupancy grid can be used to extract information about the scene such as 3D motion and 3D cue-surface correspondences. A foveal MRF ZDF algorithm permits attended object tracking and extraction, and ensures coordinated stereo fixation upon visual surfaces. Attention with active-dynamic IOR means that a short term memory of previously attended locations can be retained. Spatial and cue biassing facilitates top-down modulation of attention towards regions and cues relevant to tasks. Covert consideration of potential saccade destinations (before overt attention is deployed) provides attentional moderation. These features result in a reactive vision system and the emergence of primate-like attentional behaviours.

The synthetic vision system preferentially directs its attention towards non-suppressed salient objects/regions. Upon saccading to a new target, the MRF ZDF algorithm extracts the object that has won attention, maintaining stereo fixation on that object (smooth pursuit), regardless of its appearance or motion. Attention is maintained until a more salient scene region is detected, or until IOR allows alternate locations to win fixation.

We adopt a client-server architecture to allow concurrent serial and parallel processing. At the lowest level, a rectification server distributes rectified images and rectification parameters to dependent nodes. U and V colour chrominance images for both the left and right images are sent to the colour centre-surround ($CCS$) server for processing. Y channels are sent to the orientation ($OCS_L, OCS_R$) servers and the depth and flow ($DFCS$) server. To minimise network bandwidth, to cope with the processing load of each frame, and to prevent repetition of computations, nodes in the structure are configured simultaneously as clients of processes preceding them in cue serialisation, and as servers to nodes following them. Each node is a dual CPU hyper-threaded 3GHz PC with four virtual processors. Trade-offs exist between splitting tasks into sub-tasks, passing

sub-tasks to additional nodes, and minimising network traffic. The best performing solution involves grouping serialised tasks on each server, and that as many operations are done on the image data on the same server as possible, so that there is minimal CPU idle time and minimal network traffic. The serial nature of cue computations means that there is often no gain possible in distributing the task – in fact further network transfer of data between servers would slow performance. Implementation of the system provides insight into what capabilities may be achieved on a synthetic processing network. The low-latency real-time performance of the system indicates the feasibility of such a system. This performance, and the flexible nature of the processing network permits extensive system expansion for future additional processing tasks.

Psycho-physical experiments were conducted with humans to benchmark inter-individual trends in human gaze behaviours for evaluation of the primate-like performance of the synthetic vision system. While viewing a dynamic, repeatable, controlled 3D scene, participants' unconstrained gaze scanpaths were recorded. Parameters useful in characterising human gaze behaviours were selected based upon two pilot trials. Group statistics associated with each selected parameter were then extracted from 20 subsequent human trials. All participants were observed to react to the stimulus in a quantifiably similar manner. All participant's distance-weighted velocity magnitude histograms were distinctly bimodal, exhibiting a group of low velocities (corresponding to smooth pursuit) and a group of high velocities (corresponding to saccades) separated by a range of sparsely occupied medial velocities.

Analysis of extracted inter-individual behavioural rate parameters showed characteristic trends. In general, the ratio of saccades from periods of perturbation to periods of non-perturbation consistantly increased. The smooth pursuit distance and velocity ratios from periods of perturbation to periods of non-perturbation both consistently decreased, commensurate with the tendency for individuals to track translating stimuli. No significant change in saccade durations and velocities were detected across periods of perturbation and non-perturbation, suggesting that this parameter is not significantly dependent on the scene. The saccade length tended to increase from periods of perturbation to periods of non-perturbation. The average re-attention period for each individual varied significantly. However, object re-attention periods were approximately

constant during periods where no object was being actively perturbed, for each individual.

With the exception of the smooth pursuit velocity rate parameter, all rate parameters extracted from individual synthetic trials fell within two bootstrapped standard deviations of the bootstrapped density of means of the benchmark human trials (upper bound 95% confidence intervals). This demonstrates that all trials, in terms of the extracted rate parameters, fitted the norms of primate gaze behaviour. In particular, the configuration settings associated with Trial 4 elicited behaviours that best matched human performance, in terms of the extracted rate parameters.

The fact that all trials, all with different configuration settings, exhibited a majority of rate parameters that fell within the bootstrapped standard deviations of human benchmark parameters suggests similar performance does not rely upon the selection of configuration settings. Rather, gaze behaviours of the synthetic system are largely a product of the underlying model. Though the assumption that all trials may be treated as individual sample points is weak, and though it is not necessary to conduct a bootstrapping analysis of the four synthetic trials as a group, when treated as such, the group statistics thus formed also conform well to the human benchmarks. The density of bootstrapped synthetic means matched human benchmark means, for all parameters. The group bootstrap analysis supports the strong conformity of individual synthetic trials to the benchmark human data.

Moreover, the strong conformity of individual synthetic trials to the bootstrapped human behavioural benchmarks indicates that, in terms of the trial scenario, the primate-inspired synthetic system elicits primate-like gaze behaviours.

## 11.2 Outlook

The performance of the real-time synthetic primate vision system has been demonstrated. Further human and synthetic trials may however reveal additional interesting similarities and observations.

Complex systems can usually benefit from refinement, and this system is no exception. Some specific improvements have been suggested in the course of

presenting the system that may improve performance, or permit it to conform further with the primate vision system. For example, the active rectification step may achieve better accuracy in rectification and mosaicing if image-based techniques (such as SIFT) are incorporated to improve the estimated geometry for each pair of images. Active rectification may also benefit from projecting camera images into an alternate static reference frame, such as a sphere, instead of the planar mosaics used in the current implementation.

Improved occupancy grid resolution may be obtained by incorporating higher resolution disparity estimation. An expansion of the phase-based orientation analysis in the orientation processing server ought to provide additional cues such as symmetry and corner and edge detection, as well as an alternate method to estimate stereo disparity. This approach may in fact reduce required processing resources. The present system, and any additional processing requirements from the incorporation of additional features, could also benefit from hardware (DSP/FPGA) algorithmic implementations.

An obvious next step related to the output of the MRF ZDF algorithm is the classification and autonomous cataloging of attended objects. In determining IOR, it is likely that using the output mask from the MRF ZDF algorithm, instead of a Gaussian kernel, would yield more primate-like object-based suppression of attended objects. The attention system, and top-down search, would also benefit from knowledge of scene gist. Further, the system is presently ready for experimentation with top-down modulation of attention for target search and other tasks including object manipulation and HCI.

The application domain for the synthetic vision system is broad and exciting. Applications include autonomous robots, security and novelty detection, prosthetic vision, and human assistance systems. Primate-like machine vision will undoubtedly play an increasingly important role in the technologies of the future.

# Appendix A

# Human Trials: Ethics

## A.1 Ethics Application

# A. HUMAN TRIALS: ETHICS

***ALL APPLICATIONS TO BE TYPED*** *Version current from 1 February 2004*

**THE AUSTRALIAN NATIONAL UNIVERSITY**

**HUMAN RESEARCH ETHICS COMMITTEE**

**APPLICATION FORM**


**Surname of Researcher: Dankers**
**First name/s: Andrew Alexander**
**Title (e.g. Ms., Mr., Dr. etc,): Mr**

**Position Held (staff, postgraduate, undergraduate, etc.): Postgraduate, PhD Candidate**

**Student or Staff ID no. (if applicable): U3063322**

**Dept/School/Centre: Research School of Information Sciences and Engineering, Dept. Information Engineering.**

**Mailing address: B345, lvl 3, RSISE (Bldg 115), ANU**

**Telephone: x58685**
**Fax: x58660**
**Email: andrew.dankers@anu.edu.au**


**PROJECT TITLE: Comparision of Synthetic Stereo Active Vision Attention System with Human Attention**

**Date of this application: Nov 27, 2006**
**Anticipated start date for project: Jan 8, 2007 Anticipated end date: Feb 28, 2007**

1. ***The researcher/s***
   **Who are the investigators (including assistants) who will conduct the research and what are their qualification and experience? Please include their Department/School/Centre (or external institution for external researchers). Students should not include supervisors at this point unless they are actually participating in the research project as partner researchers.**

Andrew Dankers BSc(phys) BE, ANU: Mr Dankers is an ANU/NICTA postgraduate scholar at the RSISE, Department of Information Engineering. Mr Dankers' research focus involves the development of a biologically inspired synthetic visual attention system, based upon the characteristics of primate attention.

2. ***Understanding the national guidelines, the "National Statement on Ethical Conduct in Research Involving Humans" (1999)***
   **Can the proposer certify that the persons listed in the answer to Question 1 above have**

**been fully briefed on appropriate procedures and in particular that they have read and are familiar with the national guidelines issued by the National Health and Medical Research Council (the *National Statement on Ethical Conduct in Research Involving Humans*) (cited below as the "National Statement")? If there are guidelines from any relevant professional body with which the researcher/s are familiar they should also be listed below.**

Andrew Dankers has read, and is familiar with, the practical implications of the NHMRC National Statement on Ethical Conduct in Research Involving Humans. Although it is not anticipated that assistance from additional persons will be required, any additional persons will be required to read the NHMRC National Statement on Ethical Conduct in Research Involving Humans.

**3.** ***Purpose and design of the proposed research***

***Purpose***
**(a) Briefly describe the basic purposes of the research proposed** *(in plain language intelligible to a non-specialist).*

***Background***
A synthetic active vision system has been proposed and implemented as the focus of PhD research. The vision system is able to detect and direct its gaze towards salient visual events occurring in real scenes in real time. The system has been designed in light of observations of the primate vision system. It retains a short term memory of attended regions, such that they are not immediately re-attended, and can be biased for basic visual tasks.

The final contribution of the work will be to qualitatively compare the performance of the synthetic system to that of the primate vision system. We aim to compare timing statistics associated with the synthetic system observing a scene to that of human subjects observing the same scene.

***Desired Human Trial Outcomes***
We aim to observe humans observing a controlled dynamic 3D scene such that attentional statistics may be obtained. Such statistics include:

a) Saccade frequency - statistics associated with the time between eye gaze shifts.
b) Non-return inhibition period – once a human has attended an object, they will not attend that region again for some time. Statistics associated with this behavior are desired.
c) Time spent attending scene locations – we hope that the synthetic system spends similar periods of time fixated upon the visual regions that humans do. The order of attending locations is not important. The total time spent at each location is more relevant.

Observing these human parameters should help us tune the synthetic system to best replicate timing characteristics of human attention. This will help researchers set parameters in the synthetic system such as inhibition of return rates, inhibition response field radii, and assess whether synthetic systems behave similarly to primates upon which they are modelled. We are not interested in determining what cues were likely to have invited fixation, because we have only defined a reduced set of basic cues whereas humans have many cues that contribute to the perception of saliency. We are only interested in statistical timing parameters such as a), b) and c) above.

***Design***
**(b) Outline the design of the project** *(in plain language intelligible to a non-specialist).* **(If interviewing people or administering a survey/questionnaire, please attach either a list of the broad questions you propose to ask, or a copy of the questionnaire.)**

The trials will be conducted as follows:
1) Participants will be asked to sign a consent form (Attachment B) and consider completing a brief questionnaire (Attachment D) and visual acuity test (Attachment E) designed to establish the reliability of the trial to be conducted.
2) FaceLab, a commercial product from SeeingMachines, will be calibrated for each candidate and used to log the participant's eye gaze direction. In this manner, we can record the candidates' gaze scanpaths. Facelab is a passive system that uses static cameras to record eye gaze.

3) Participants will be asked to sit in front of a curtain with a 100x100cm window cut into it, such that they may observe the controlled, bounded 3D scene beyond.
4)We will ask the participant to undertake a simple search task to reduce the effect of emotional/expectational noise on visual attention. An example visual task would be"find a cherry". This may, for example, make the participant more responsive to small, red objects. We choose simple tasks that we can also bias the synthetic system towards. We will then introduce random non-iconic dynamic stimulus into the visual workspace. We may or may not enable the candidate to successfully perform the task (for example, we may not actually show a cherry) during a trial.
5) 5 different but similar trials will be logged by FaceLab and simultaneously recorded by video camera, per participant.
6) Statistics of attentional timing parameters will be determined offline.

**Considerations**:
1) Repeatability of trials is highly desirable, but it is not crucial that they are all identical. We will endeavor to ensure a high degree of similarity (in stimulus locations and appearance timings) across participants.
2) Stimulus will be non-iconic and may include simple objects like moving coloured balls and Leggo pieces. We avoid stimulus that may elicit emotional responses such as a doll – the apparent mood on the face of a doll, for example, may affect the participant's attention.

**4.** ***Sources of data involving humans***
**To ensure compliance with privacy legislation the committee needs to know your sources of information, i.e. where you are obtaining data involving humans. If you are using individual participants, tick at (a). If you are accessing personal records held by government departments or agencies, or by other bodies, e.g. private sector organisations, please tick and complete the relevant sections (b), (c) and/or (d) below.**

| | |
|---|---|
| (a) Individual subjects | ( √ ) |
| (b) Commonwealth Department/s or agency *(specify)** | ( )……. |
| (c) State/Territory Department/s or agency *(specify** | ( )……. |
| (d) Other sources (*specify*) | ( )…..…… |

*Please include an estimate of how many records you expect to access:…….

**5.** ***Personal identifiable data for medical/health research***
**Are you obtaining personal identifiable data specifically for medical/health research that is held by a government or private sector agency? (The committee needs this information to determine whether it needs to comply with relevant National Health and Medical Research Council guidelines relating to privacy legislation.)**

**NO**

**6.** ***Recruitment***
**Describe how participants will be recruited for this project. Indicate how many participants are likely to be involved, how initial contact will be made, and how participants will be invited to take part in this project.** A copy of any relevant correspondence should be attached to this application. **Does the recruitment process raise any privacy issues, e.g. does the researcher plan to access personal information to identify potential participants without their knowledge or consent? Describe the steps to be taken to ensure that participation or refusal to participate will not impair any existing relationship between participants and researcher or institution involved.**

We aim to conduct human trials using 20-30 subjects. Subjects will be between ages 18 to 50 years. An email advertisement (Attachment A) will be sent out to research peers and other personnel at the ANU, for consideration. Interested persons who respond to the advertisement will be sent a further information sheet (Attachment C) describing in more detail what to expect and how to get involved. Upon scheduling a trial, participants will be sent a consent form (Attachment B) that they should read, complete, and bring to the trial. The consent form will ask participants not to disclose the exact nature of the trials to other prospective participants who have not yet participated, such that expectational biases are not induced in prospective participants. Upon attending a trial, participants will first be asked to answer a brief questionnaire about their vision (Attachment D), and conduct a brief visual acuity test (Attachment E) – both of which they may choose to decline and still be permitted to participate. Participants are not required to have their name associated with the trial data. Participants may withdraw from the trials at any stage during the trial, and may request all data associated with the trial be destroyed.

**7.** ***Arrangements for access to identifiable data held by another party***
**In cases where participants are identified from information held by another party (e.g. government department, non-governmental organisation, private company, community association, doctor, hospital) describe the arrangement whereby you will gain access to this information. Attach any relevant correspondence.**

N/A

**8.** ***Vulnerable participants***
**Will participants include students, children, the mentally ill or others in a dependent relationship? If so, provide details.**

The study will not include students, children, the mentally ill or others in a dependent relationship.

**9.** ***Payment***
**Will payment be made to any participants? If so, give details of arrangements.**

Payment will not be provided to participants.

**10.** ***Consent***
**Describe the consent issues involved in this proposal (***see the National Statement, in particular Section 1.7-12, and other sections relevant to your research***). Describe the procedures to be followed in obtaining the informed consent of participants and/or of others responsible. Attach any relevant documents such as a consent form, information sheet, letter of invitation etc. If you do not propose to obtain written consent (e.g. if working with non-literate people) give a detailed explanation of the reasons for seeking oral consent, describe the procedure you intend to adopt, and specify the information to be provided to participants. If you have answered YES to Question 8 above please address any issues of consent and the possibility of coercion.**

Consent will be sought from participants prior to the trials to a) ensure that the participant fully understands what they will be required to do in the assessment, b) to ensure that the participant understands what the information they may provide will be used for, and c) to ensure that their agreement to participate is based on an accurate interpretation of participant requirements.

An email advertisement (Attachment A) will be sent to prospective participants.

Prior to participation, an information sheet (Attachment C) and a consent form (Attachment B) will be provided to potential participants. The consent form will be returned to the researchers prior to trial commensement. Immediately prior to the commencement of trials, the purpose of the study will be again outlined, and the participants will be reminded that they may withdraw from the study at any time without prejudice, and that all data relating to their participation can be destroyed, at their request. Participants are reassured that there are no correct or incorrect reactions to each trial.

At the beginning of the trial, participants will be given ample opportunity to ask any questions they may have about the study. The researchers will again explain the purpose and nature of the study,

and check that the participant understands the consent form and has completed it correctly. Participants will then be guided through a brief visual health questionnaire (Attachment D), and a brief visual acuity eye chart test (Attachment E). This is designed to obtain suitability of data associated with each participant. Participants are to liberty to decline participation, or further participation, in the guided questionnaire and/or visual acuity test.

**11.** ***Protection of privacy (confidentiality)***
**Describe the confidentiality issues involving in this proposal. Give details of the measures that will be adopted to protect confidential information about participants, both in handling and storing raw research data and in any publications. Blanket guarantees of confidentiality are not helpful. If the term "confidential" is used in information provided to participants, a full description of what precisely confidentiality means in the context of this research should be given. You should be aware that, under Australian law, any data you collect can potentially be subpoenaed. Depending on the nature of your research, it may be helpful to qualify promises of confidentiality with terms such as "as far as possible" or "as far as the law allows".** *[See the National Statement, in particular Sections 1.19, 18 and Appendix II]*

On presenting to the testing session, participants will be assigned a participant number, and all data associated with a participant will subsequently be identified by that number alone. Participants may choose to associate their name and contact details with a data number such that they may view their raw results after participation, or such that they may be contacted to participate in extra trials, if they so choose. Such names and/or contact details will be removed from the data set and destroyed within one month of the completion date of the trials, or beforehand at the request of the participant.

The investigators will ensure that data obtained during the trials:
- will not be released to anyone outside the study;
- will not be used for purposes other than those given in the consent form; and,
- will be published in statistical summary form such that individuals are not be identifiable.

Raw data, including questionnaire responses, will be kept confidential as far as the law allows. Raw data will be stored on a password protected computer only as long as is required for statistical analysis. Thereafter, only statistical summaries will be kept, and the raw data will be destroyed. Raw data will not be kept for a period longer than one year. Raw data will not be published.

**12.** ***Cultural or social considerations***
**Comment on any cultural or social considerations that may affect the design of the research.** *[See the National Statement, in particular Sections 1.2 and 1.19]***.**

This research does not address issues of cultural difference. However, should such issues arise, they will be considered in an ethical and sensitive way.

**13.** ***How the research might impact on participants***
**Describe and discuss any possible impact of the proposed research on the participants or their communities that you can foresee. This might include psychological, health, social, economic or political changes or ramifications. Discuss how you will try to miminise any impact.** *[See the National Statement, in particular Sections 1.3 to 1.6 and Section 1.14]*

It is expected that most participants will enjoy the experience of participating. However, it is possible that some individuals may find the laboratory based testing unpleasant or tedious. The researchers will minimise this by providing refreshment breaks as appropriate. Participants will be reminded during the assessment that they are free to leave at any time. Participants are free to ask any questions at any stage during the trials, and to seek satisfactory answers.

**14.** ***Other ethical and any legal considerations***
**Comment on any other ethical considerations that are involved in this proposal, including any potential for legal difficulties to arise for participants.**

N/A

**15.** ***Benefits versus risks***
**Describe the possible benefit/s to be gained from the proposed research. Explain why these benefits outweigh or justify any possible discomforts and risks to participants. In framing your explanation make explicit reference to the ethical considerations mentioned in your answers to previous questions on this form.** *[See the National Statement, in particular Sections 1.3-6 and 1.13-14]*

The knowledge gained from this study will be used to assess the validity of the synthetic visual attention architecture as a model of primate visual attention. It will help to assess the performance of other synthetic visual systems, in terms of statistical similarity to primate attention.

A conceivable risk to participants is that they may become anxious during the trial. We will endeavour to keep participants as relaxed as possible. The researches will adhere to the applicable OH&S guidelines associated with the care of persons present on ANU campus. It is not foreseen that precautions additional to those considered by the applicable ANU OH&S guidelines are necessary.

We see few risks associated with the trials and are confident that the associated benefits far outweighs any risk.

**16.** ***Handling possible problems arising from the research***
**Describe the arrangements you have made to handle concerns and complaints by participants, or emergencies involving participants or researchers.**

The information sheet will include the names and phone numbers of researchers to contact about any research issues and of the Secretary of the Human Research Ethics Committee to contact concerning ethical complaints.

**17. RESEARCH PROTOCOL CHECKLIST**

There are some key ethical principles that need to be addressed in your protocol (as an ethics application is known). In particular the committee needs to see how you have addressed the issue of informed consent and the issue of confidentiality, i.e. how the identities of participants will be protected in the raw research data and in published material. The usual way to obtain informed consent is in writing, by use of a consent form that is signed by the participant and retained by you. Because you retain the consent form the same information needs to be included in an information sheet that participants retain. Both the consent form and the information sheet should include your name, contact details, title and brief description of the project, details on how the identities of participants will be protected (both when storing the raw research data and in its published form), a statement that participation is voluntary and participants can withdraw at any time, and contact details for the Human Research Ethics Committee in case of any ethical concerns. If you do not propose to seek written consent, you need to explain why oral consent will be sufficient and how you propose to obtain it.

**Please tick the relevant boxes below to indicate what has been included in your protocol:**

| | | |
|---|---|---|
| Outline of proposal and purpose | Yes ☐ | No ☐ |
| Measures to be taken to protect confidentiality | Yes ☐ | No ☐ |
| Explanation of how written informed consent will be obtained | Yes ☐ | No ☐ |

If written consent is not being sought, justification of a verbal consent procedure is included Yes ☐

| | | |
|---|---|---|
| Full details on investigators (name, institution, etc.) | Yes ☐ | No ☐ |
| All researchers on this project are familiar with the national guidelines (*National Statement*) | | |

Yes ☐ No ☐

Details re how participants will be recruited Yes ☐ No ☐

Is personal data from a Commonwealth department/agency or private sector organisation being used?

Yes ☐ No ☐

Details on how cultural and social sensitivities will be addressed Yes ☐ No ☐

Consideration of likely risk to participants (e.g. psychological stress; cultural, social, political or economic ramifications) Yes ☐ No ☐

Do your research participants include:

| | | |
|---|---|---|
| Aboriginal or Torres Strait Islander peoples | Yes ☐ | No ☐ |
| Children and young people (i.e. minors under the age of 18) | Yes ☐ | No ☐ |
| People with an intellectual or mental impairment | Yes ☐ | No ☐ |
| People highly dependent on medical case | Yes ☐ | No ☐ |

People in dependent or unequal relationships Yes ☐ No ☐

Do you intend to pay participants? Yes ☐ No ☐

Description of method and amount is included Yes ☐

Description of clinical facilities (for medical research) Yes ☐ No ☐

Period of research Yes ☐ No ☐

**SUPPORTING DOCUMENTATION:** *The committee requires copies of all relevant documents*

Consent form to be signed by participants Yes ☐ No ☐

Information sheet for participants to retain Yes ☐ No ☐

Dot point list of the points that will be made when seeking verbal consent Yes ☐

List of interview questions Yes ☐ No ☐

Copy of questionnaire/s Yes ☐ No ☐

Invitation or introductory letter/s Yes ☐ No ☐

Publicity material (posters etc.) Yes ☐ No ☐

Other (*specify*) Yes ☐ No ☐

**18. SIGNATURES AND UNDERTAKINGS**

**PROPOSER OF THE RESEARCH**

**I certify that the above is as accurate a description of my research proposal as possible and that the research will be conducted in accordance with the *National Statement on Ethical Conduct in Research Involving Humans* (version current at time of application). I also agree to adhere to the conditions of approval stipulated by the ANU Human Research Ethics Committee (HREC) and will cooperate with HREC monitoring requirements. I agree to notify the Committee in writing immediately of any significant departures from this protocol and will not continue the research if ethical approval is withdrawn and will comply with any special conditions required by the HREC.**

Name and title (please print): Mr Andrew Alexander Dankers
(Proposer of research)

Signed:  Date: 27/11/2006

**ANU SUPERVISOR**

***Where the proposal is from a student, the ANU Supervisor is asked to certify the accuracy of the above account.***

**I certify that I shall provide appropriate supervision to the student to ensure that the project is undertaken in accordance with the undertakings above:**

Name and title (please print): Dr. Nick Barnes
(ANU Supervisor)

ANU Department/School/Centre: ANU Dept. Info Eng / RSISE

Signed:…….. Date: 27/11/2006

**COMMENT ON PROJECT FROM HEAD OF ANU DEPARTMENT/GROUP/CENTRE:**

***The Head of ANU Department/School/Centre is asked to certify that this proposal has his/her support:***

**I certify that:**

- **I am familiar with this project and endorse its undertakings;**
- **the resources required to undertake this project are available; and**
- **the investigators have the skill and expertise to undertake this project appropriately.**

Any additional comments (optional):

Name and title (please print):…….
(Head of ANU Department/Group/Centre)

ANU Department/School/Centre: …….

Signed:……. Date:……..

---

**Applications should be submitted as follows:**
**(a) 15 hard copies (one master copy with original signatures + 14 photocopies) and all supporting documentation**
**PLUS**
**(b) an identical email version emailed to Human.Ethics.Officer@anu.edu.au.**

**Hard copies of the completed protocol form, together with all supporting documents, should be sent to**:
The Secretary
Human Research Ethics Committee
Research Services Office
Chancelry 10B

The Australian National University ACT 0200

Tel: 6125-7945
Fax: 6125-4807
Email: Human.Ethics.Officer@anu.edu.au

**Please ensure that the application includes (a) your signature (b) signature of Head of ANU School, Department or Centre; and (c) signature of ANU supervisor (for students).**
**All copies of your application must be secured. Do not send loose pages.**

-10-

List of Attachments:

Attachment A – Advertisement
Attachment B – Consent form
Attachment C – Information sheet
Attachment D – Brief Questionnaire
Attachment E – Sample Snellen Chart

**Attachment A**
**Advertisement**


**Comparing the Performance of a Synthetic Vision System to Human Performance**

Researchers at the Research School of Information Sciences and Engineering (RSISE) at the Australian National University are seeking volunteers to participate in experiments designed to observe human attentive reactions to certain dynamic visual stimulus.

We are looking for participants able to come to the RSISE to participate in the brief trials. Participation should not require more than 20-30 minutes per person. Data from the trials will be used for a statistical comparison between the attentive behaviours of the synthetic and human vision systems.

Trials will be conducted in January and February 2007. Volunteers are asked to express interest in participating by contacting Andrew Dankers by phone at: x58685, or by email at: andrew.dankers@anu.edu.au. You will be sent an information sheet that may answer some of your questions. Any other questions or enquiries welcome!

**Attachment B**
**Consent Form**

**CONSENT FORM**

I, _____ agree to participate in the research trials aimed to observe statistics associated with the human visual attention.

- I understand that I am free to withdraw from the study at any time without needing to give any reason.
- I understand that if I withdraw from the study I can choose to have the information that I have provided destroyed, or I can contribute it to the study.
- I understand that trial data will be stored on a password protected computer at the Research School of Information Sciences and Engineering at the Australian National University.
- I understand that this study has been approved by the Australian National University Ethics Committee, and that if I have an questions regarding this study I may contact the researchers:

  Andrew Dankers
  Research School of Information Sciences and Engineering
  The Australian National University ACT 0200

- I understand that if I have any queries about the ethics of this research I can contact the Australian National University Human Ethics Officer at the ANU Research Office on (02) 6125 2900.
- I will endeavour not to tell prospective participants the exact nature of the trials, until they have completed their trial.

Signature: …… Date: ……..

Please return this consent form to [Research Assistant]

**Attachment C**
**Information Sheet**

**INFORMATION SHEET**

**Comparing the Performance of a Synthetic Vision System to Human Performance**

Thankyou for your expression of interest to participate in experiments designed to observe human attentive reactions to dynamic visual stimulus. Below is additional information that may help you to participate. We hope you are able to attend!

**Location**
Trials will be conducted in the Vision Lab, level 3 RSISE (Bldg 115), ANU.

**What to Expect**
We have prepared a small dynamic 3D scene that we want you to observe. We may ask you to perform a basic visual task (like "count any yellow balls that enter the scene"). However, we are not interested in the correctness of your answers, or performance of each task. We are more interested in observing where you instinctively direct your gaze as novel visual events occur while performing the task

During the trial, you will be seated in front of the scene (it's a bit like watching a puppet show!), and your gaze directions will be recorded by an automated eye-tracking system.

**Confidentiality**
Participants may withdraw from the trials at any time, and request that data associated with their trial be destroyed. Before the trials you will be asked to read and sign a participation consent form. You will be asked some non-specific yes/no questions about your vision condition that you may choose not to answer. You will also be asked to conduct a brief visual acuity test that you may choose to decline. The questions and acuity test should take less than a minute. Results will be associated with a trial number, not your identity. You may, however, choose to associate your identity with your trial data if you are interested in viewing your results at a later date. You may also provide contact details if you wish to be contacted in the event that we require additional trials. In these events, your name will be kept on file no longer than one month. All trial data will be kept no longer than one year. Thereafter, only statistical summaries will be retained.

**Questions**
If you have any questions relating to the trials, please don't hesitate to ask the Research Contact listed below. If you have ethical questions or concerns, you may contact:

Human Research Ethics Committee
Research Services Office
Chancelry 10B
The Australian National University ACT 0200
Tel: 6125-7945
Fax: 6125-4807
Email: Human.Ethics.Officer@anu.edu.au

**Time**
Trials should take no longer than 2 minutes each. We hope participants can complete up to 5 such 2-minute trials. Trials will be conducted during January and February 2007.

**Confirming Participation**
If you would like to participate, please contact the Research Contact and specify a time that's best for you!

**Research Contact**
Andrew Dankers, x58685, andrew.dankers@anu.edu.au

Thankyou!
Andrew

**Attachment D**
**Brief Questionnaire**

TRIAL ID NUMBER _____

**Questionnaire**

Thank you for agreeing to participate in our study. This questionnaire provides us with information that may be used to consider the reliability of trial data.

**Section A**

1. What is your date of birth? _____

2. Are you currently taking medication that may cause drowsyness, or that may affect your vision?

Y / N

3. To your knowledge, do you experience any vision conditions, such as colour blindness, tunnel vision, short sightedness, or other, that may affect your vision?

Y / N

4. Do you normally wear prescription glasses?

Y / N

5. Please rate your current level of alertness:

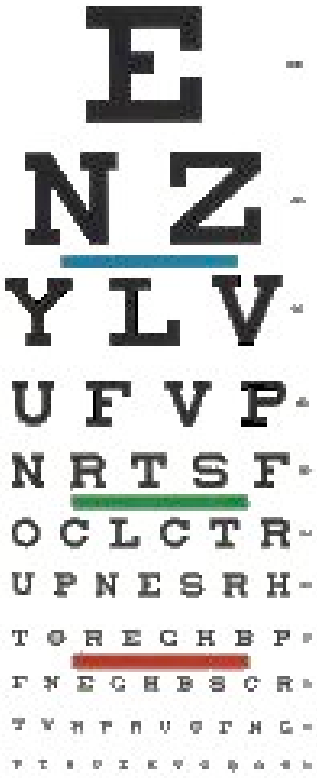
(Drowsy) 1 2 3 4 5 (Alert)

**Section B**

Please read the five Snellen visual acuity chart letters as requested by technician (Attachment E).

1) Correct / Incorrect
2) Correct / Incorrect
3) Correct / Incorrect
4) Correct / Incorrect
5) Correct / Incorrect

**Attachment E**
**Sample Visual Acuity Test Chart (not to scale)**



## A.2 Ethics Approval


THE AUSTRALIAN NATIONAL UNIVERSITY

RESEARCH OFFICE

CANBERRA ACT 0200 AUSTRALIA
TELEPHONE:(02) 6125 7945
FACSIMILE: (02) 6125 4807
EMAIL: Yolanda.Shave@anu.edu.au

Ms Yolanda Shave
Secretary, Human Research Ethics Committee

4 January 2007

Mr Andrew Alexander Dankers
Postgraduate Student,
Systems Engineering
Research School of Information Sciences and Engineering
The Australian National University
ACT 0200

Dear Mr Dankers,

**Protocol 2006/344**
***Comparison of synthetic stereo active vision attention system with human attention***

On behalf of the Human Research Ethics Committee I am pleased to advise that the above protocol has been approved as per the attached *Outcome of Consideration of Protocol.*

For your information:

1. Under the NHMRC/AVCC *National Statement on Ethical Conduct in Research Involving Humans* we are required to follow up research that we have approved. Once a year (or sooner for short projects) we shall request a brief report on any ethical issues which may have arisen during your research and whether it proceeded according to the plan outlined in the above protocol.
2. Please notify the Committee of any changes to your protocol in the course of your research, and when you complete or cease working on this project.
3. The validity of this current approval is five years' maximum from the date shown on the attached *Outcome of Consideration of Protocol* form. For longer projects you are required to seek renewed approval from the Committee.

Yours sincerely,

JRadcliffe

per

Ms Yolanda Shave
Secretary, Human Research Ethics Committee

# A. HUMAN TRIALS: ETHICS

THE AUSTRALIAN NATIONAL UNIVERSITY

THE AUSTRALIAN NATIONAL UNIVERSITY

HUMAN RESEARCH ETHICS COMMITTEE

Outcome of Consideration of Protocol

**Researcher:** Mr Andrew Alexander Dankers
**Contact details:** Postgraduate Student, Systems Engineering, Research School of Information Sciences and Engineering
**Protocol No.** 2006/344
**Title:** Comparison of synthetic stereo active vision attention system with human attention
**Date on application:** 27 November 2006 **Date received in Research Office:** 30 November 2006

**On behalf of the Human Research Ethics Committee,**

**I approve/do not approve the above protocol.**

**Approval is subject to the following conditions:**

.....

.....

.....

**Reasons for non-approval:** .....

.....

.....

.....

**Review due:** .....

**Chairperson:** ..... **Date:** 21/12/06

Prof Lawrence Cram

# Appendix B

# Trial Results

## B.1 Human Trials

### B.1.1 Individual trial results

#### B.1.1.1 Pilot 1

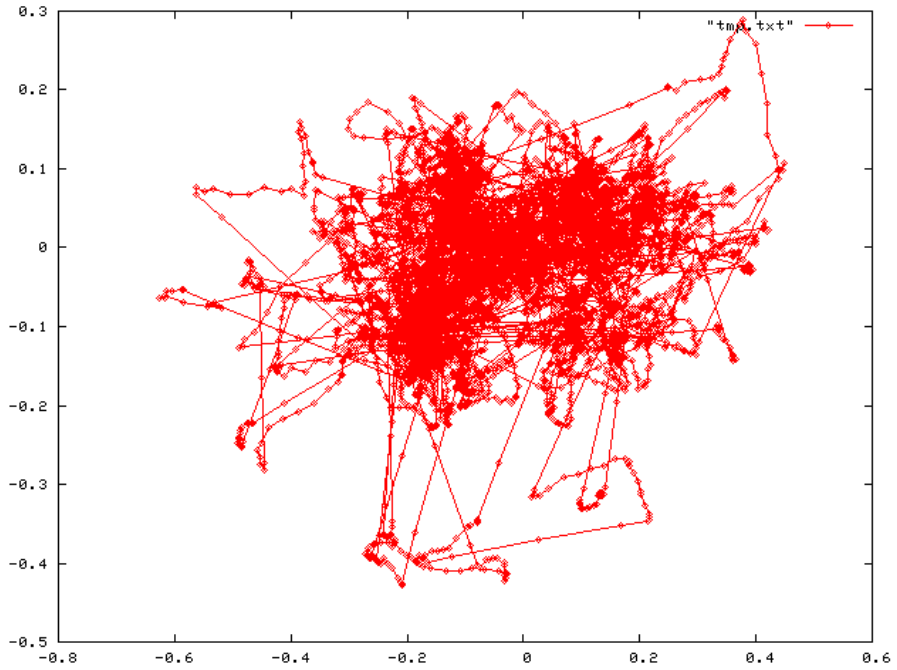



Figure B.1: Complete scan path, Pilot 1.

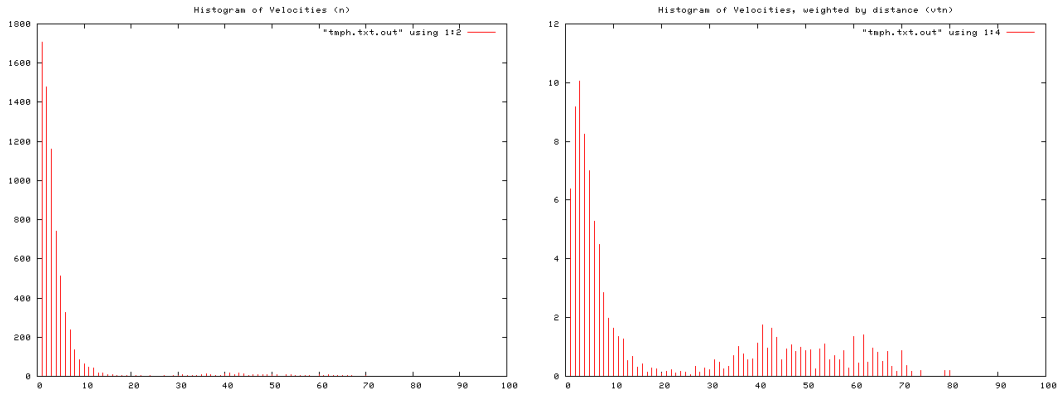



Figure B.2: Histogram of velocity magnitudes, Pilot 1 (left). Histogram of distance weighted velocities, Pilot 1 (right).

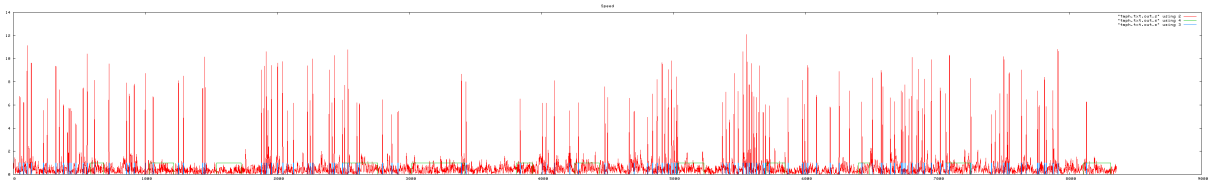


Figure B.3: Velocity profile. Velocity magnitude of each frame, Pilot 1.

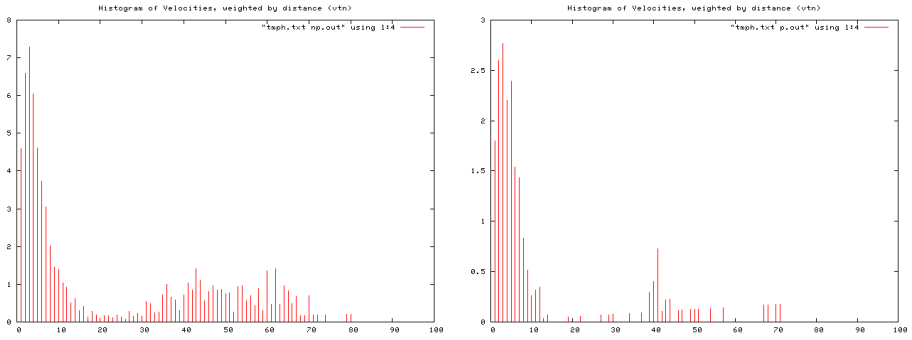



Figure B.4: Histogram of velocities during periods of no perturbation (left), and perturbation (right), Pilot 1.

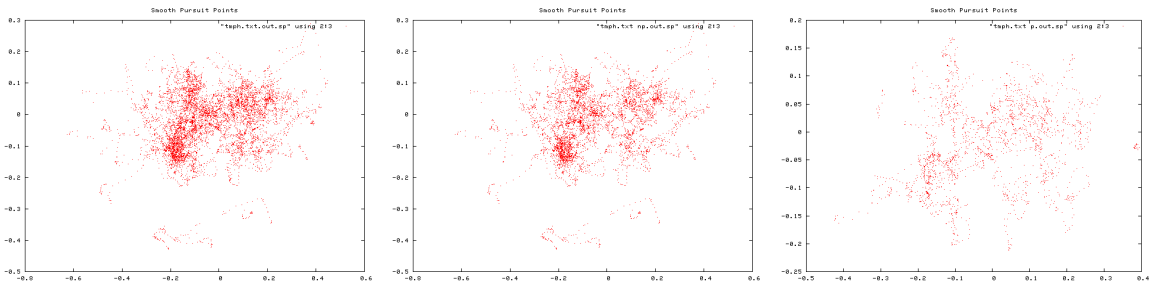



Figure B.5: Smooth pursuit gaze locations, Pilot 1. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

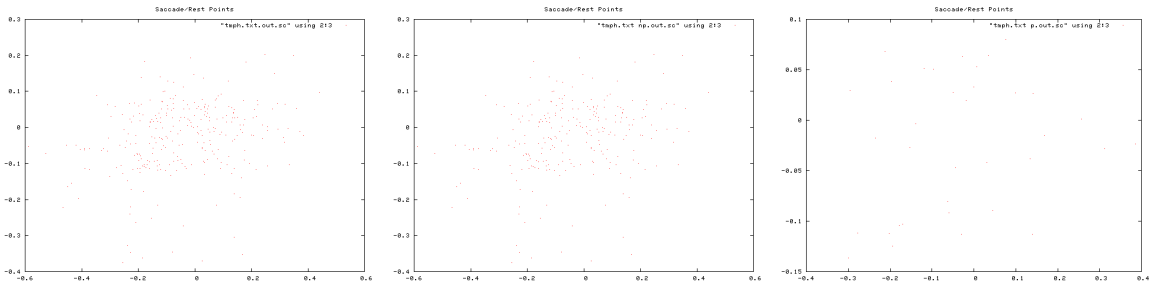



Figure B.6: Saccade gaze locations, Pilot 1. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

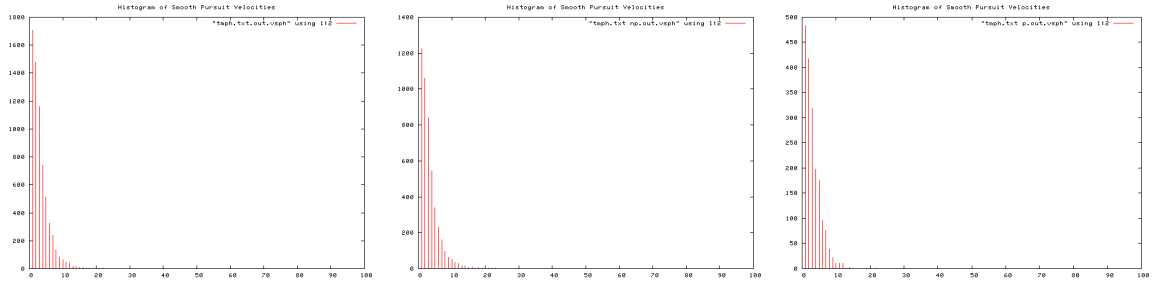



Figure B.7: Histogram of smooth pursuit velocities, Pilot 1. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

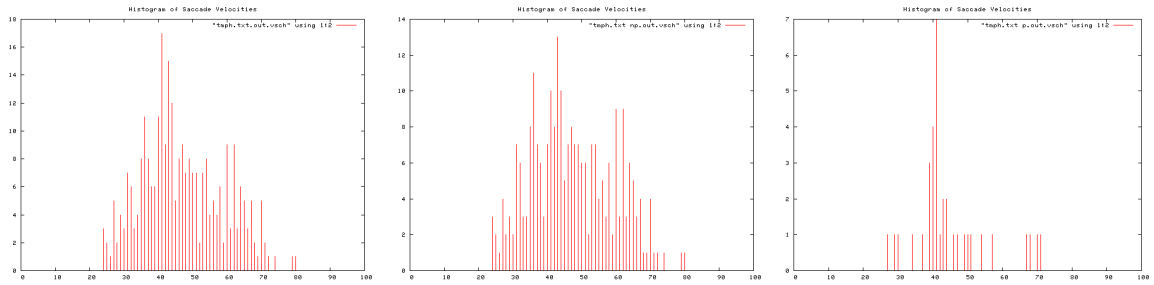



Figure B.8: Histogram of Saccade velocities, Pilot 1. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

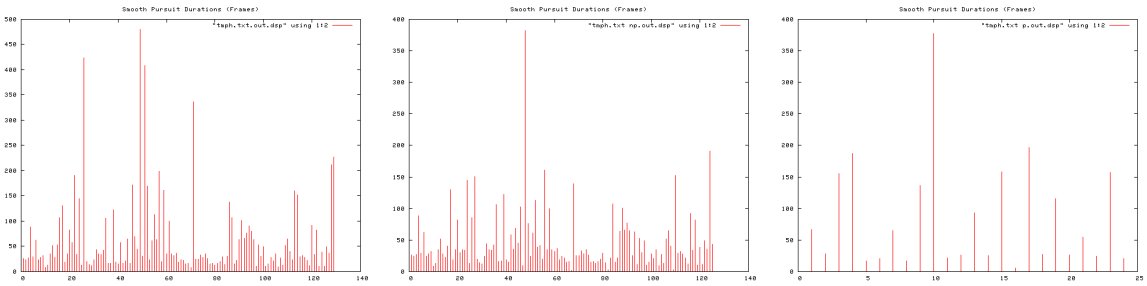



Figure B.9: Smooth pursuit durations, Pilot 1. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

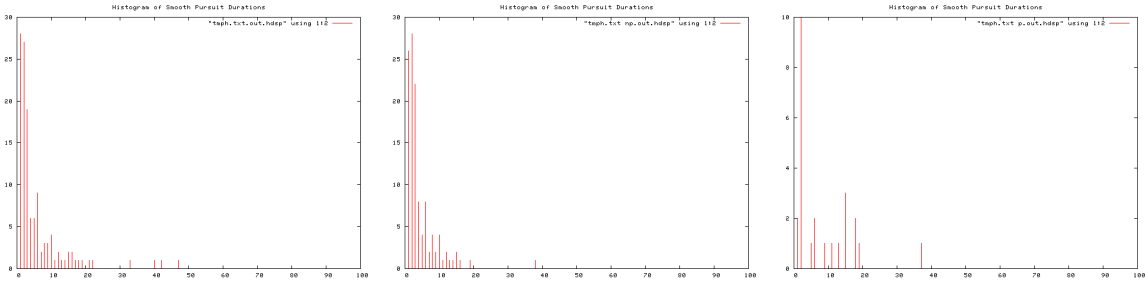



Figure B.10: Histogram of Smooth pursuit durations, Pilot 1. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

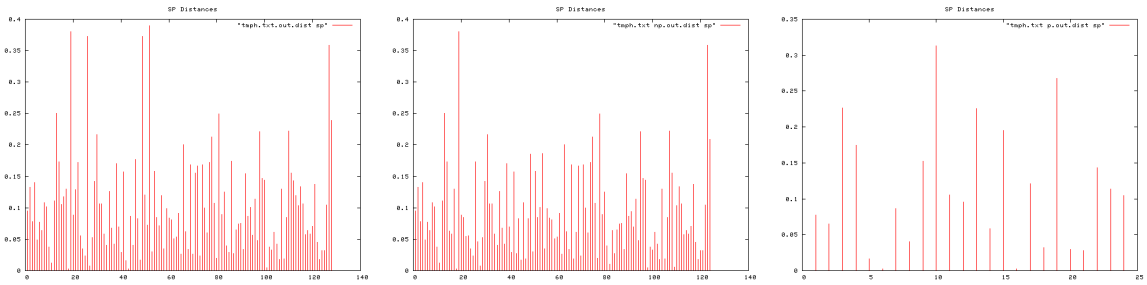



Figure B.11: Smooth pursuit distances, Pilot 1. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

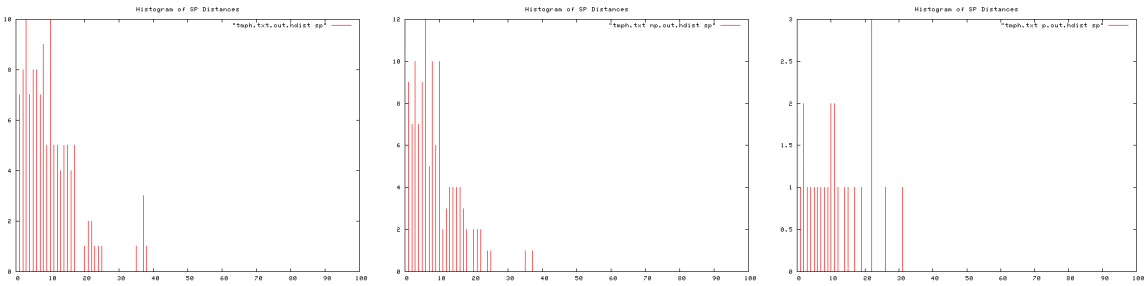



Figure B.12: Histogram of smooth pursuit distances, Pilot 1. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

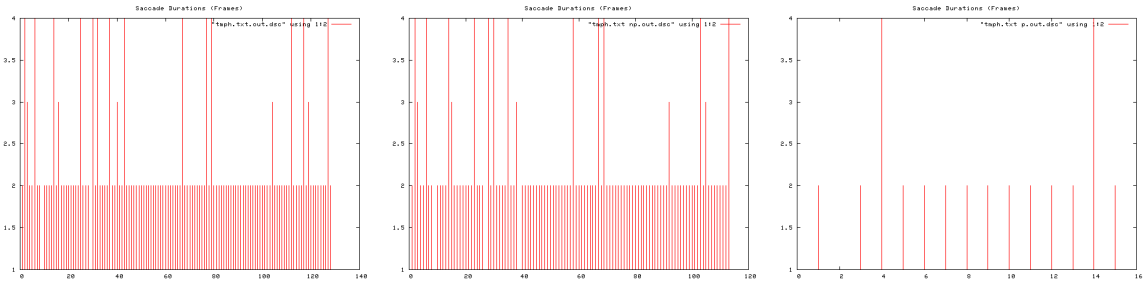



Figure B.13: Saccade durations, Pilot 1. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

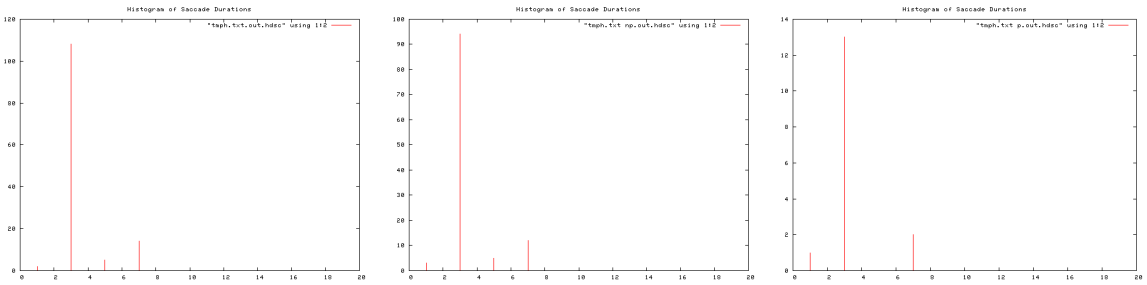



Figure B.14: Histogram of saccade durations, Pilot 1. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

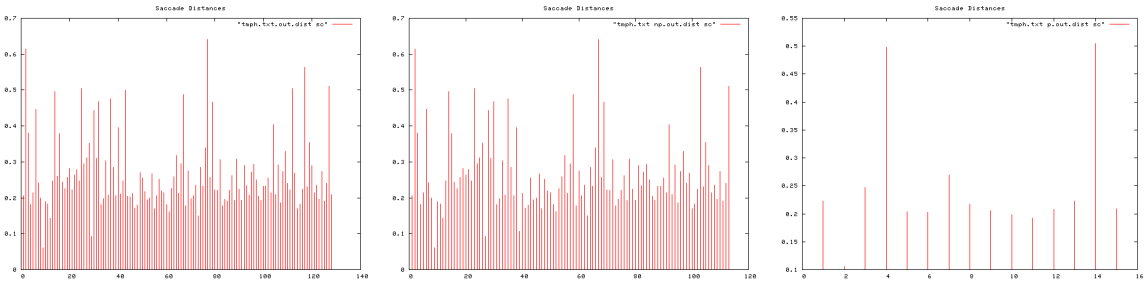



Figure B.15: Saccade distances, Pilot 1. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

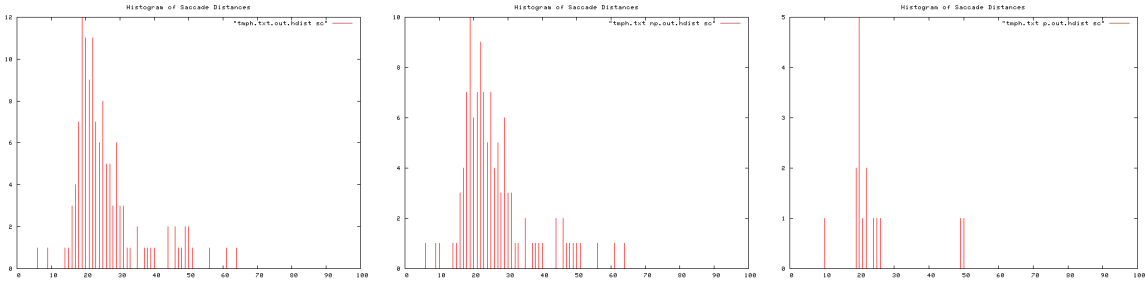



Figure B.16: Histogram of saccade distances, Pilot 1. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

| Ti | To | T | I/O Event | Pear | Apple | Peach | Orange | |
|---|---|---|---|---|---|---|---|---|
| 12 | 17 | 5 | Orange In | | | | 1 | |
| 20 | 23 | 3 | Pear In | 1 | | | 1 | |
| 29 | 43 | 4 | | 6 | | | 5 | |
| 46 | 49 | 3 | Peach In | 1 | | 1 | 1 | |
| 57 | 1:03 | 6 | | 2 | | 1 | 1 | |
| 1:07 | 1:12 | 5 | | 1 | | 1 | 1 | |
| 1:16 | 1:23 | 7 | | 2 | | 3 | 2 | |
| 1:29 | 1:35 | 6 | Apple In, Peach Out | 1 | 2 | 1 | 1 | |
| 1:39 | 1:47 | 8 | Orange Out | 2 | 2 | | 3 | |
| 1:49 | 1:58 | 9 | Pear Out | 3 | 5 | | | |
| 2:02 | 2:15 | 13 | Apple Out | | 3 | | | |
| | | | TOTAL Rets | 19 | 12 | 7 | 16 | |
| | | | TOTAL T | 51 | 36 | 27 | 47 | **SD** |
| | | | **Av. Re-attention Period** | **2.7** | **3** | **3.8** | **2.9** | 0.48304589 |

Figure B.17: Re-attention period statistics, Pilot 1.

### B.1.1.2 Pilot 2

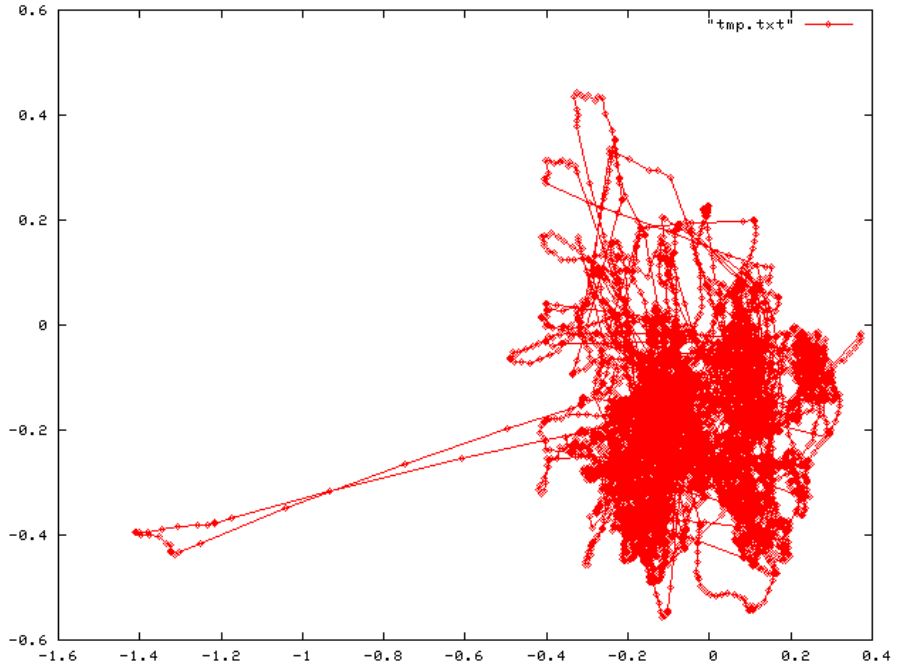



Figure B.18: Complete scan path, Pilot 2.

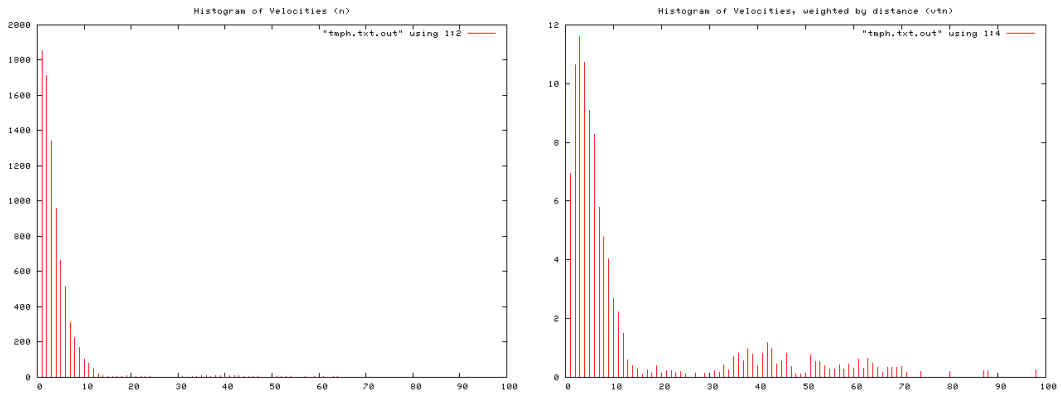



Figure B.19: Histogram of velocity magnitudes, Pilot 2 (left). Histogram of distance weighted velocities, Pilot 2 (right).

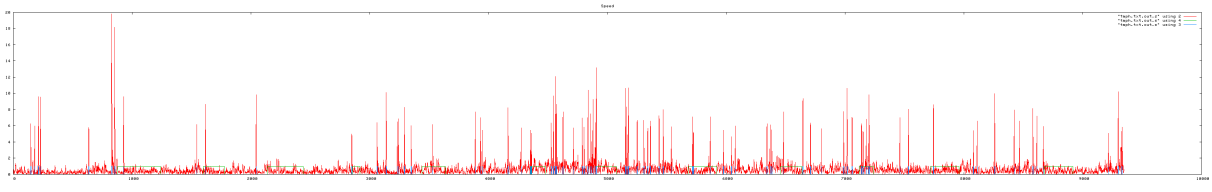


Figure B.20: Velocity profile. Velocity magnitude of each frame, Pilot 2.

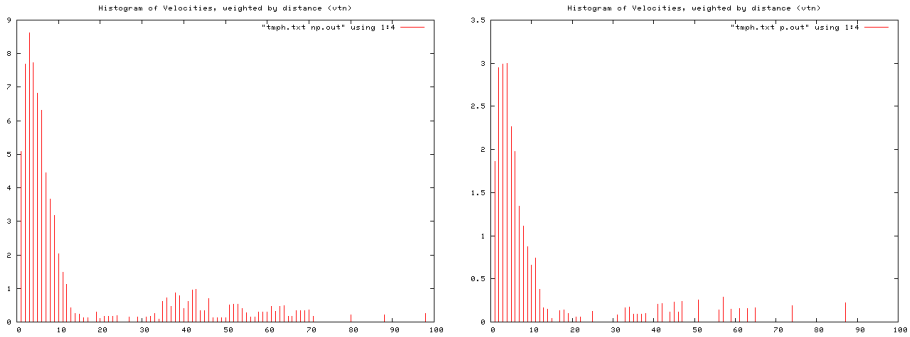



Figure B.21: Histogram of velocities during periods of no perturbation (left), and perturbation (right), Pilot 2.

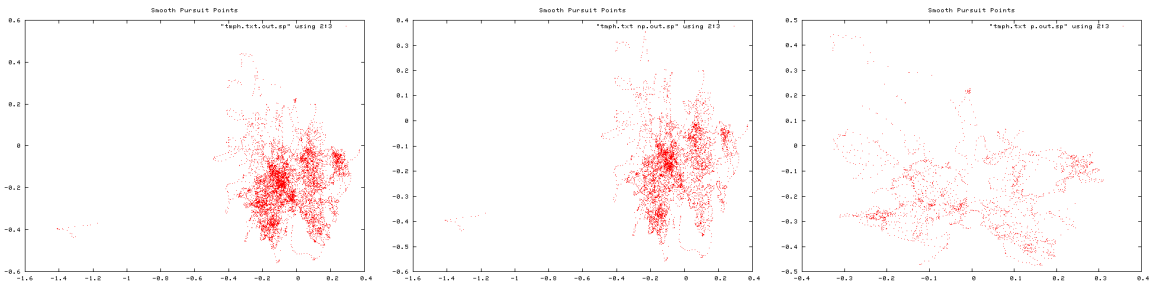



Figure B.22: Smooth pursuit gaze locations, Pilot 2. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

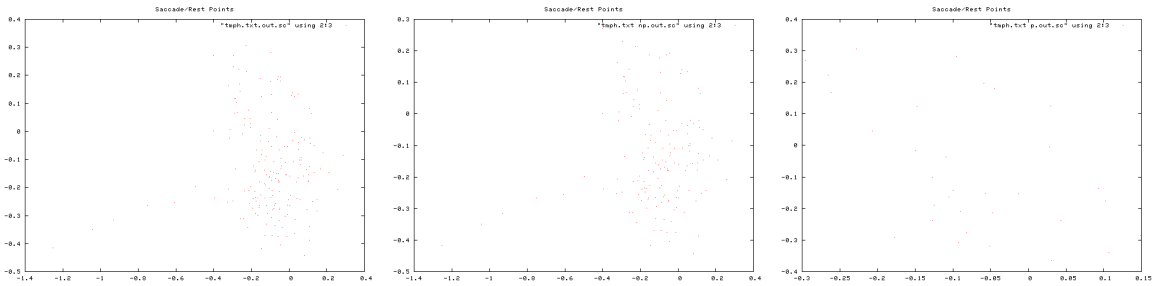



Figure B.23: Saccade gaze locations, Pilot 2. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

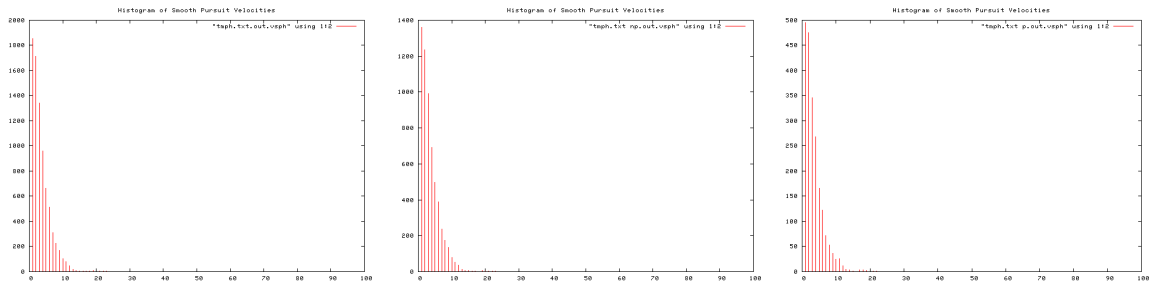



Figure B.24: Histogram of smooth pursuit velocities, Pilot 2. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

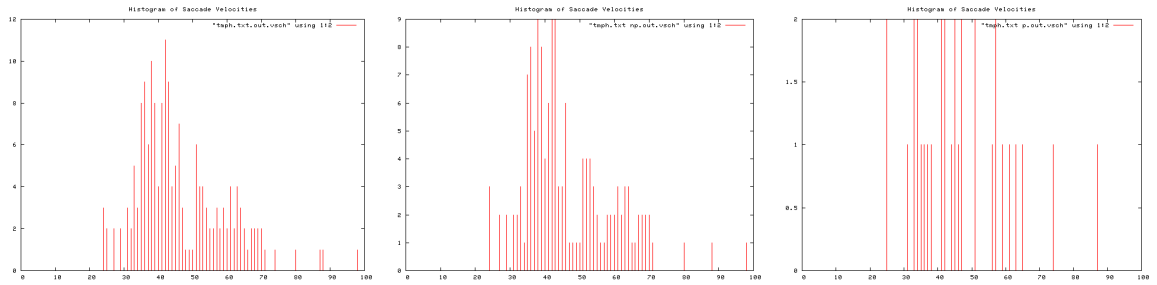



Figure B.25: Histogram of Saccade velocities, Pilot 2. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

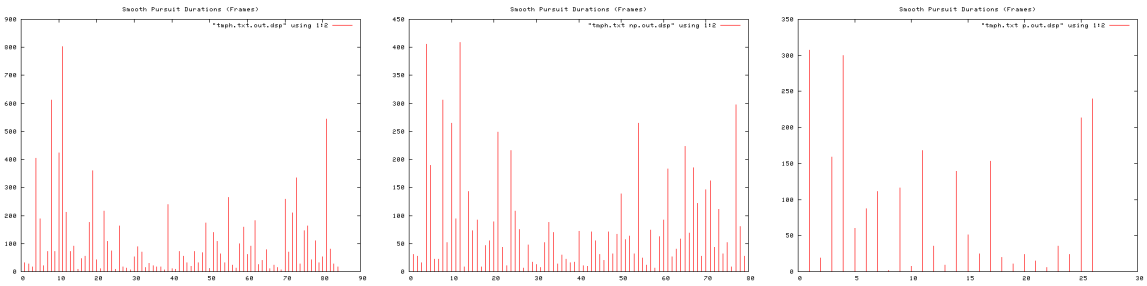



Figure B.26: Smooth pursuit durations, Pilot 2. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

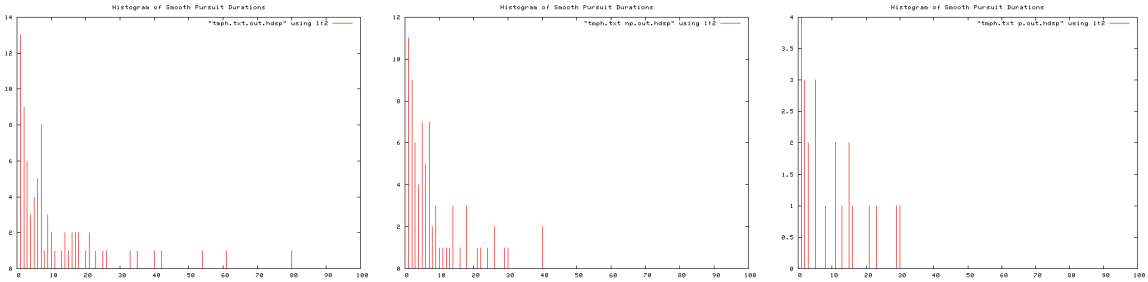



Figure B.27: Histogram of Smooth pursuit durations, Pilot 2. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

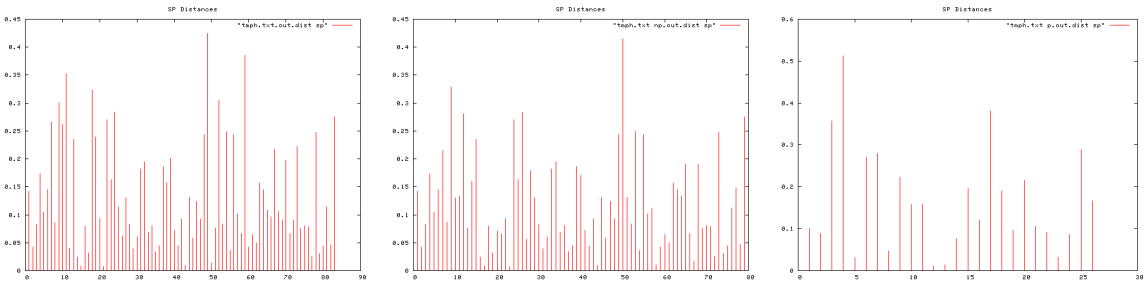



Figure B.28: Smooth pursuit distances, Pilot 2. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

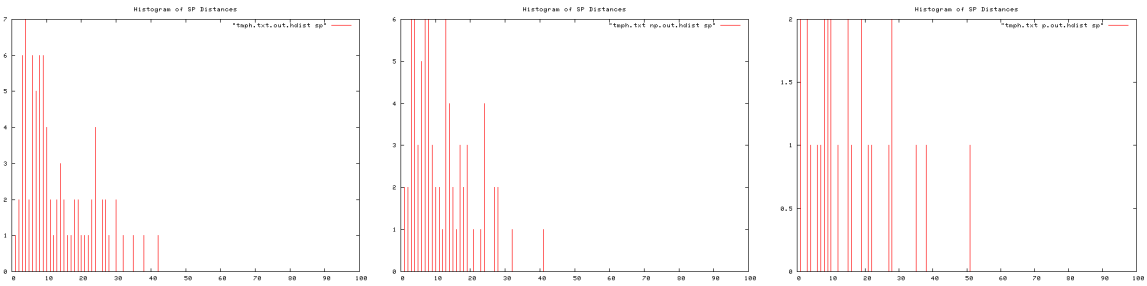



Figure B.29: Histogram of smooth pursuit distances, Pilot 2. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

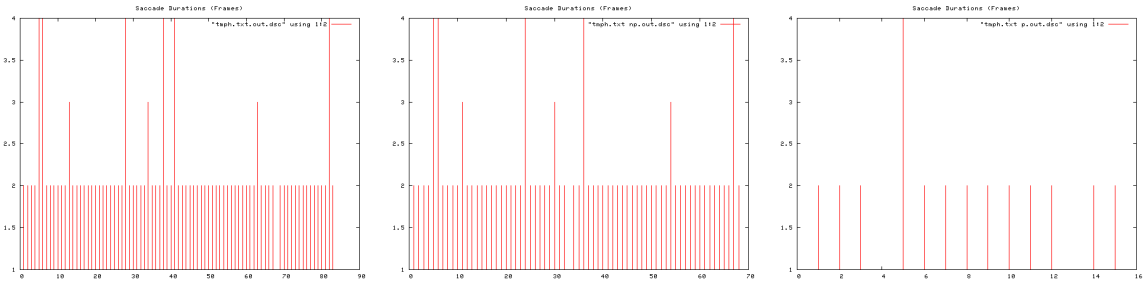



Figure B.30: Saccade durations, Pilot 2. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

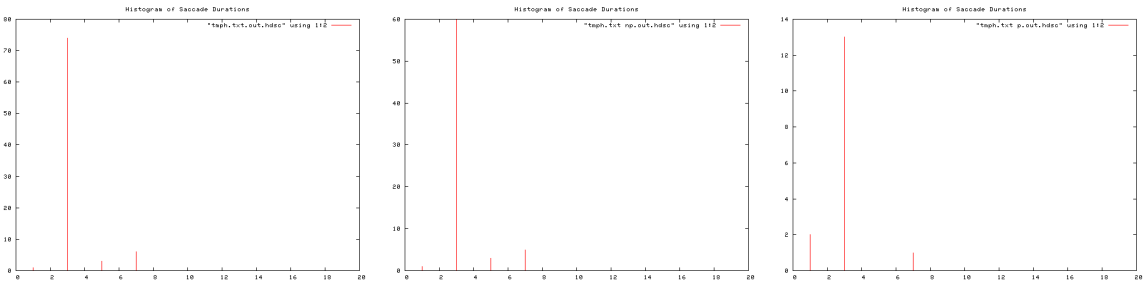



Figure B.31: Histogram of saccade durations, Pilot 2. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

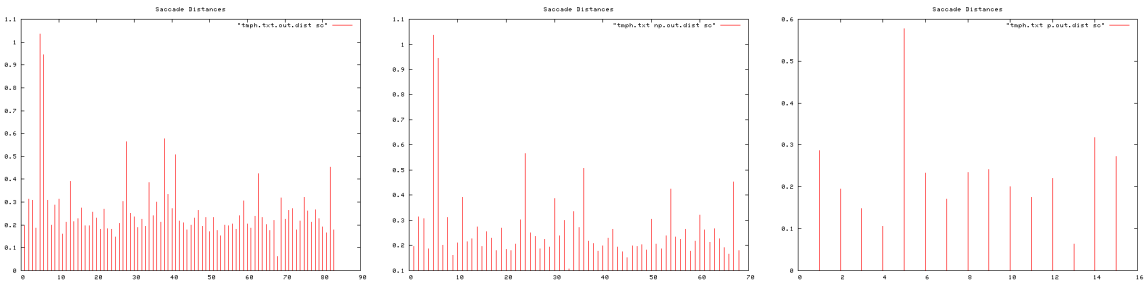



Figure B.32: Saccade distances, Pilot 2. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

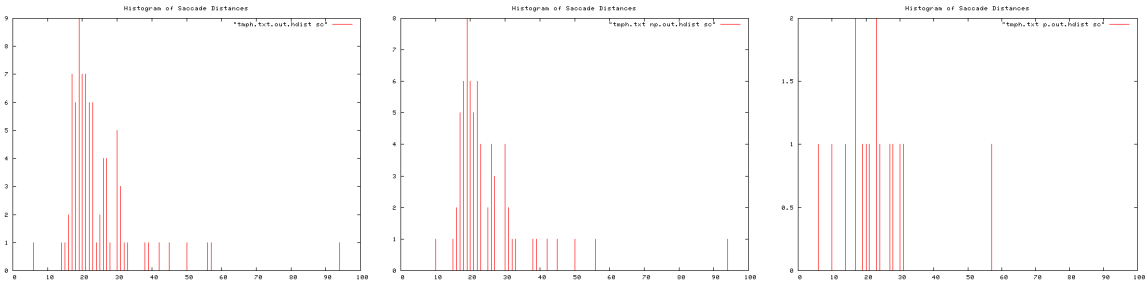



Figure B.33: Histogram of saccade distances, Pilot 2. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

| Ti | To | T | I/O Event | Pear | Apple | Peach | Orange | |
|---|---|---|---|---|---|---|---|---|
| 20 | 26 | 6 | Orange In | | | | 1 | |
| 29 | 33 | 4 | Pear In | 1 | | | 1 | |
| 40 | 47 | 7 | | 1 | | | 1 | |
| 49 | 55 | 6 | Peach In | 3 | | 2 | 2 | |
| 1:00 | 1:11 | 11 | | 2 | | 3 | 1 | |
| 1:13 | 1:20 | 7 | | 1 | | 1 | 1 | |
| 1:24 | 1:34 | 10 | | 3 | | 3 | 3 | |
| 1:38 | 1:47 | 9 | Apple In, Peach Out | 0 | 1 | 0 | 1 | |
| 1:51 | 1:58 | 7 | Orange Out | 2 | 3 | | 2 | |
| 1:59 | 2:08 | 9 | Pear Out | 2 | 2 | | | |
| 2:12 | 2:23 | 11 | Apple Out | | 2 | | | |
| | | | TOTAL Rets | 15 | 8 | 9 | 13 | |
| | | | TOTAL T | 81 | 36 | 43 | 67 | |
| | | | R/T | 0.19 | 0.22 | 0.21 | 0.19 | **SD** |
| | | | **Av. Re-attention Period** | **5.4** | **4.5** | **4.8** | **5.15** | 0.39449335 |

Figure B.34: Re-attention period statistics, Pilot 2.

### B.1.1.3 Trial 1

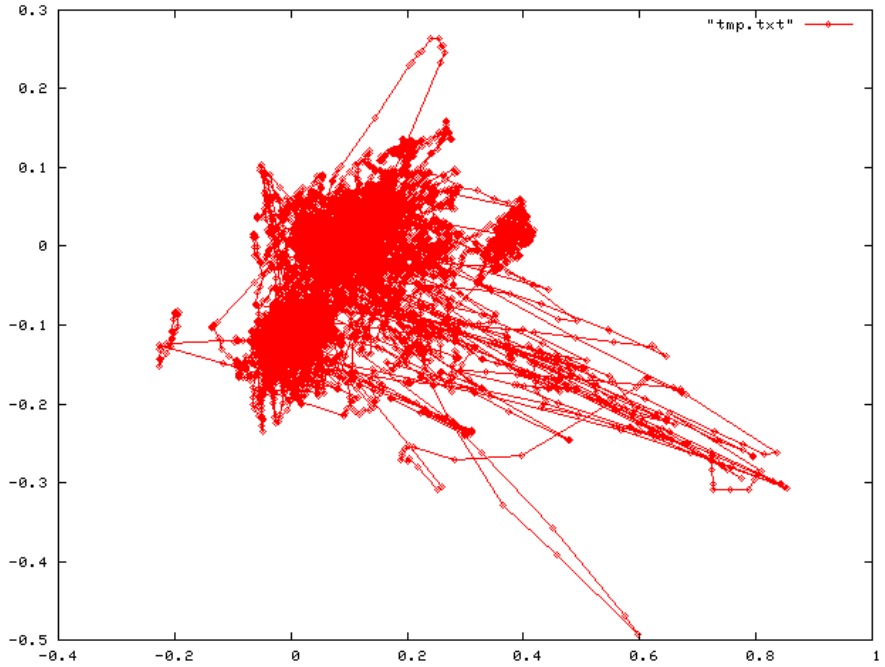



Figure B.35: Complete scan path, Trial 1.

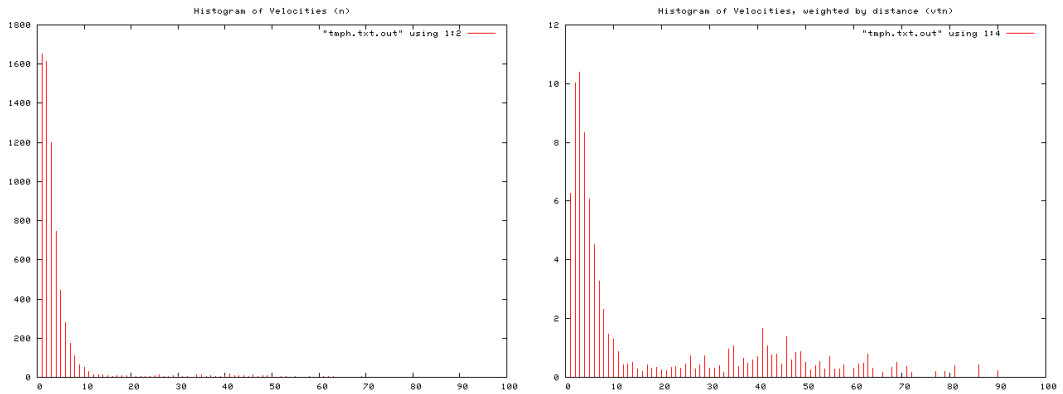



Figure B.36: Histogram of velocity magnitudes, Trial 1 (left). Histogram of distance weighted velocities, Trial 1 (right).

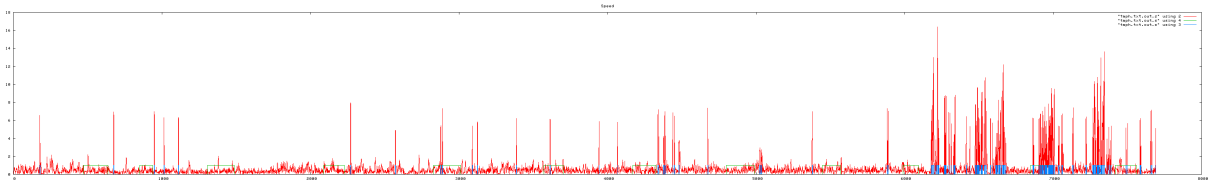


Figure B.37: Velocity profile. Velocity magnitude of each frame, Trial 1.

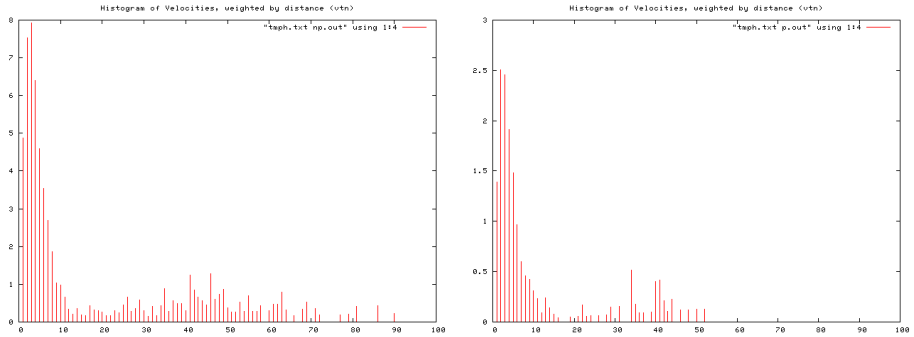



Figure B.38: Histogram of velocities during periods of no perturbation (left), and perturbation (right), Trial 1.

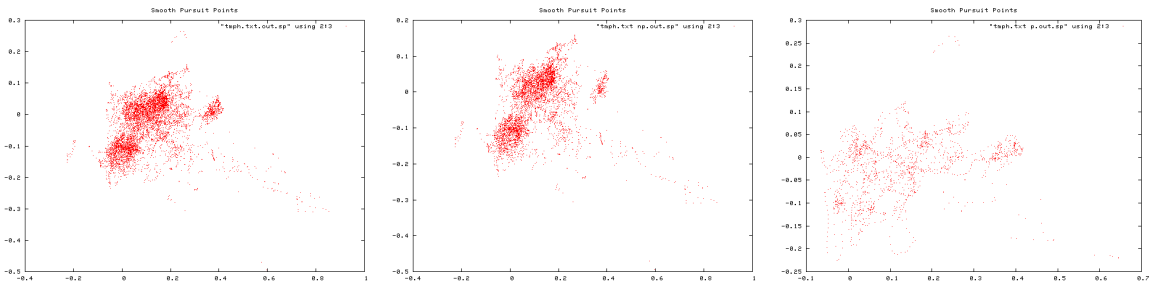



Figure B.39: Smooth pursuit gaze locations, Trial 1. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

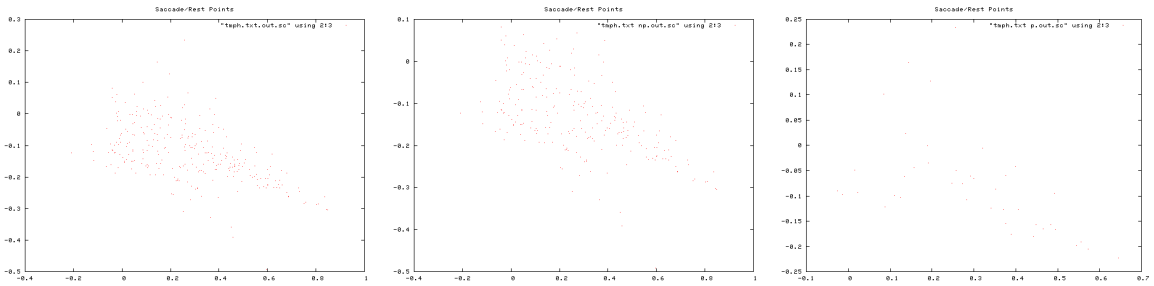



Figure B.40: Saccade gaze locations, Trial 1. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

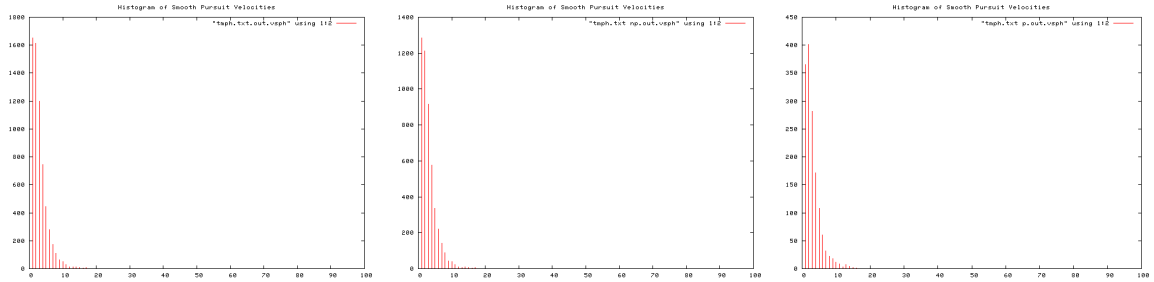



Figure B.41: Histogram of smooth pursuit velocities, Trial 1. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

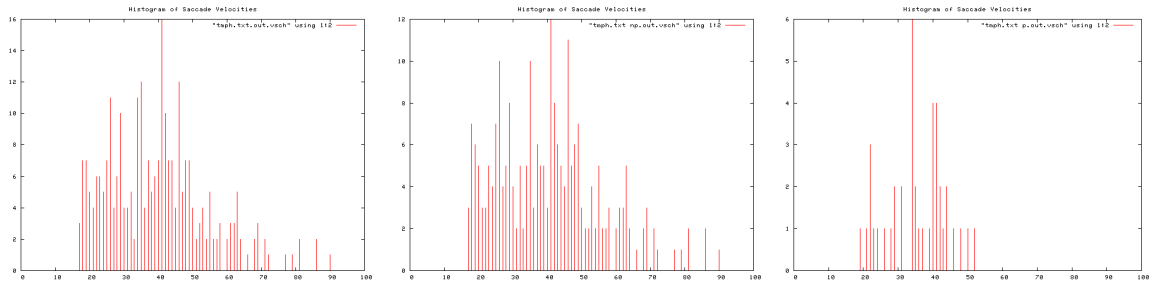



Figure B.42: Histogram of Saccade velocities, Trial 1. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

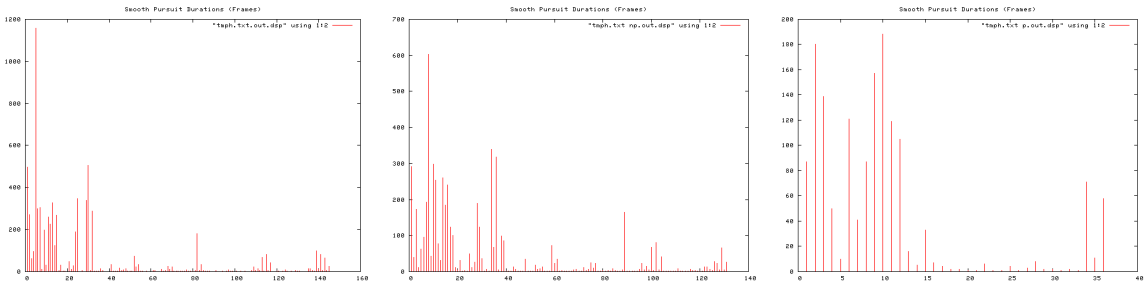



Figure B.43: Smooth pursuit durations, Trial 1. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

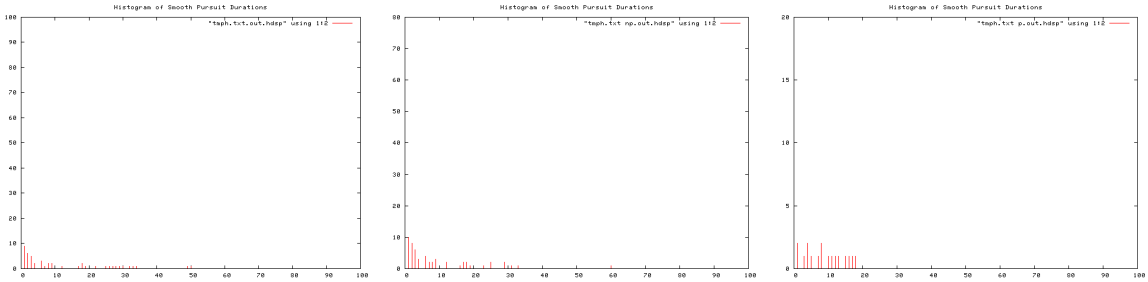



Figure B.44: Histogram of Smooth pursuit durations, Trial 1. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

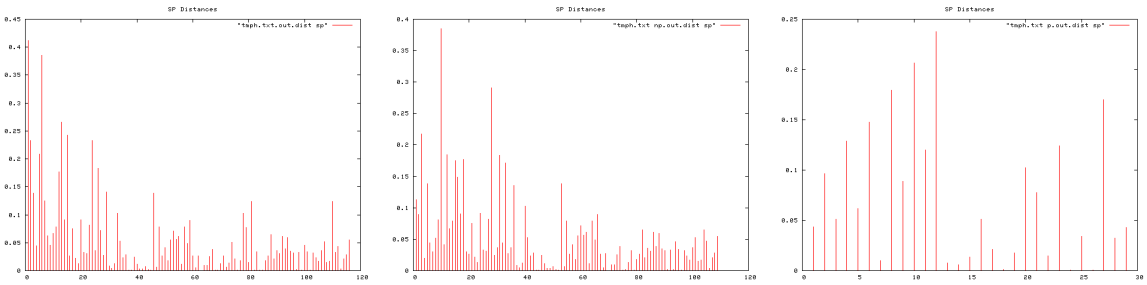



Figure B.45: Smooth pursuit distances, Trial 1. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

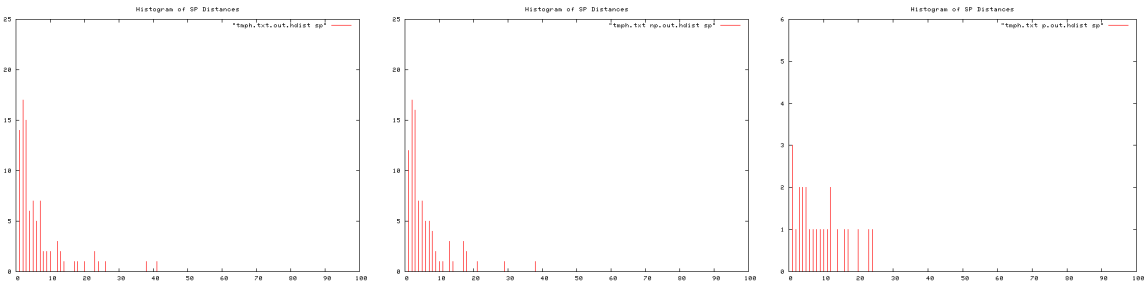



Figure B.46: Histogram of smooth pursuit distances, Trial 1. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

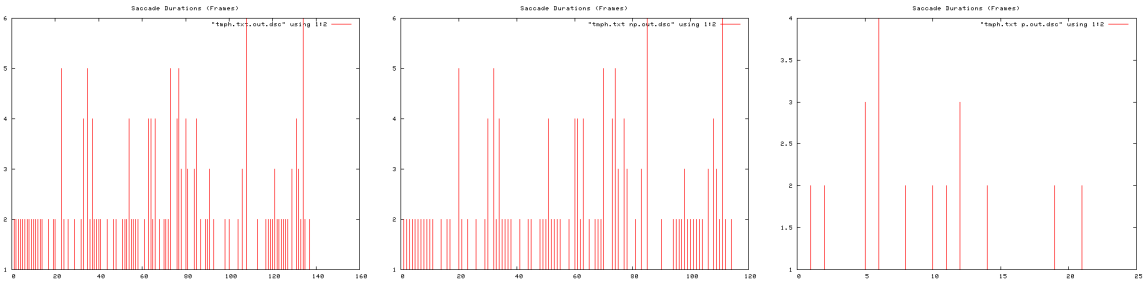



Figure B.47: Saccade durations, Trial 1. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

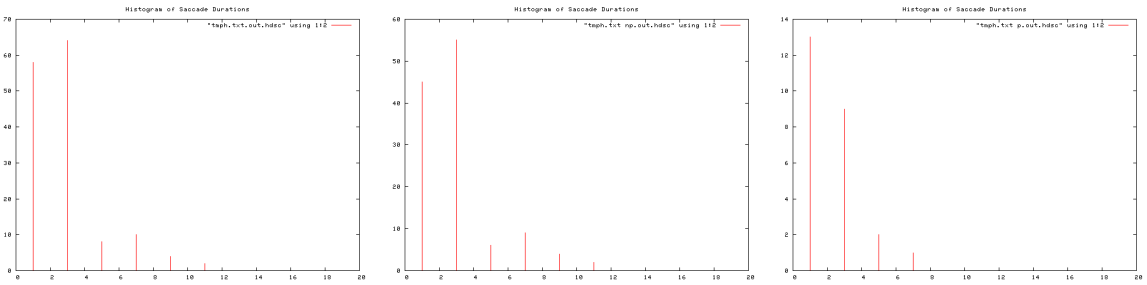



Figure B.48: Histogram of saccade durations, Trial 1. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

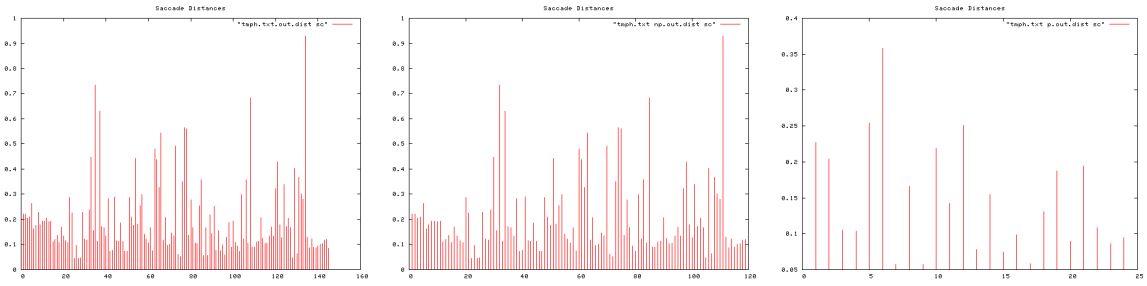



Figure B.49: Saccade distances, Trial 1. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

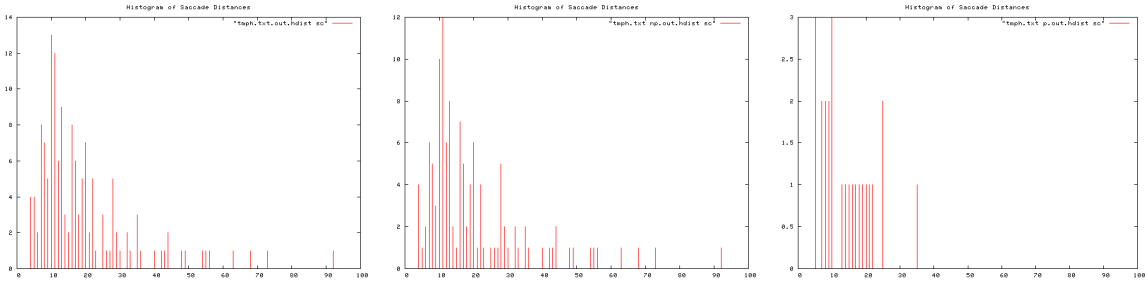



Figure B.50: Histogram of saccade distances, Trial 1. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

| Ti | To | T | I/O Event | Pear | Apple | Peach | Orange | |
|---|---|---|---|---|---|---|---|---|
| 11 | 15 | 6 | Orange In | | | | 1 | |
| 17 | 22 | 5 | Pear In | 1 | | | 1 | |
| 26 | 36 | 10 | | 2 | | | 2 | |
| 40 | 47 | 7 | Peach In | 2 | | 1 | 2 | |
| 49 | 0:58 | 9 | | 3 | | 2 | 2 | |
| 1:02 | 1:10 | 8 | | 3 | | 2 | 2 | |
| 1:14 | 1:19 | 5 | | 1 | | 2 | 2 | |
| 1:24 | 1:30 | 6 | Apple In, Peach Out | 0 | 2 | 1 | 1 | |
| 1:33 | 1:40 | 7 | Orange Out | 0 | 1 | | 1 | |
| 1:42 | 1:53 | 9 | Pear Out | 2 | 2 | | | |
| 1:58 | 2:02 | 4 | Apple Out | | 1 | | | |
| | | | TOTAL Rets | 14 | 6 | 7 | 14 | |
| | | | TOTAL T | 66 | 26 | 35 | 63 | **SD** |
| | | | **Av. Re-attention Period** | **4.7** | **4.3** | **4.4** | **4.5** | 0.17078251 |

Figure B.51: Re-attention period statistics, Trial 1.

### B.1.1.4 Trial 2

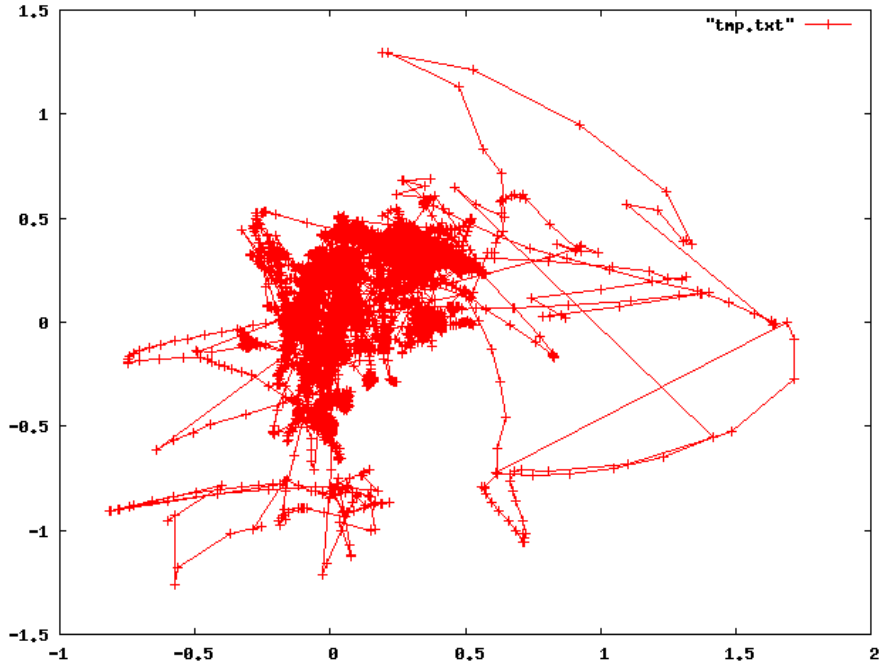



Figure B.52: Complete scan path, Trial 2.

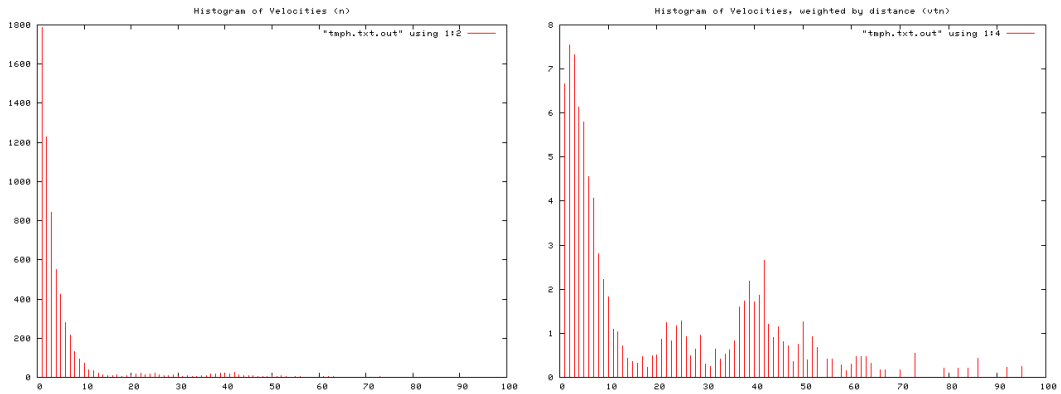



Figure B.53: Histogram of velocity magnitudes, Trial 2 (left). Histogram of distance weighted velocities, Trial 1 (right).

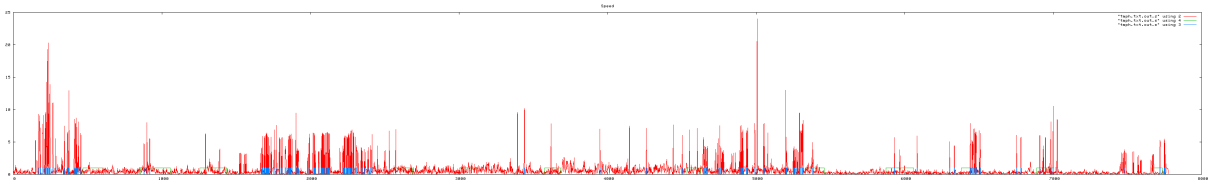


Figure B.54: Velocity profile. Velocity magnitude of each frame, Trial 2.

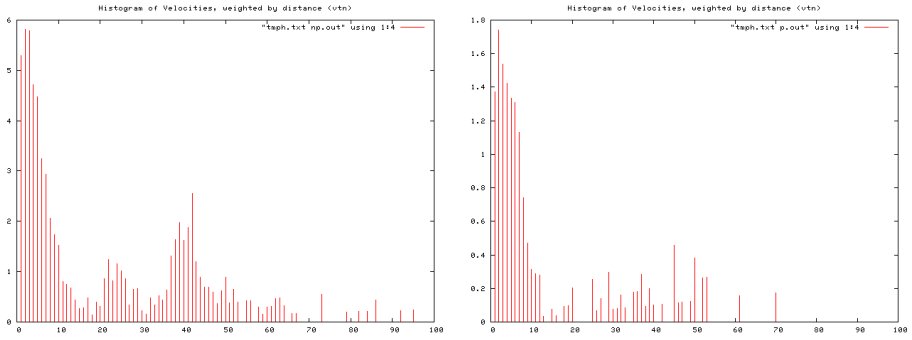



Figure B.55: Histogram of velocities during periods of no perturbation (left), and perturbation (right), Trial 2.

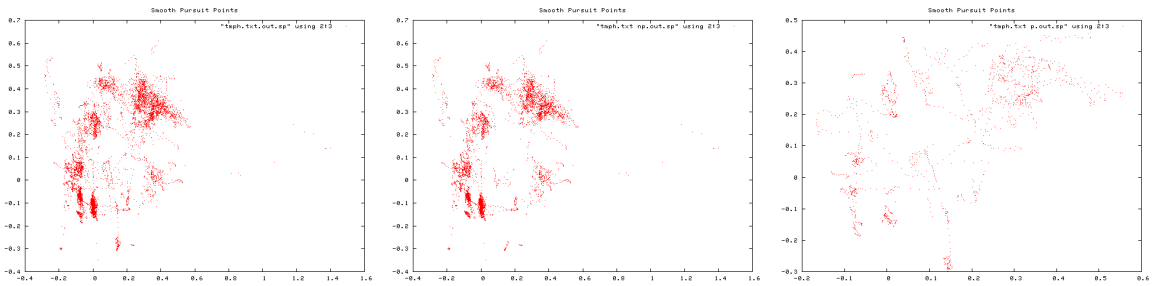



Figure B.56: Smooth pursuit gaze locations, Trial 2. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

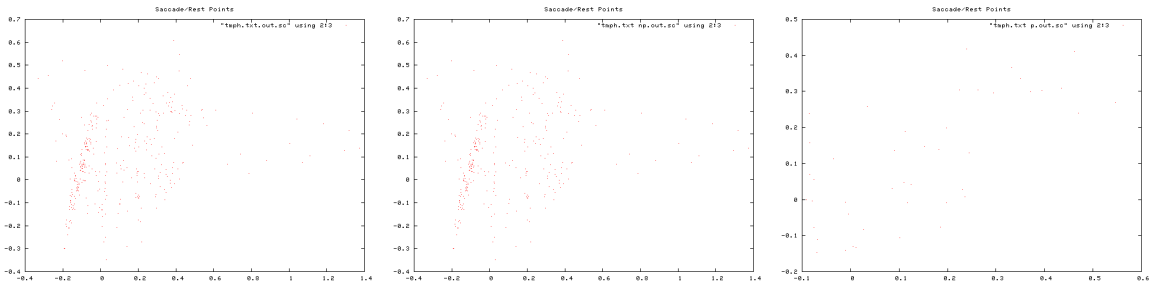



Figure B.57: Saccade gaze locations, Trial 2. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

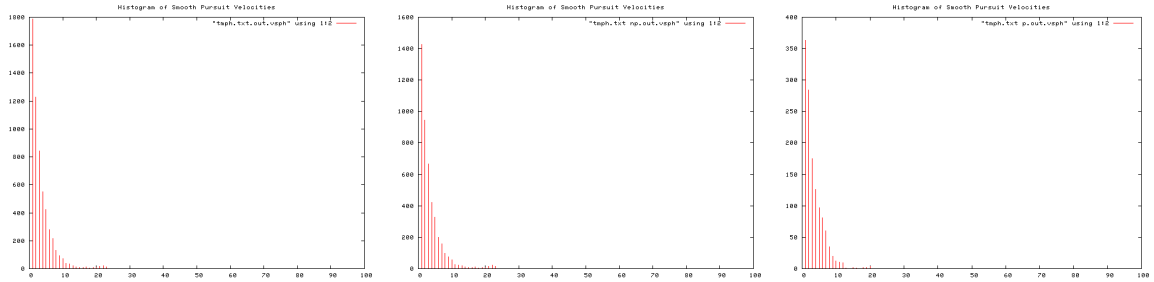



Figure B.58: Histogram of smooth pursuit velocities, Trial 2. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

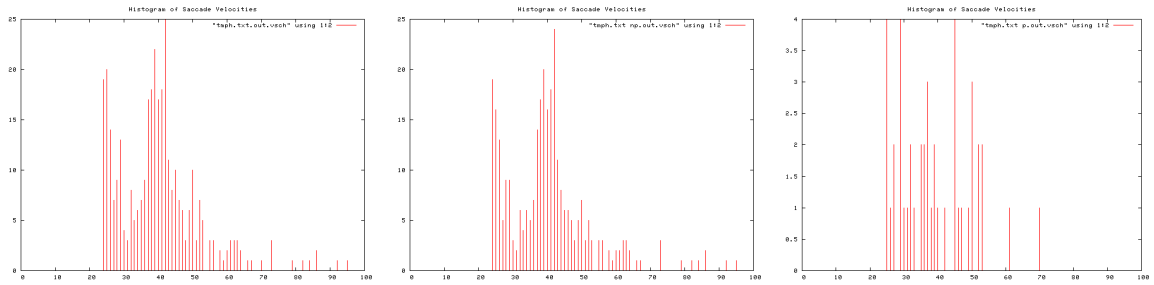



Figure B.59: Histogram of Saccade velocities, Trial 2. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

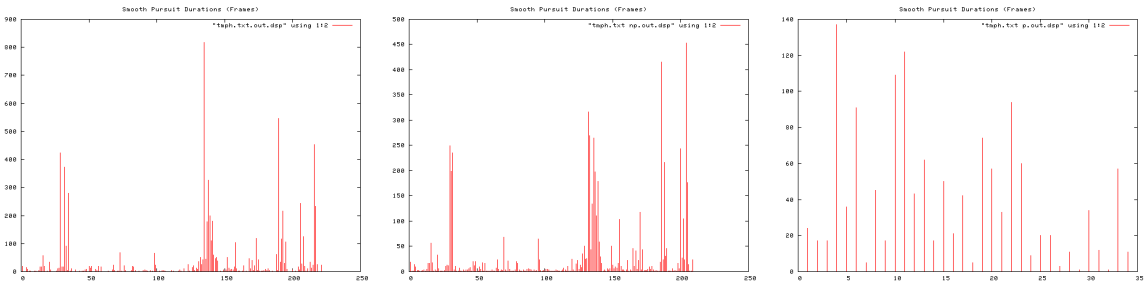



Figure B.60: Smooth pursuit durations, Trial 2. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

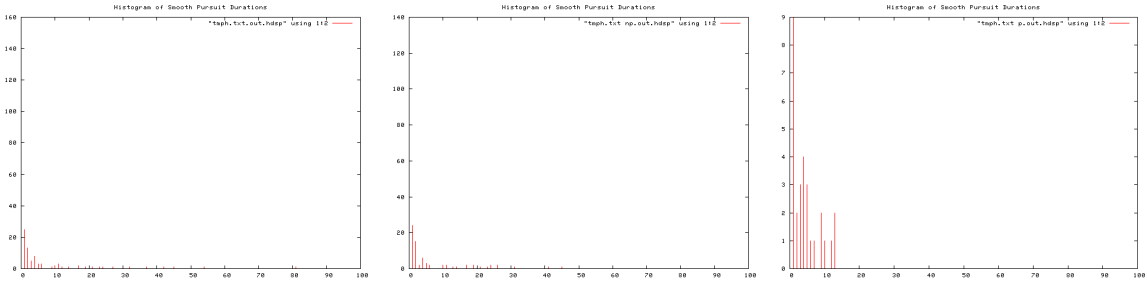



Figure B.61: Histogram of Smooth pursuit durations, Trial 2. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

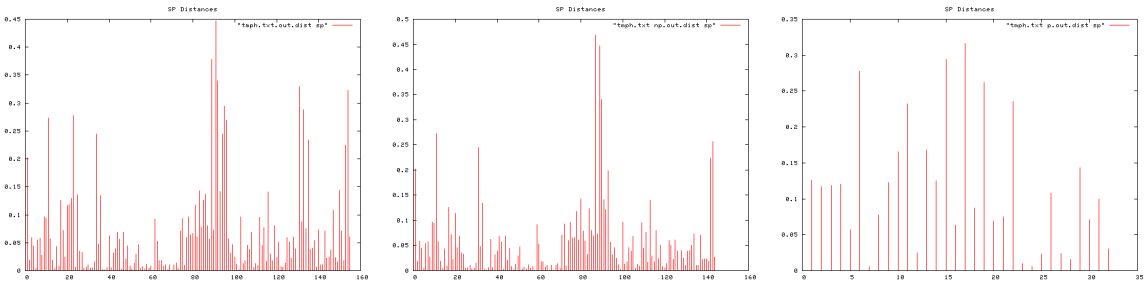



Figure B.62: Smooth pursuit distances, Trial 2. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

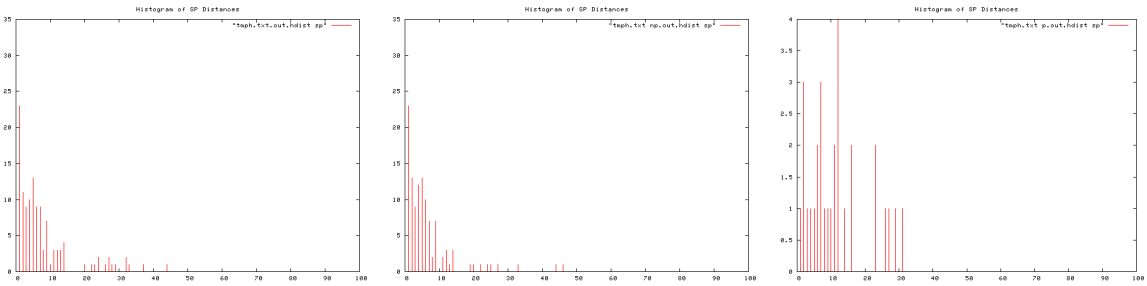



Figure B.63: Histogram of smooth pursuit distances, Trial 2. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

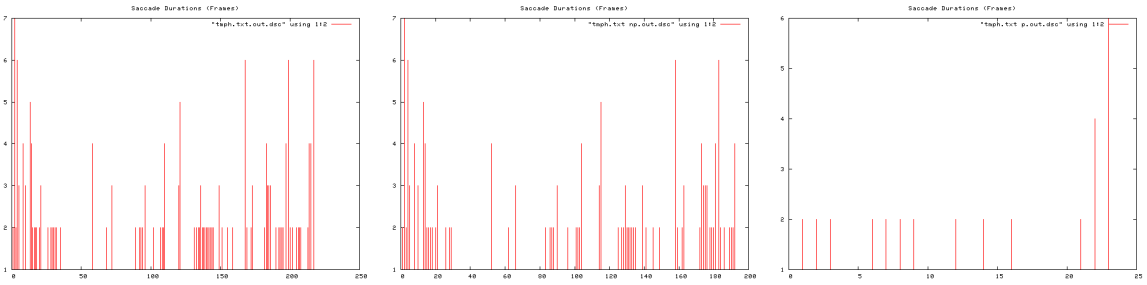



Figure B.64: Saccade durations, Trial 2. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

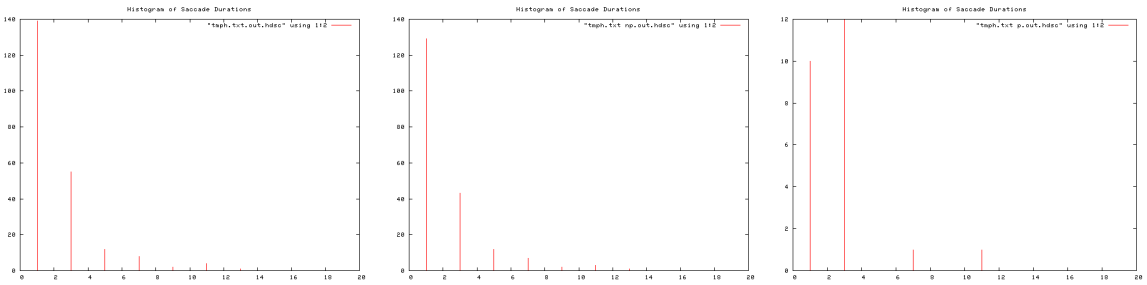



Figure B.65: Histogram of saccade durations, Trial 2. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

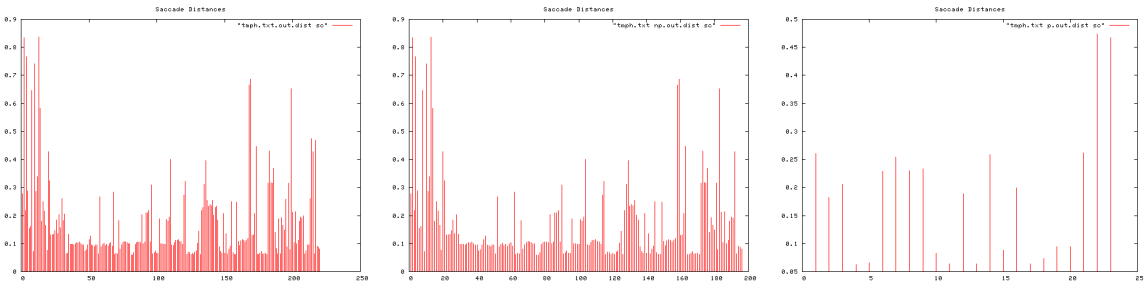



Figure B.66: Saccade distances, Trial 2. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

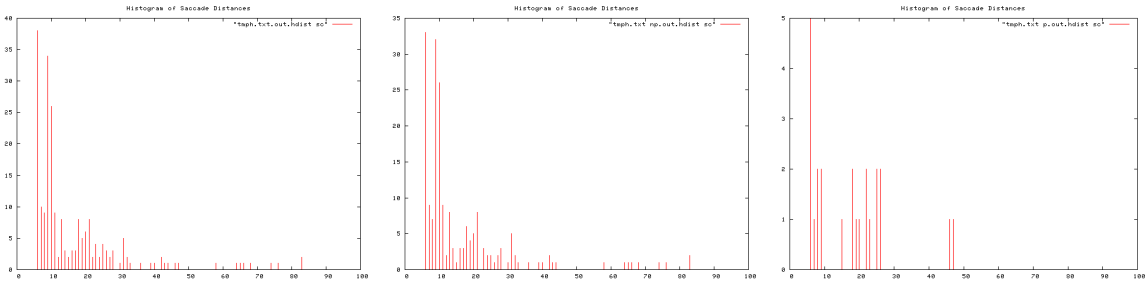



Figure B.67: Histogram of saccade distances, Trial 2. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

| Ti | To | T | I/O Event | Pear | Apple | Peach | Orange | |
|---|---|---|---|---|---|---|---|---|
| 11 | 13 | 2 | Orange In | | | | 0 | |
| 16 | 21 | 5 | Pear In | 1 | | | 1 | |
| 25 | 41 | 16 | | 4 | | | 5 | |
| 43 | 50 | 7 | Peach In | 0 | | 1 | 0 | |
| 54 | 0:58 | 4 | | 1 | | 0 | 1 | |
| 1:01 | 1:15 | 4 | | 2 | | 2 | 1 | |
| 1:18 | 1:27 | 9 | | 2 | | 3 | 1 | |
| 1:31 | 1:38 | 7 | Apple In, Peach Out | 0 | 0 | 0 | 0 | |
| 1:40 | 1:45 | 5 | Orange Out | 1 | 2 | | 1 | |
| 1:48 | 1:53 | 5 | Pear Out | 1 | 1 | | | |
| 1:57 | 2:05 | 8 | Apple Out | | 1 | | | |
| | | | TOTAL Rets | 12 | 4 | 6 | 10 | |
| | | | TOTAL T | 62 | 25 | 31 | 59 | **SD** |
| | | | **Av. Re-attention Period** | **5.2** | **6.3** | **5.2** | **5.9** | 0.54467115 |

Figure B.68: Re-attention period statistics, Trial 2.

### B.1.1.5 Trial 3

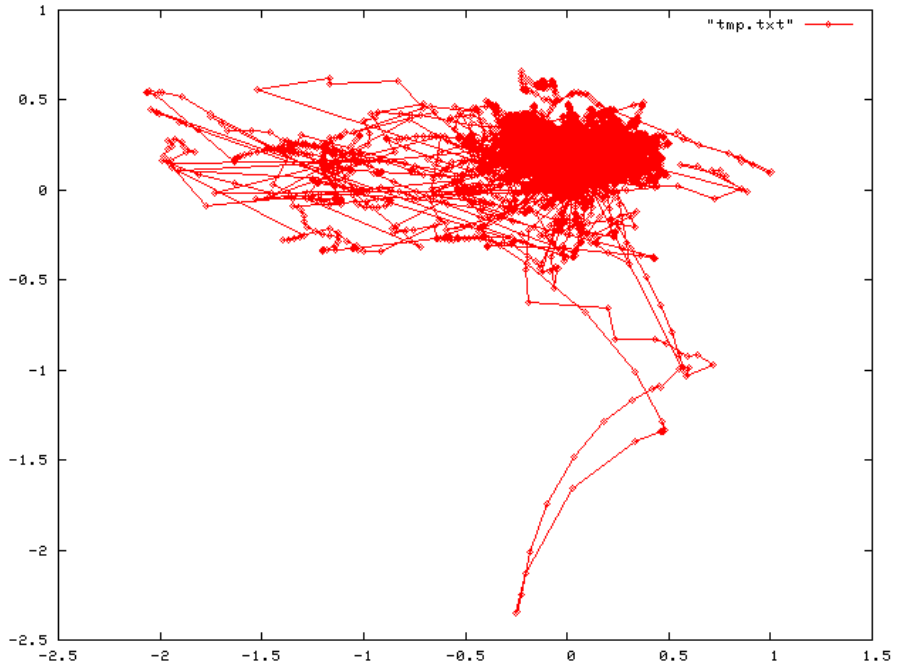



Figure B.69: Complete scan path, Trial 3.

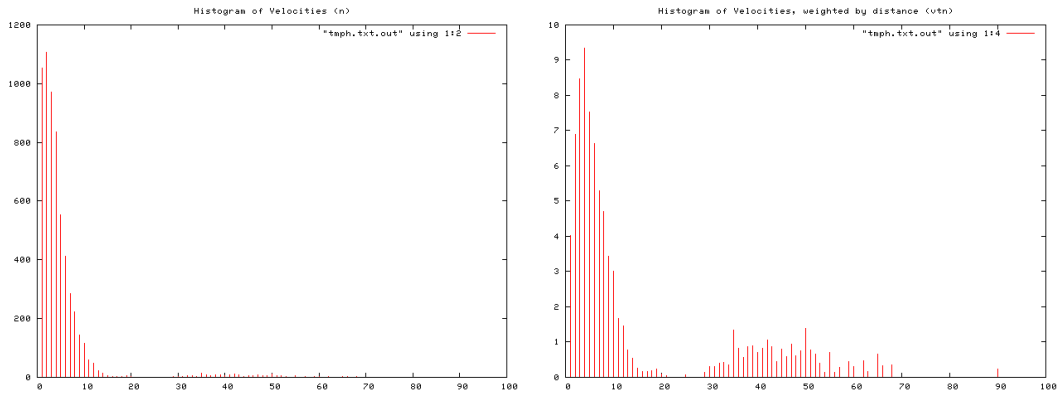



Figure B.70: Histogram of velocity magnitudes, Trial 3 (left). Histogram of distance weighted velocities, Trial 3 (right).

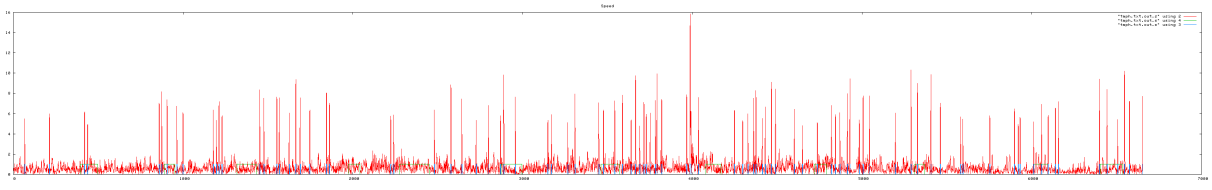


Figure B.71: Velocity profile. Velocity magnitude of each frame, Trial 3.

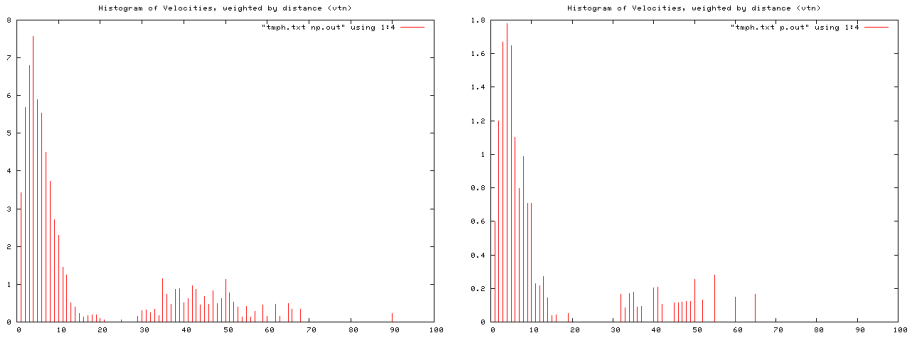



Figure B.72: Histogram of velocities during periods of no perturbation (left), and perturbation (right), Trial 3.

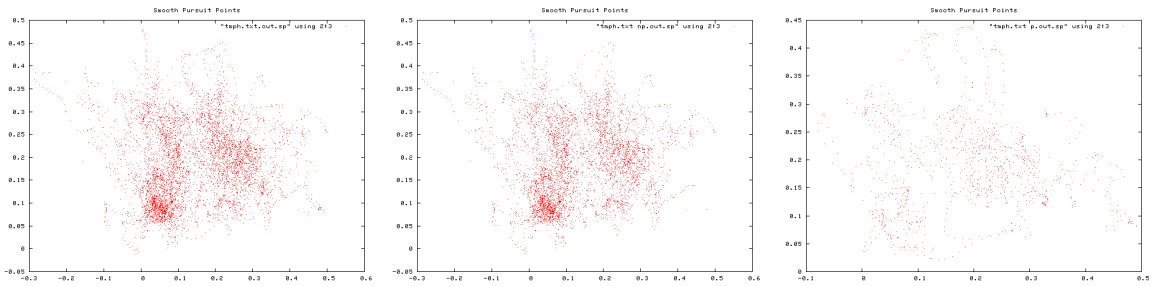



Figure B.73: Smooth pursuit gaze locations, Trial 3. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

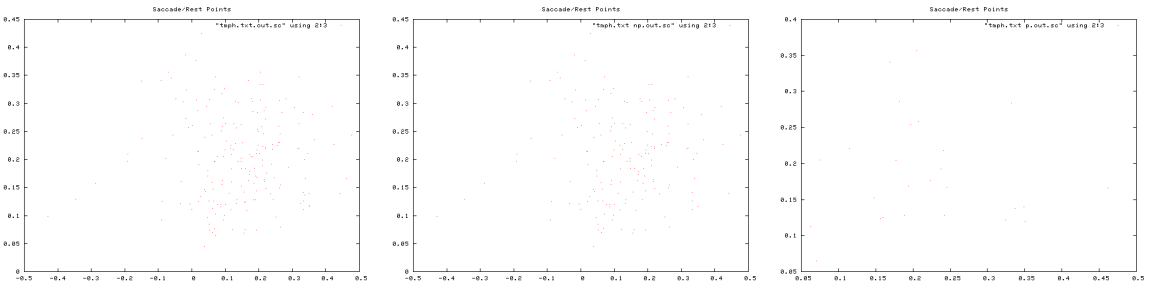



Figure B.74: Saccade gaze locations, Trial 3. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

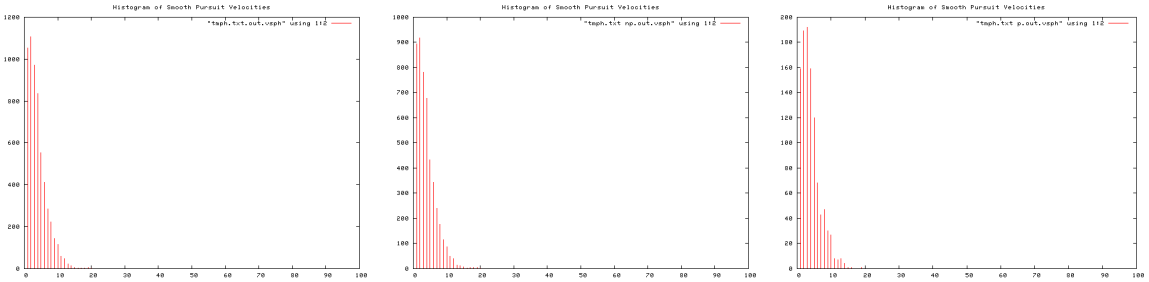



Figure B.75: Histogram of smooth pursuit velocities, Trial 3. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

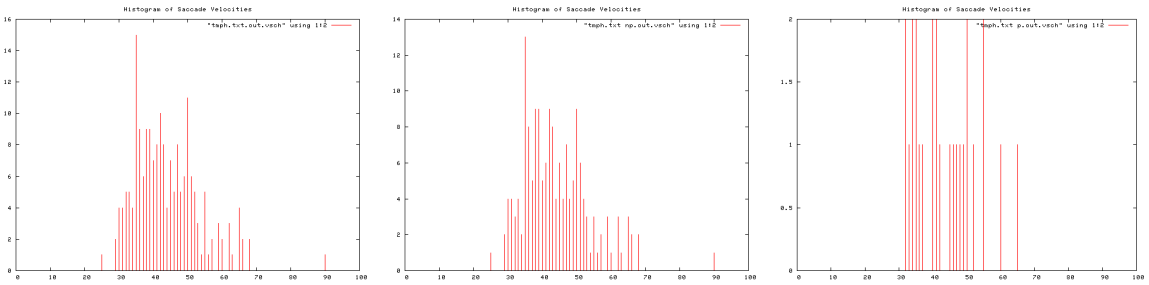



Figure B.76: Histogram of Saccade velocities, Trial 3. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

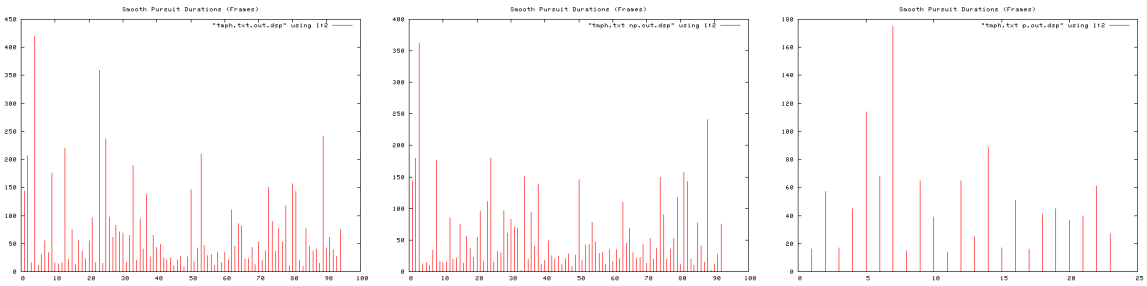



Figure B.77: Smooth pursuit durations, Trial 3. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

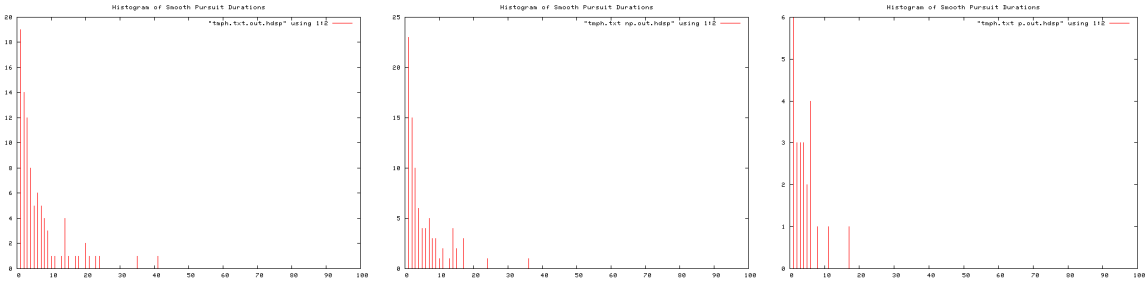



Figure B.78: Histogram of Smooth pursuit durations, Trial 3. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

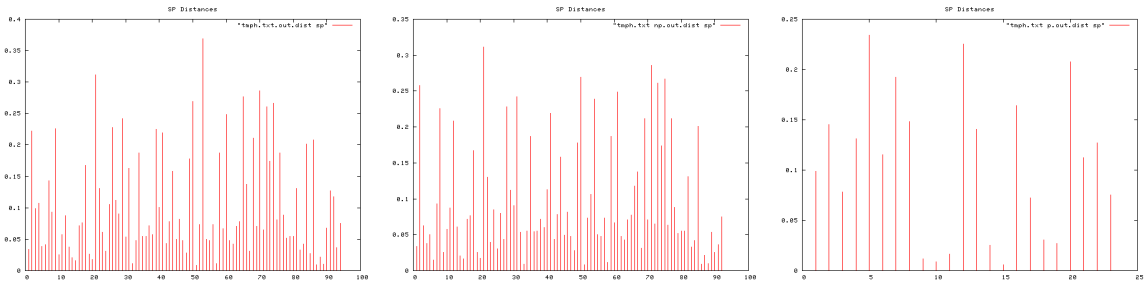



Figure B.79: Smooth pursuit distances, Trial 3. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

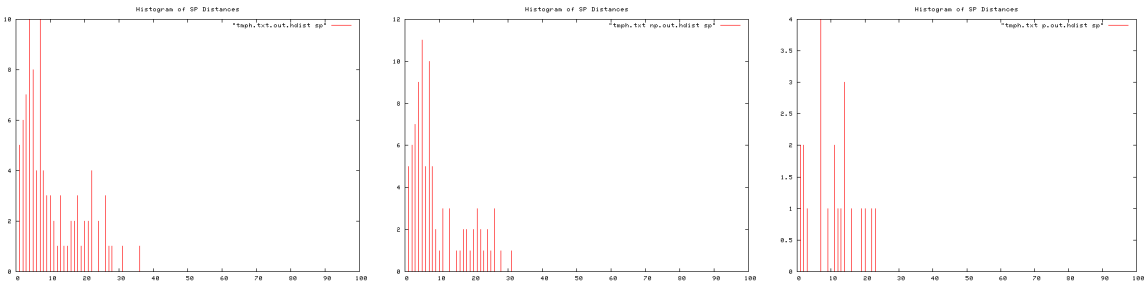



Figure B.80: Histogram of smooth pursuit distances, Trial 3. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

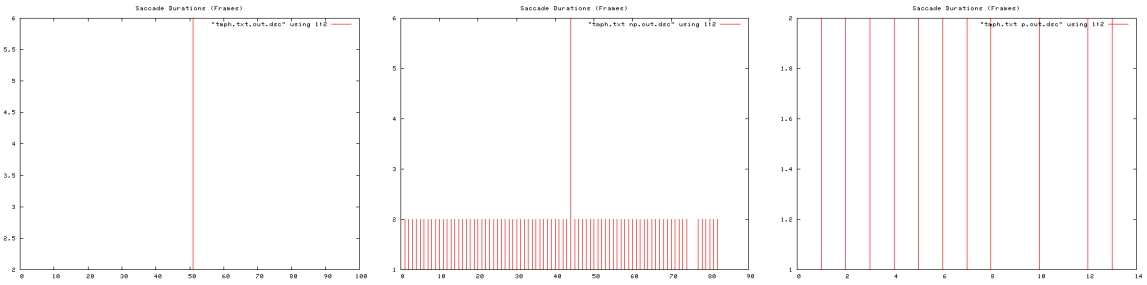



Figure B.81: Saccade durations, Trial 3. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

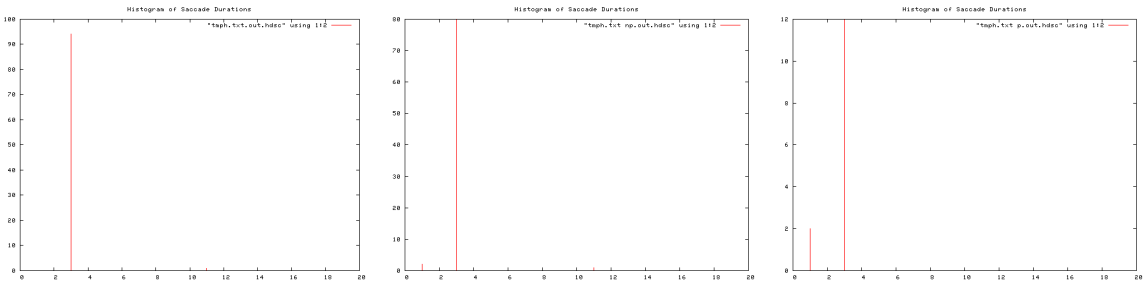



Figure B.82: Histogram of saccade durations, Trial 3. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

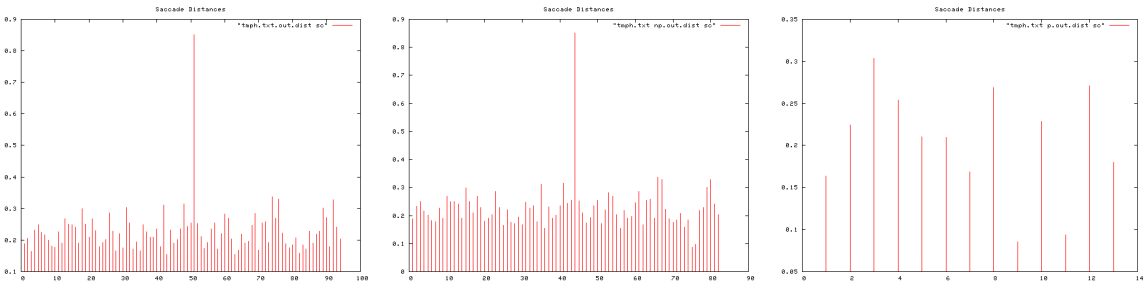



Figure B.83: Saccade distances, Trial 3. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

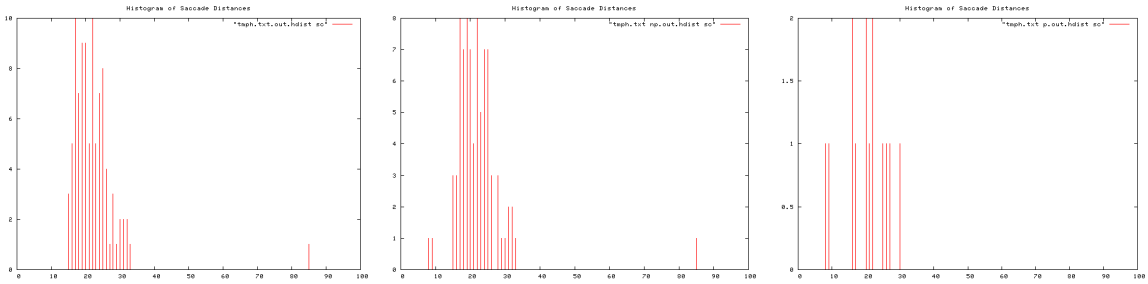



Figure B.84: Histogram of saccade distances, Trial 3. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

| Ti | To | T | I/O Event | Pear | Apple | Peach | Orange | |
|---|---|---|---|---|---|---|---|---|
| 8 | 14 | 6 | Orange In | | | | 1 | |
| 16 | 22 | 6 | Pear In | 2 | | | 2 | |
| 25 | 32 | 7 | | 3 | | | 2 | |
| 34 | 39 | 5 | Peach In | 2 | | 1 | 1 | |
| 41 | 0:48 | 7 | | 2 | | 2 | 3 | |
| 0:51 | 0:58 | 7 | | 2 | | 2 | 2 | |
| 1:00 | 1:08 | 8 | | 1 | | 4 | 3 | |
| 1:10 | 1:18 | 8 | Apple In, Peach Out | 2 | 2 | 2 | 2 | |
| 1:20 | 1:30 | 10 | Orange Out | 1 | 2 | | 4 | |
| 1:31 | 1:40 | 9 | Pear Out | 4 | 4 | | | |
| 1:42 | 1:47 | 5 | Apple Out | | 1 | | | |
| | | | TOTAL Rets | 19 | 9 | 11 | 20 | |
| | | | TOTAL T | 73 | 32 | 35 | 64 | **SD** |
| | | | **Av. Re-attention Period** | **3.8** | **3.5** | **3.2** | **3.2** | 0.28722813 |

Figure B.85: Re-attention period statistics, Trial 3.

### B.1.1.6 Trial 4

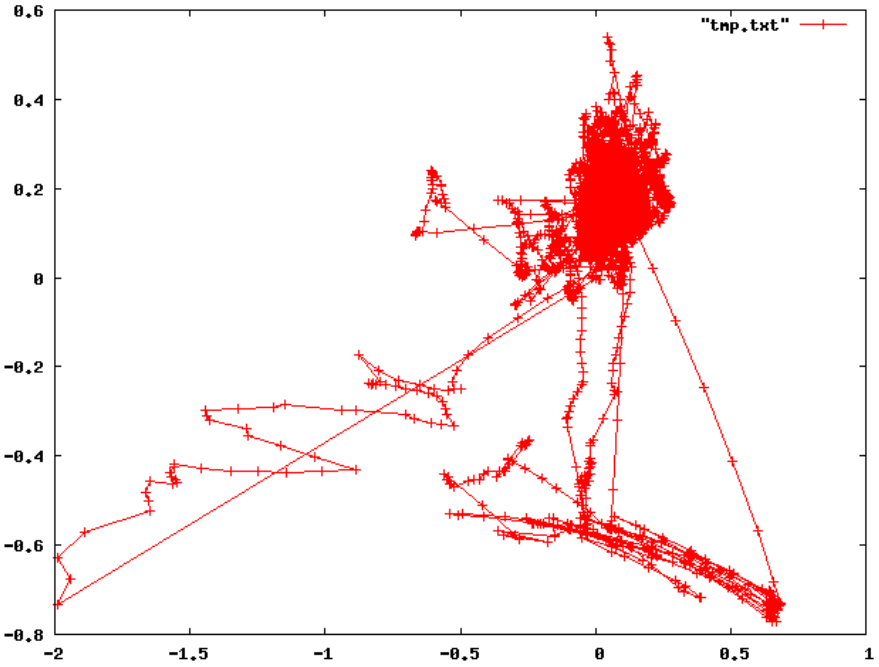



Figure B.86: Complete scan path, Trial 4.

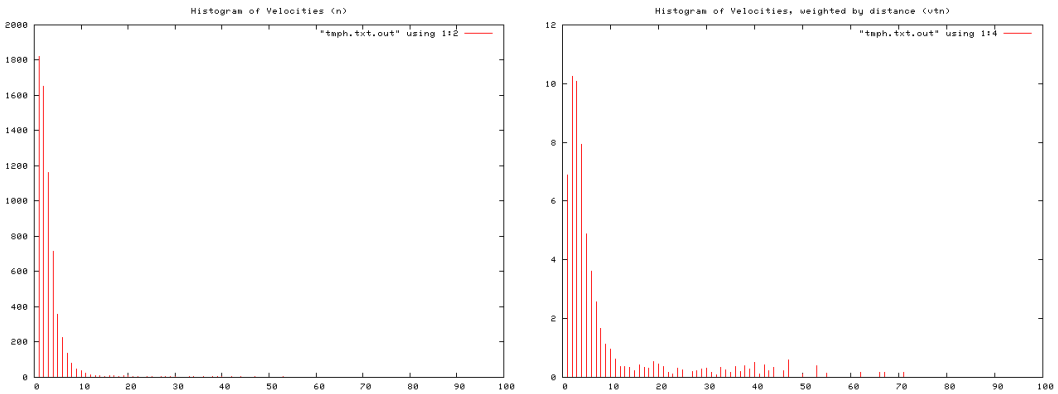



Figure B.87: Histogram of velocity magnitudes, Trial 4 (left). Histogram of distance weighted velocities, Trial 1 (right).

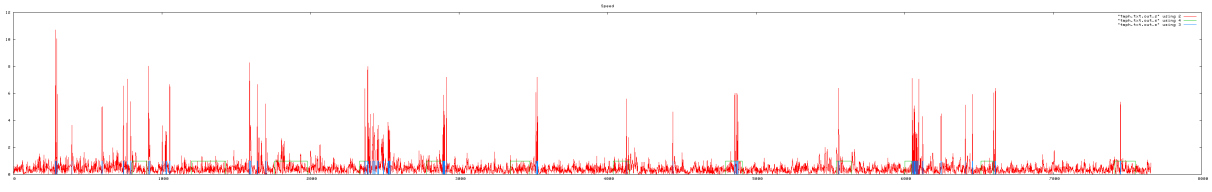


Figure B.88: Velocity profile. Velocity magnitude of each frame, Trial 4.

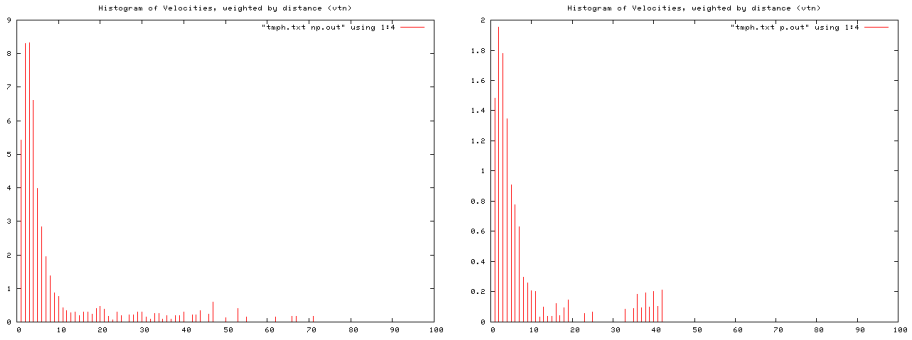



Figure B.89: Histogram of velocities during periods of no perturbation (left), and perturbation (right), Trial 4.

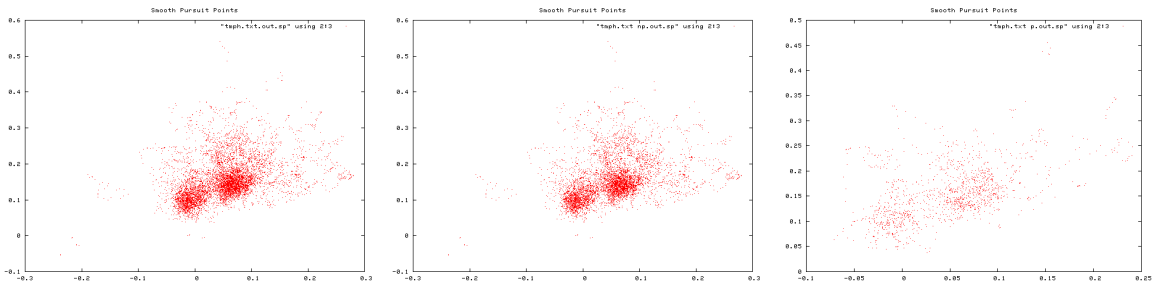



Figure B.90: Smooth pursuit gaze locations, Trial 4. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

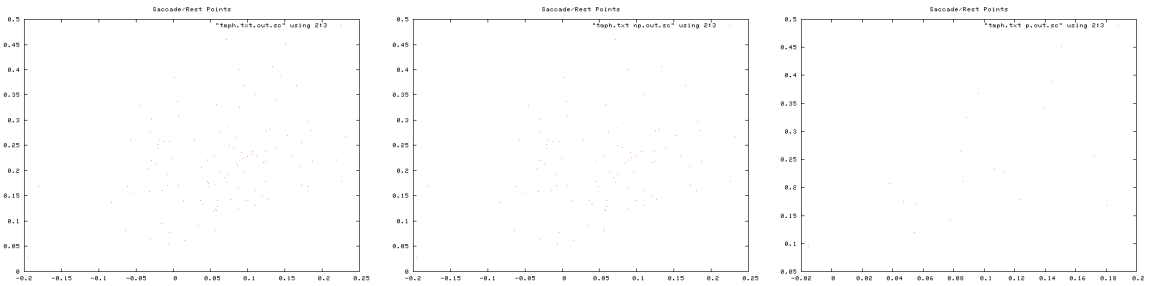



Figure B.91: Saccade gaze locations, Trial 4. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

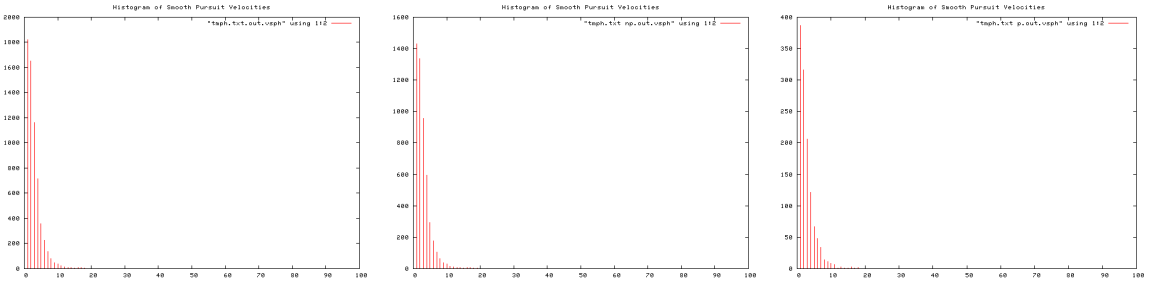



Figure B.92: Histogram of smooth pursuit velocities, Trial 4. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

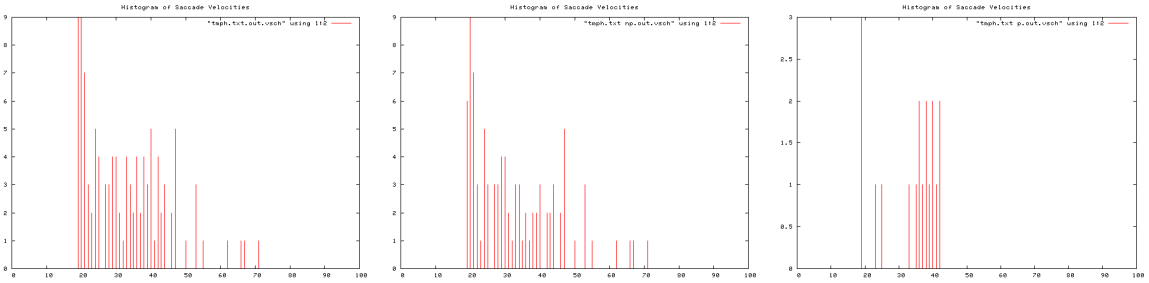



Figure B.93: Histogram of Saccade velocities, Trial 4. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

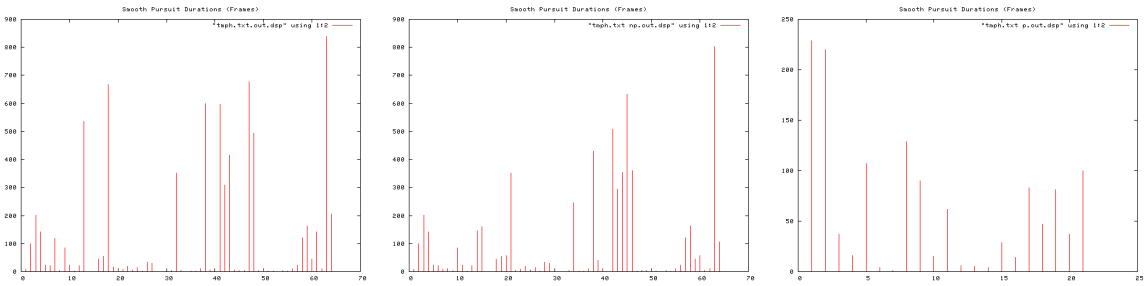



Figure B.94: Smooth pursuit durations, Trial 4. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

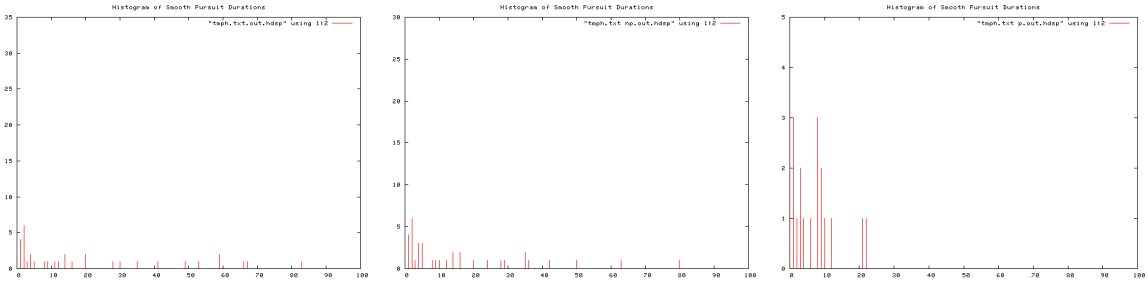



Figure B.95: Histogram of Smooth pursuit durations, Trial 4. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).


Figure B.96: Smooth pursuit distances, Trial 4. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

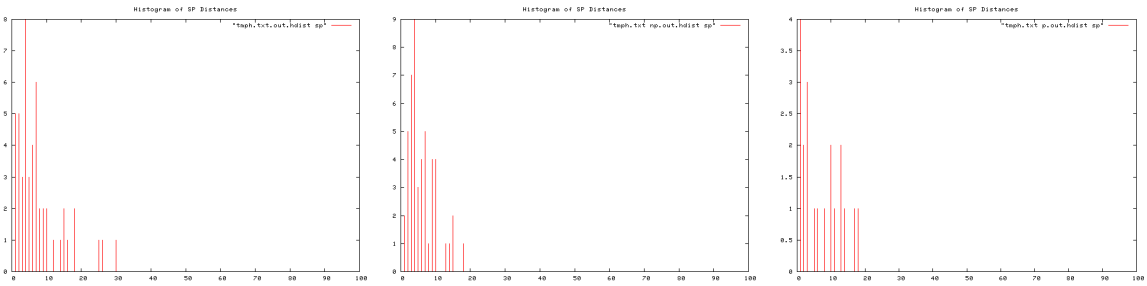



Figure B.97: Histogram of smooth pursuit distances, Trial 4. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

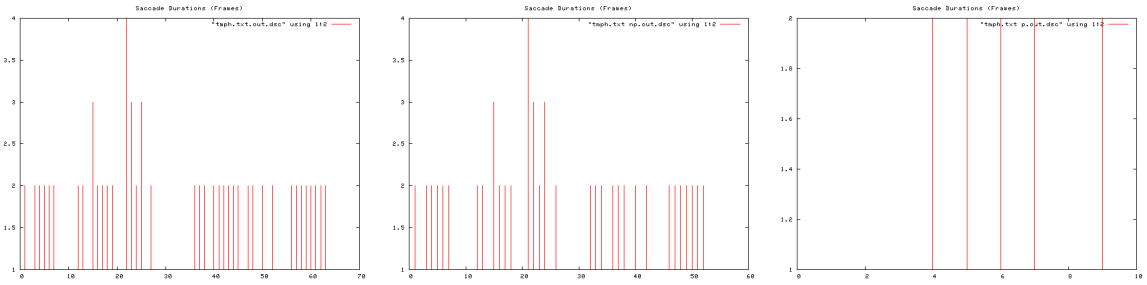



Figure B.98: Saccade durations, Trial 4. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

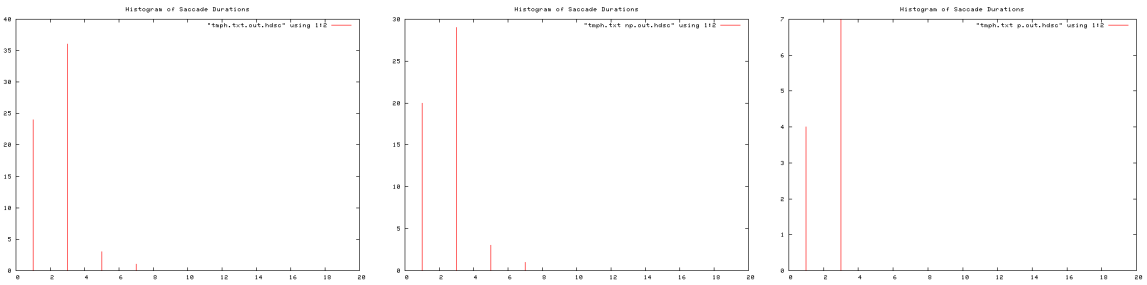



Figure B.99: Histogram of saccade durations, Trial 4. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

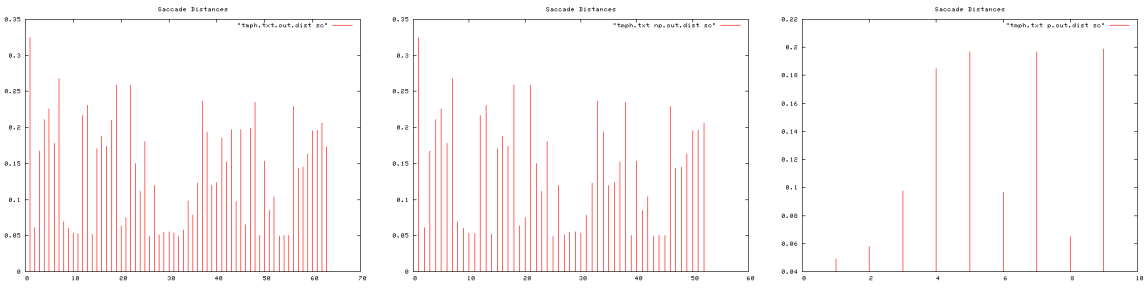



Figure B.100: Saccade distances, Trial 4. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

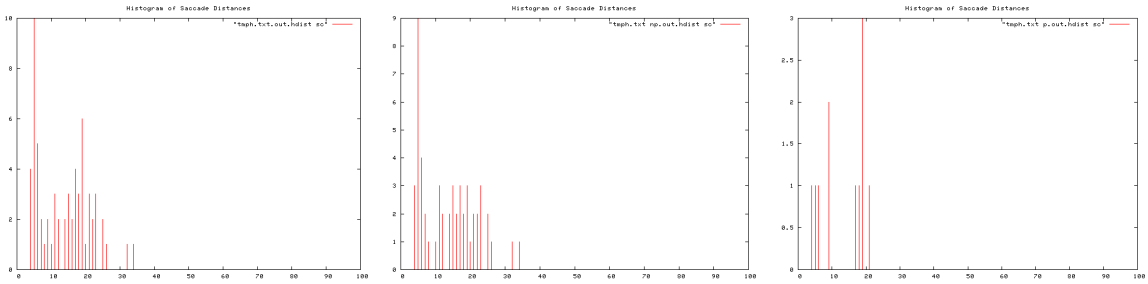



Figure B.101: Histogram of saccade distances, Trial 4. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

| Ti | To | T | I/O Event | Pear | Apple | Peach | Orange | |
|---|---|---|---|---|---|---|---|---|
| 15 | 22 | 7 | Orange In | | | | 1 | |
| 24 | 29 | 5 | Pear In | 1 | | | 1 | |
| 31 | 39 | 8 | | 1 | | | 1 | |
| 40 | 47 | 7 | Peach In | 1 | | 1 | 1 | |
| 49 | 0:56 | 7 | | 1 | | 1 | 1 | |
| 0:58 | 1:08 | 10 | | 0 | | 1 | 1 | |
| 1:10 | 1:21 | 11 | | 1 | | 2 | 1 | |
| 1:22 | 1:30 | 8 | Apple In, Peach Out | 1 | 1 | 1 | 2 | |
| 1:31 | 1:40 | 9 | Orange Out | 1 | 1 | | 0 | |
| 1:41 | 1:48 | 7 | Pear Out | 3 | 2 | | | |
| 1:50 | 2:02 | 12 | Apple Out | | 1 | | | |
| | | | TOTAL Rets | 10 | 5 | 6 | 9 | |
| | | | TOTAL T | 72 | 36 | 43 | 72 | **SD** |
| | | | **Av. Re-attention Period** | **7.2** | **7.2** | **7.2** | **8** | 0.4 |

Figure B.102: Re-attention period statistics, Trial 4.

### B.1.1.7 Trial 5

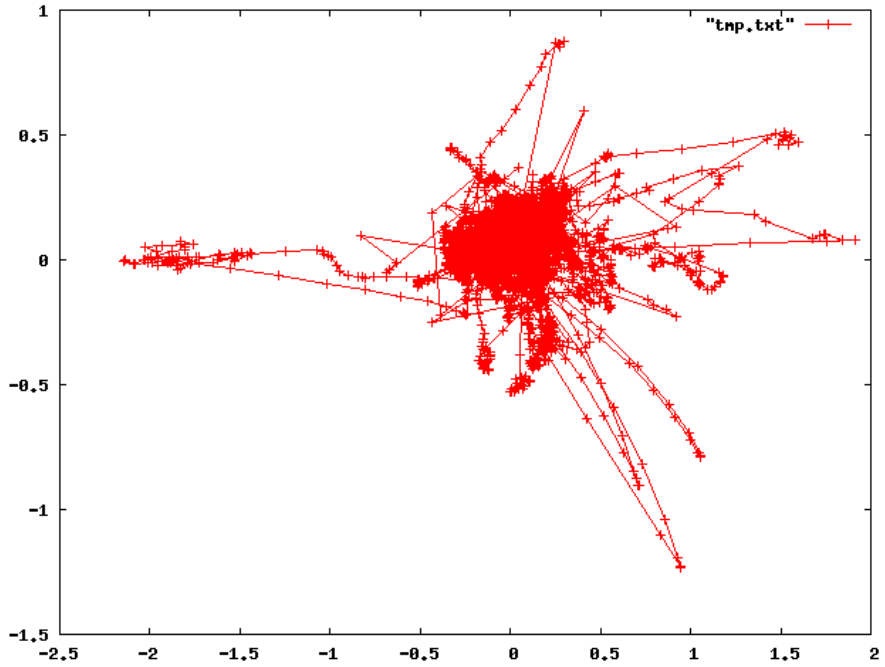



Figure B.103: Complete scan path, Trial 5.

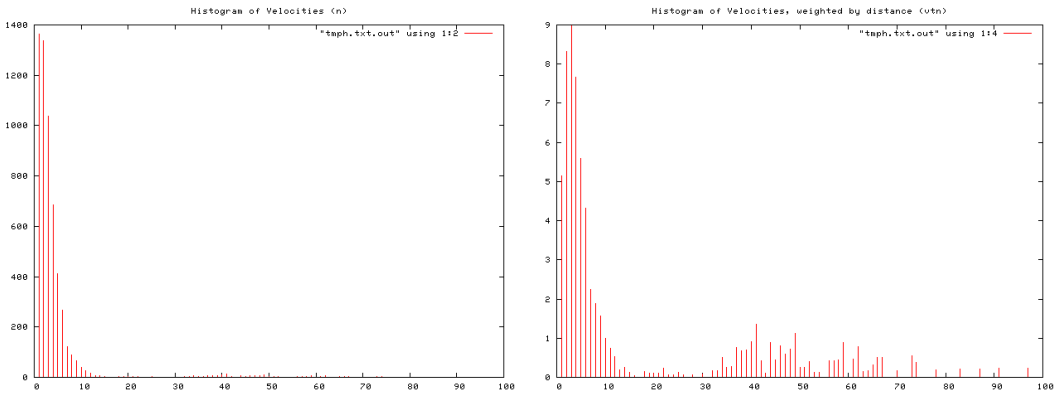



Figure B.104: Histogram of velocity magnitudes, Trial 5 (left). Histogram of distance weighted velocities, Trial 5 (right).

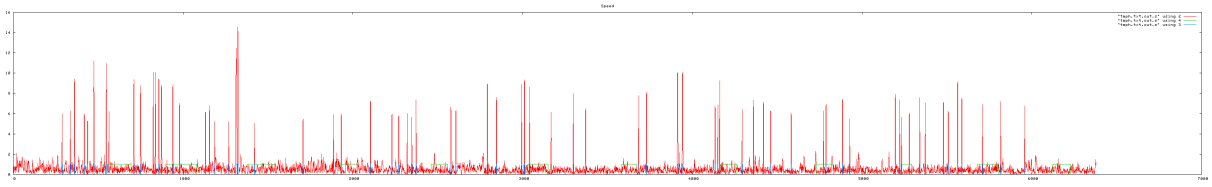


Figure B.105: Velocity profile. Velocity magnitude of each frame, Trial 5.

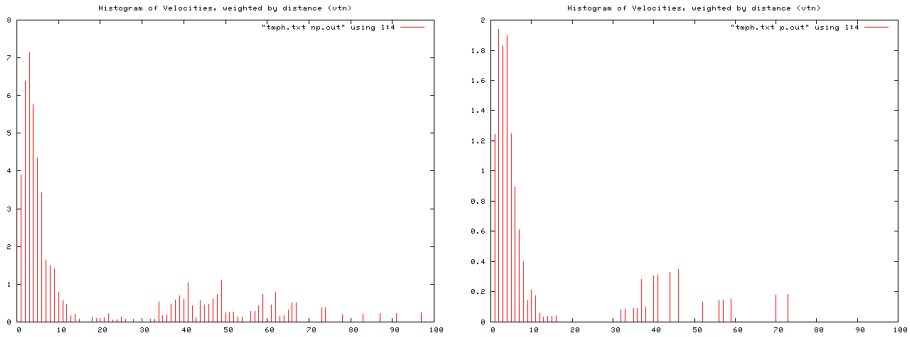



Figure B.106: Histogram of velocities during periods of no perturbation (left), and perturbation (right), Trial 5.

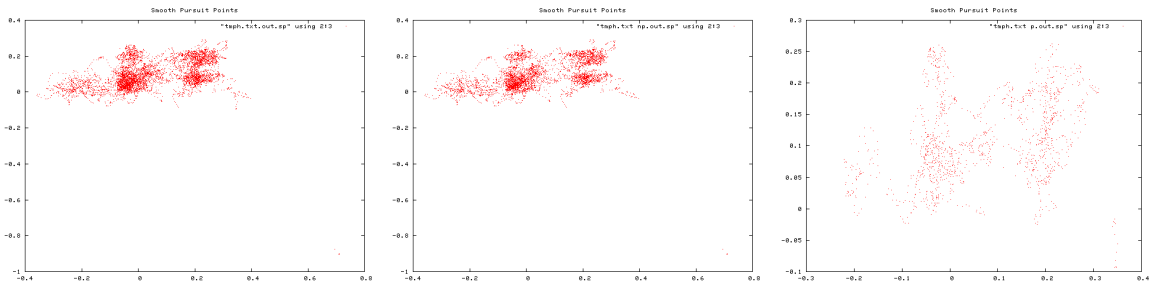



Figure B.107: Smooth pursuit gaze locations, Trial 5. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

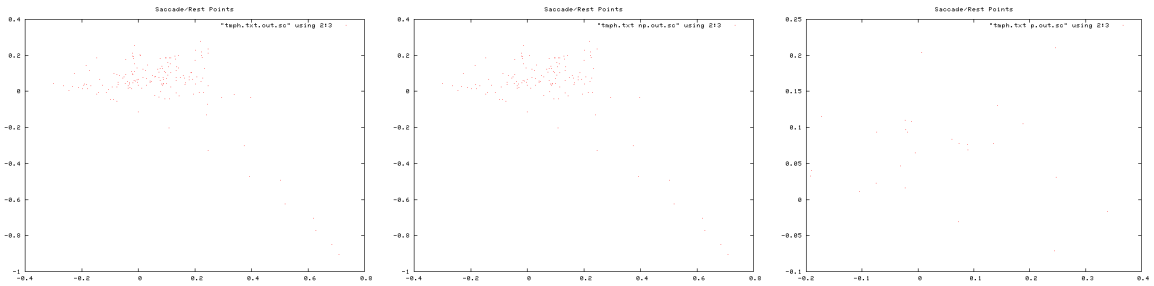



Figure B.108: Saccade gaze locations, Trial 5. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

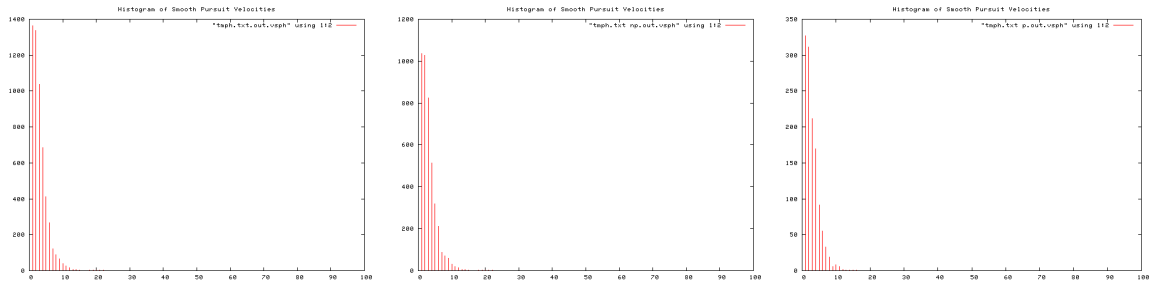



Figure B.109: Histogram of smooth pursuit velocities, Trial 5. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

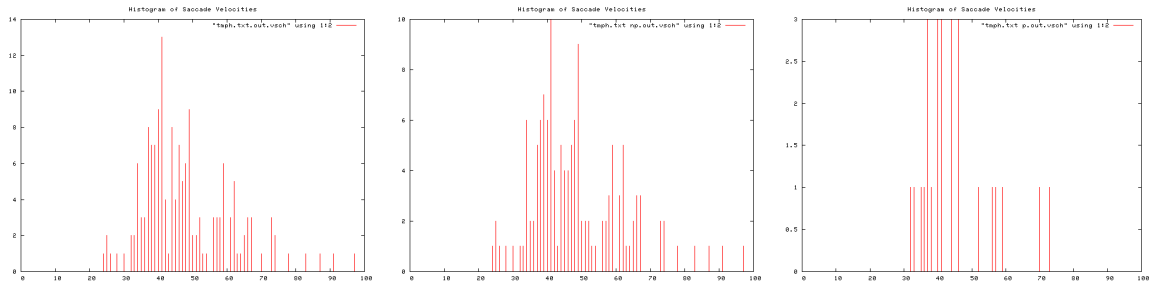



Figure B.110: Histogram of Saccade velocities, Trial 5. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

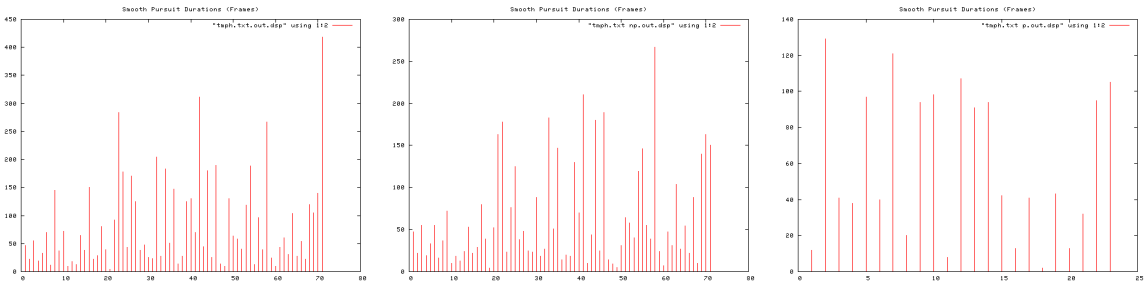



Figure B.111: Smooth pursuit durations, Trial 5. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

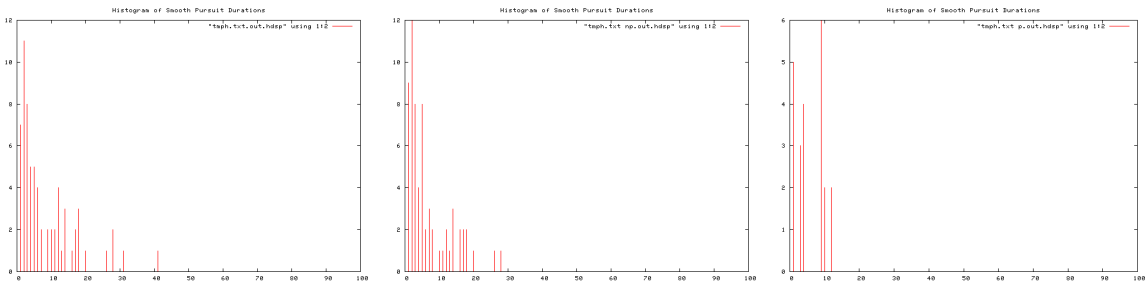



Figure B.112: Histogram of Smooth pursuit durations, Trial 5. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

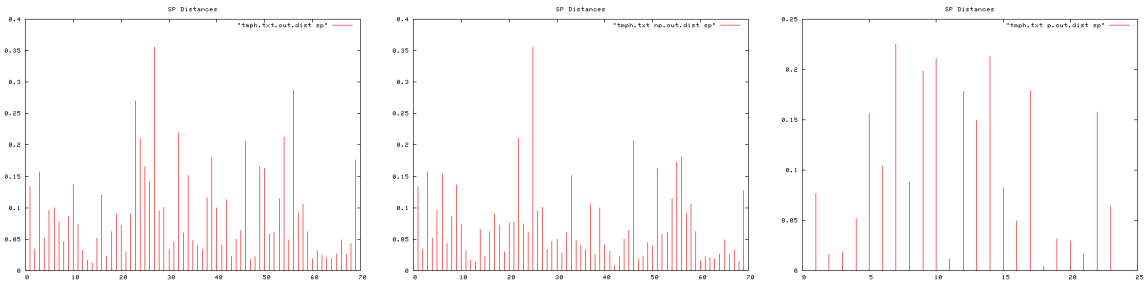



Figure B.113: Smooth pursuit distances, Trial 5. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

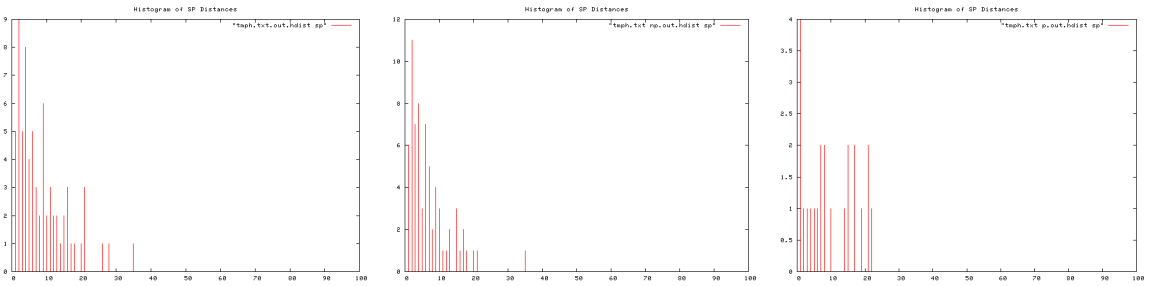



Figure B.114: Histogram of smooth pursuit distances, Trial 5. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

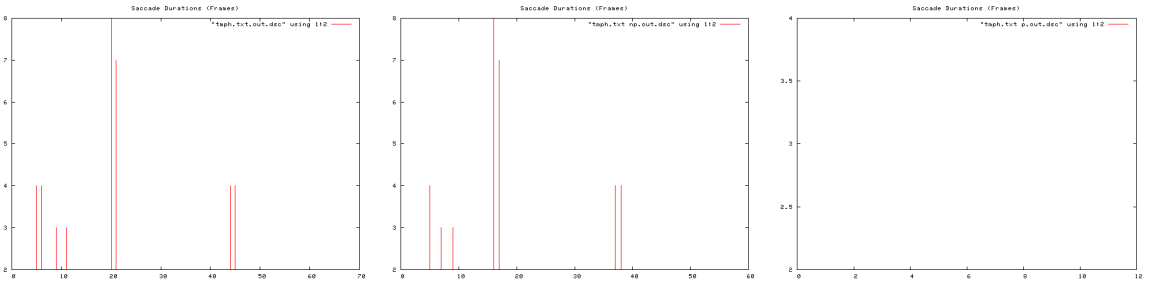



Figure B.115: Saccade durations, Trial 5. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

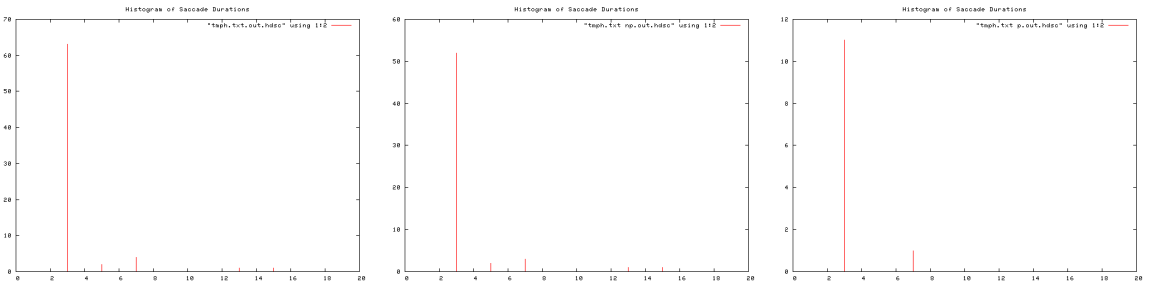



Figure B.116: Histogram of saccade durations, Trial 5. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

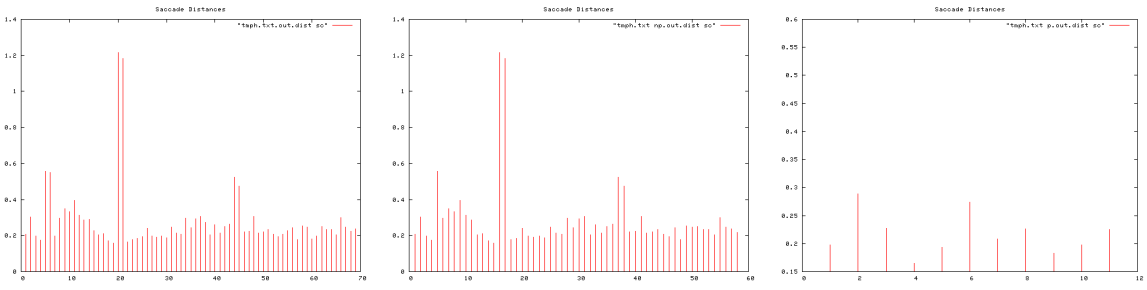



Figure B.117: Saccade distances, Trial 5. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

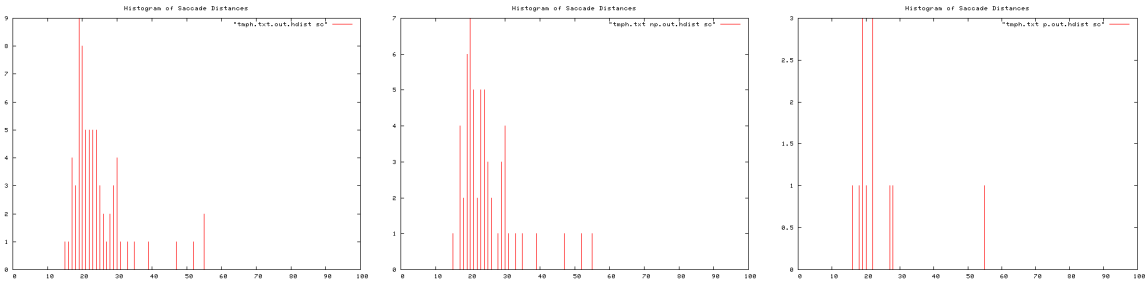



Figure B.118: Histogram of saccade distances, Trial 5. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

| Ti | To | T | I/O Event | Pear | Apple | Peach | Orange | |
|---|---|---|---|---|---|---|---|---|
| 8 | 13 | 5 | Orange In | | | | 3 | |
| 15 | 20 | 5 | Pear In | 2 | | | 2 | |
| 23 | 30 | 7 | | 3 | | | 2 | |
| 31 | 39 | 8 | Peach In | 3 | | 3 | 3 | |
| 41 | 0:49 | 8 | | 2 | | 2 | 2 | |
| 0:51 | 0:58 | 7 | | 1 | | 1 | 1 | |
| 1:00 | 1:08 | 8 | | 2 | | 2 | 2 | |
| 1:09 | 1:16 | 7 | Apple In, Peach Out | 2 | 4 | 2 | 2 | |
| 1:19 | 1:25 | 6 | Orange Out | 2 | 3 | | 3 | |
| 1:26 | 1:34 | 8 | Pear Out | 4 | 4 | | | |
| 1:35 | 2:05 | 10 | Apple Out | | 2 | | | |
| | | | TOTAL Rets | 21 | 13 | 10 | 20 | |
| | | | TOTAL T | 64 | 31 | 38 | 61 | **SD** |
| | | | **Av. Re-attention Period** | **3** | **2.4** | **3.8** | **3.1** | 0.57373048 |

Figure B.119: Re-attention period statistics, Trial 5.

### B.1.1.8 Trial 6

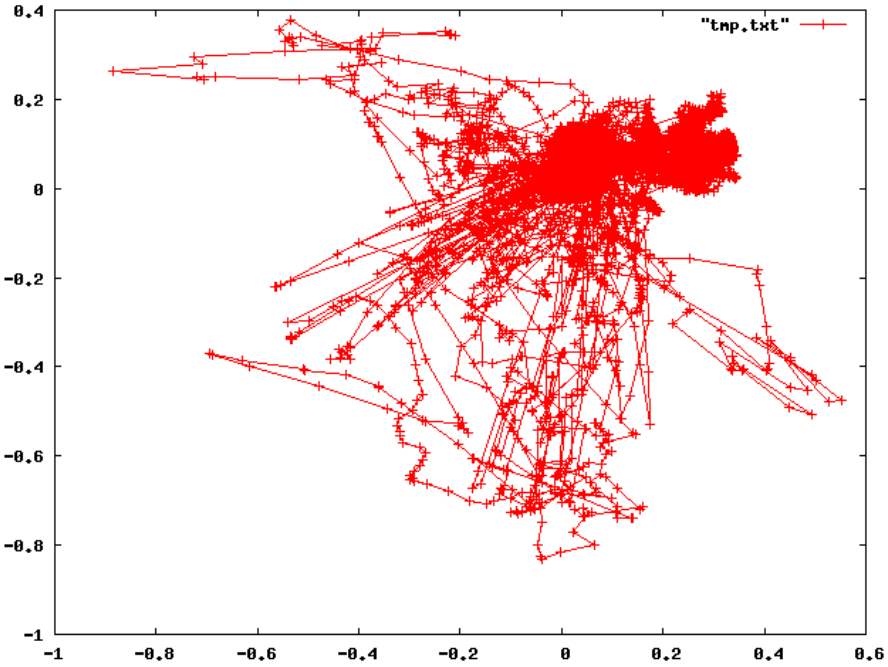



Figure B.120: Complete scan path, Trial 6.

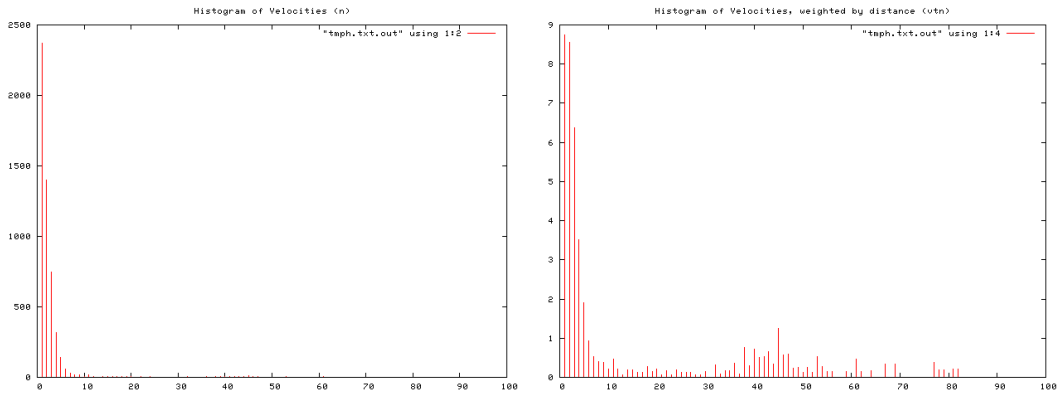



Figure B.121: Histogram of velocity magnitudes, Trial 6 (left). Histogram of distance weighted velocities, Trial 5 (right).

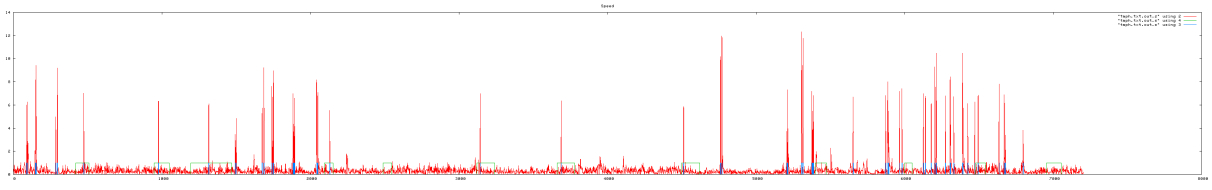


Figure B.122: Velocity profile. Velocity magnitude of each frame, Trial 6.

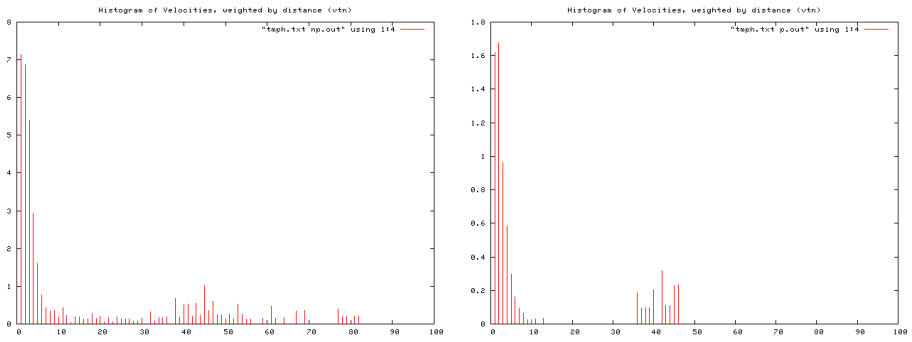



Figure B.123: Histogram of velocities during periods of no perturbation (left), and perturbation (right), Trial 6.

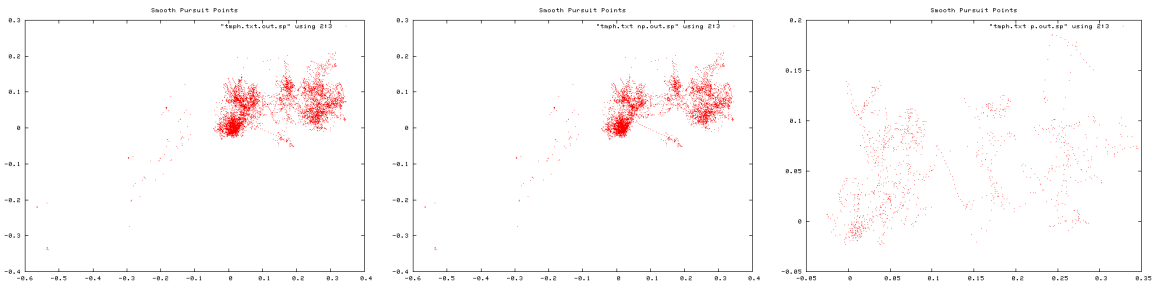



Figure B.124: Smooth pursuit gaze locations, Trial 6. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

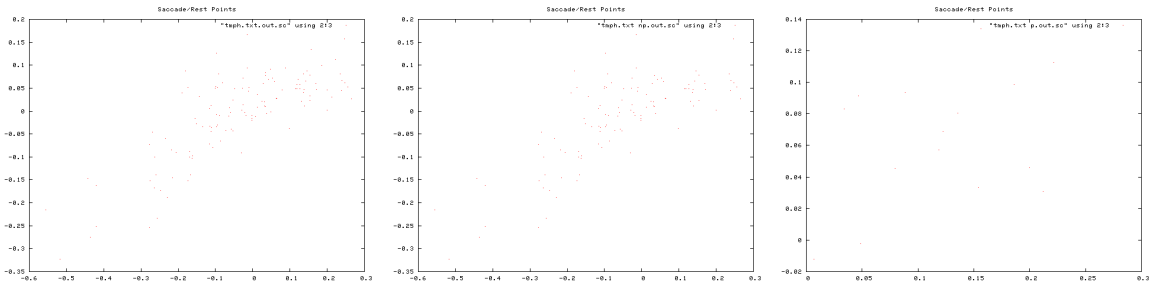



Figure B.125: Saccade gaze locations, Trial 6. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

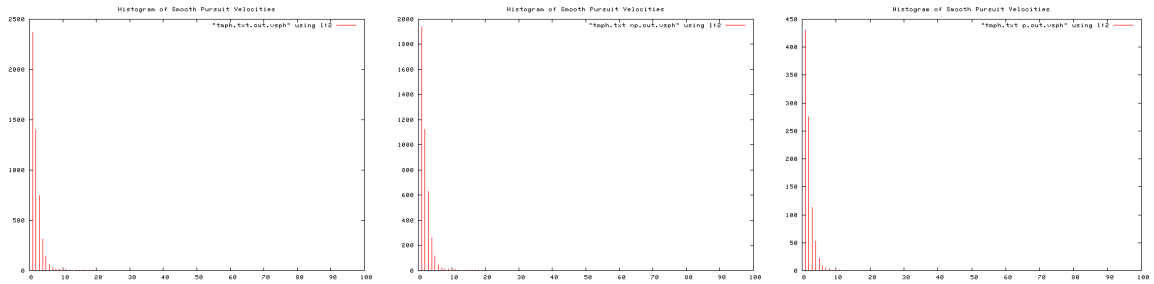



Figure B.126: Histogram of smooth pursuit velocities, Trial 6. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

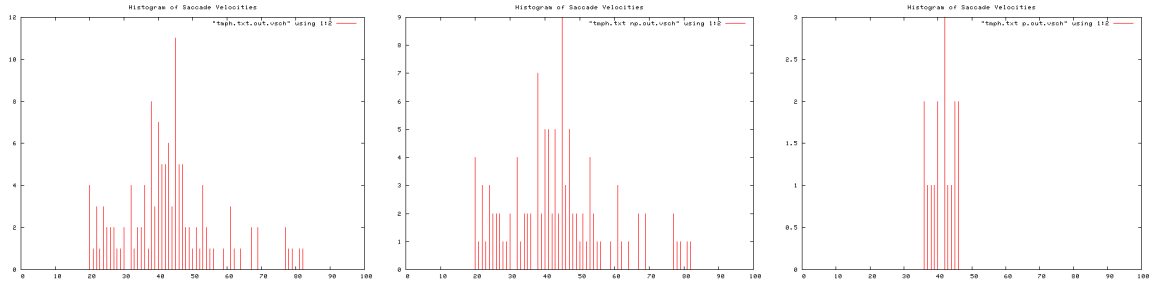



Figure B.127: Histogram of Saccade velocities, Trial 6. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

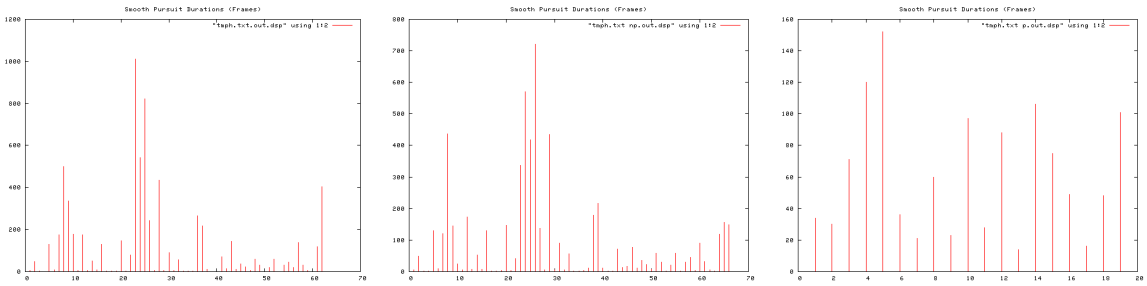



Figure B.128: Smooth pursuit durations, Trial 6. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

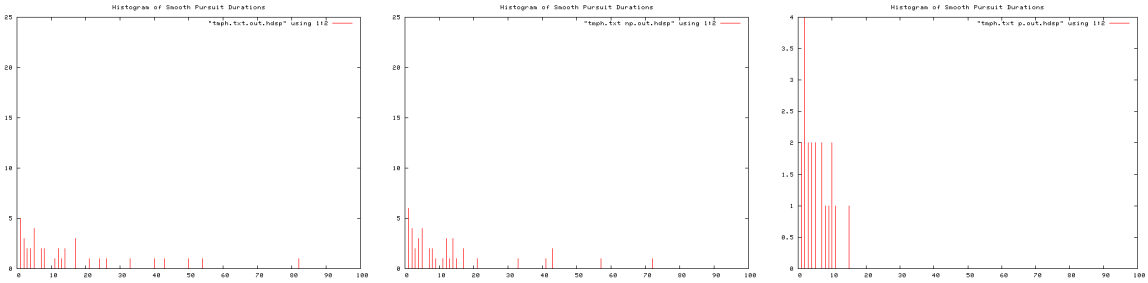



Figure B.129: Histogram of Smooth pursuit durations, Trial 6. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

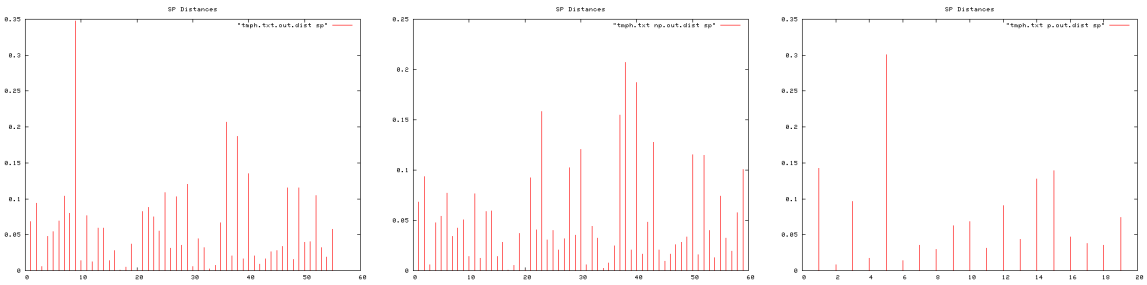



Figure B.130: Smooth pursuit distances, Trial 6. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

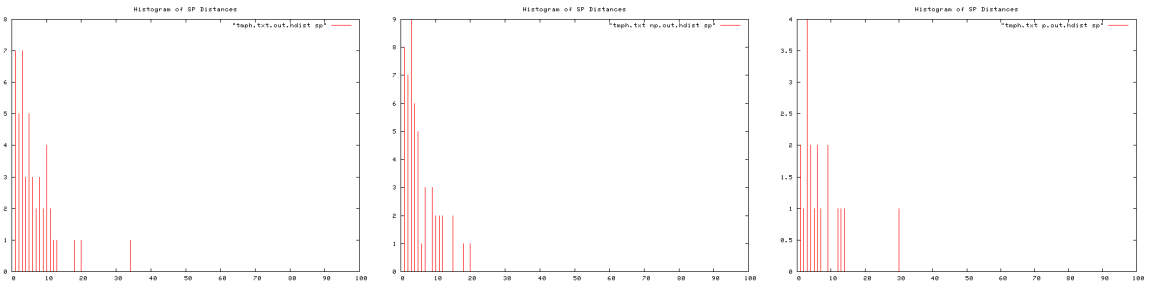



Figure B.131: Histogram of smooth pursuit distances, Trial 6. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

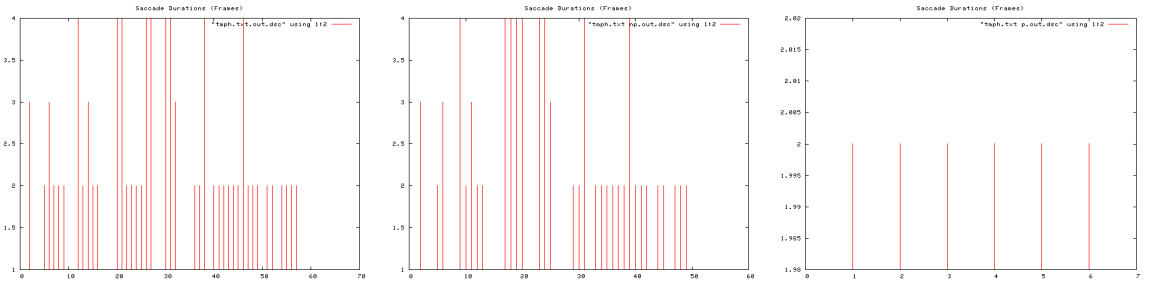



Figure B.132: Saccade durations, Trial 6. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

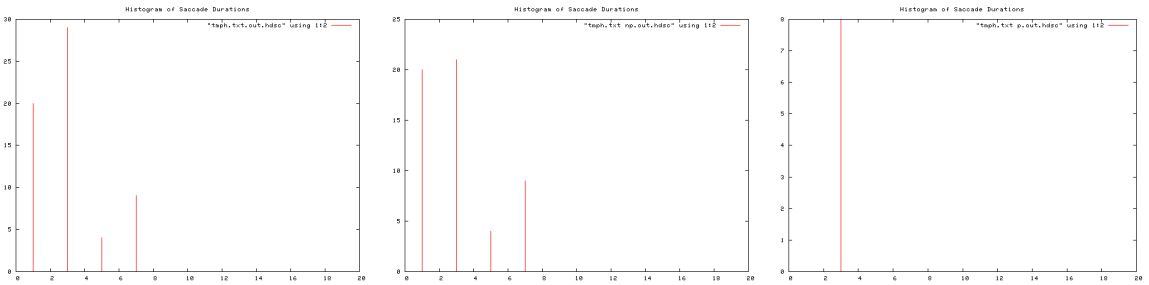



Figure B.133: Histogram of saccade durations, Trial 6. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

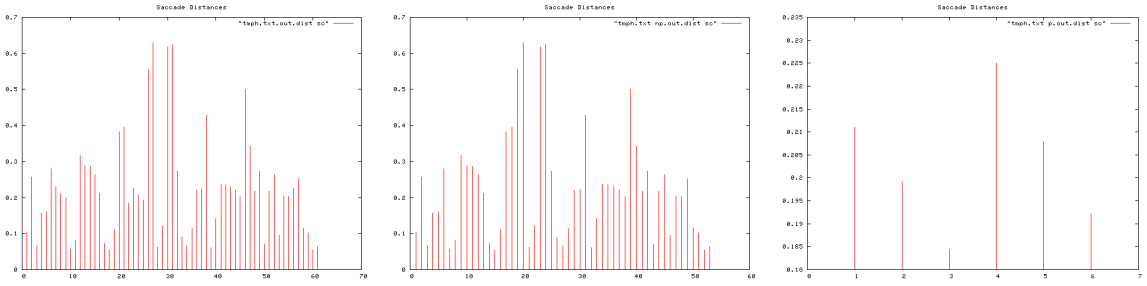



Figure B.134: Saccade distances, Trial 6. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

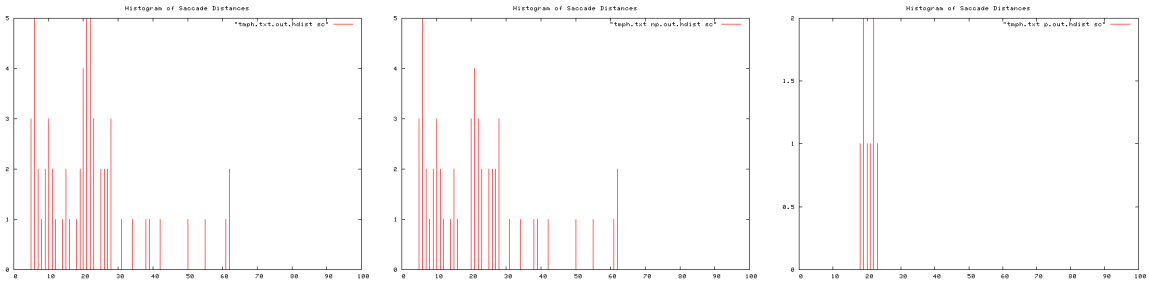



Figure B.135: Histogram of saccade distances, Trial 6. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

| Ti | To | T | I/O Event | Pear | Apple | Peach | Orange | |
|---|---|---|---|---|---|---|---|---|
| 8 | 16 | 8 | Orange In | | | | 2 | |
| 17 | 21 | 4 | Pear In | 1 | | | 1 | |
| 24 | 35 | 11 | | 1 | | | 2 | |
| 36 | 42 | 6 | Peach In | 1 | | 1 | 0 | |
| 43 | 0:52 | 9 | | 0 | | 2 | 1 | |
| 0:54 | 1:01 | 7 | | 1 | | 0 | 0 | |
| 1:03 | 1:14 | 11 | | 1 | | 2 | 2 | |
| 1:17 | 1:29 | 12 | Apple In, Peach Out | 0 | 0 | 2 | 1 | |
| 1:32 | 1:40 | 8 | Orange Out | 3 | 2 | | 2 | |
| 1:41 | 1:48 | 7 | Pear Out | 4 | 3 | | | |
| 1:50 | 1:58 | 8 | Apple Out | | 1 | | | |
| | | | TOTAL Rets | 12 | 6 | 7 | 11 | |
| | | | TOTAL T | 75 | 35 | 45 | 76 | **SD** |
| | | | **Av. Re-attention Period** | **6.3** | **5.8** | **6.4** | **6.9** | 0.45092498 |

Figure B.136: Re-attention period statistics, Trial 6.

### B.1.1.9 Trial 7

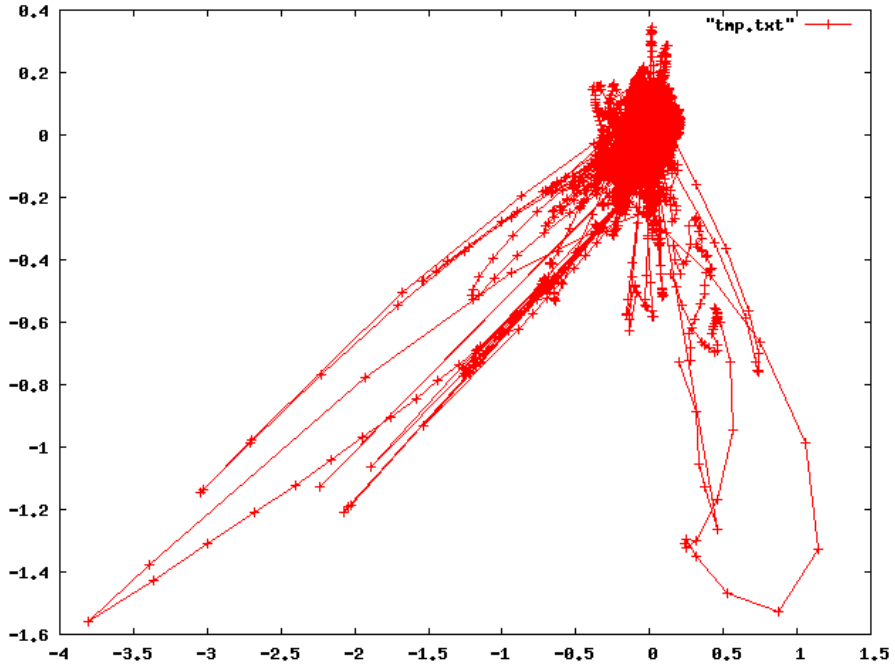



Figure B.137: Complete scan path, Trial 7.

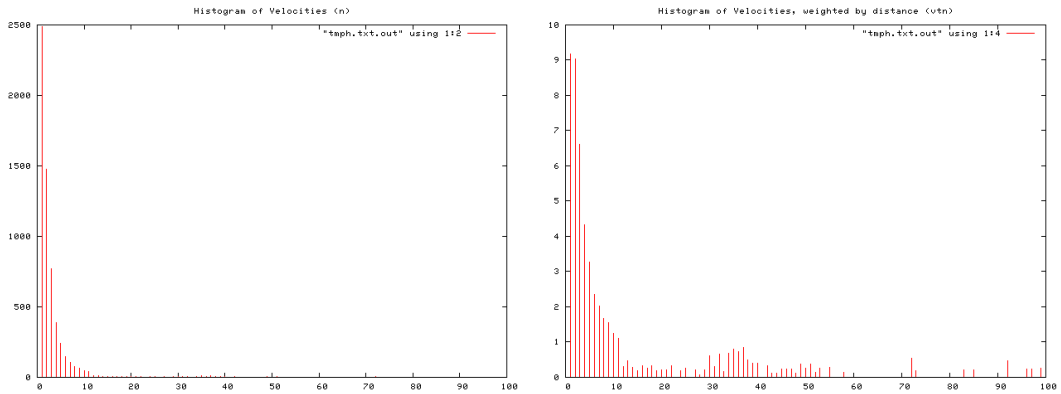



Figure B.138: Histogram of velocity magnitudes, Trial 7 (left). Histogram of distance weighted velocities, Trial 7 (right).

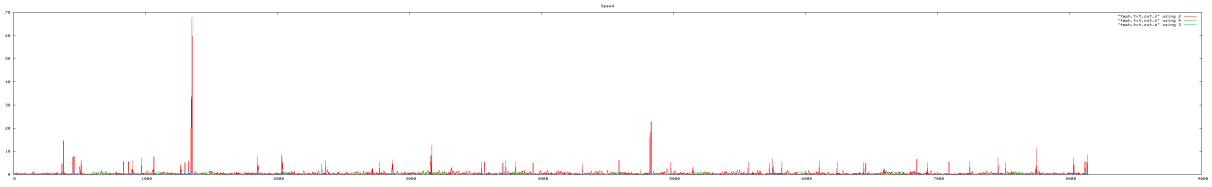


Figure B.139: Velocity profile. Velocity magnitude of each frame, Trial 7.

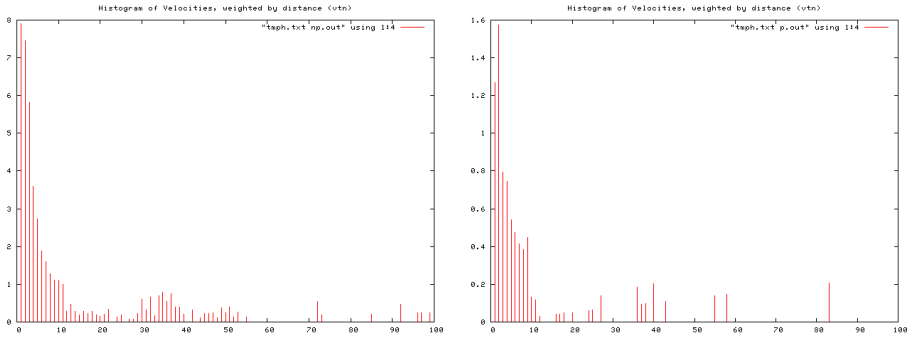



Figure B.140: Histogram of velocities during periods of no perturbation (left), and perturbation (right), Trial 7.

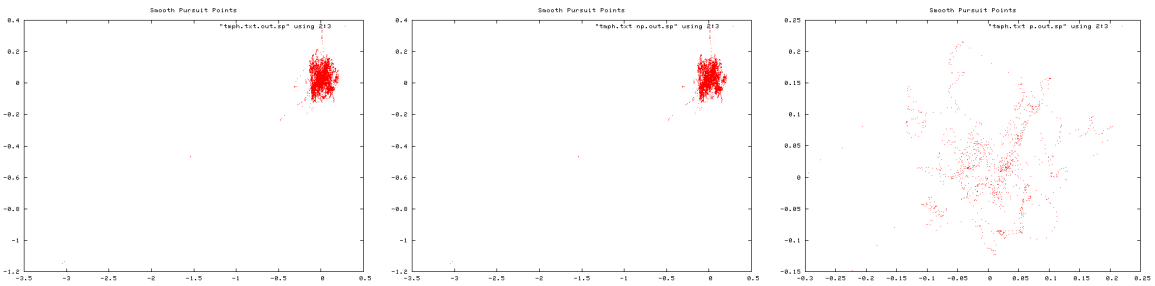



Figure B.141: Smooth pursuit gaze locations, Trial 7. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

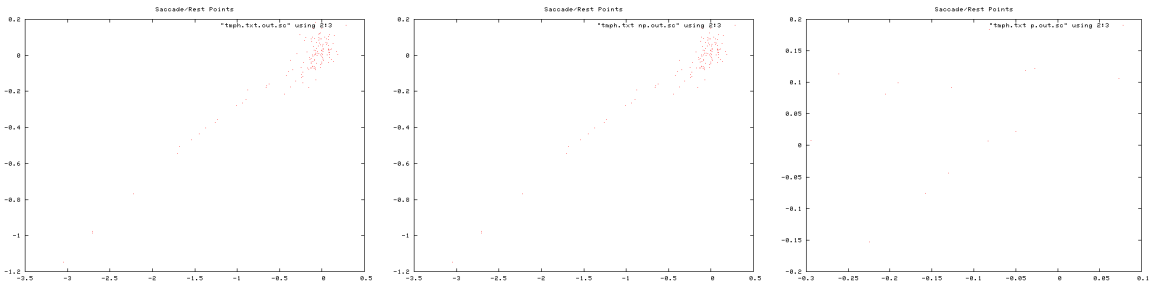



Figure B.142: Saccade gaze locations, Trial 7. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

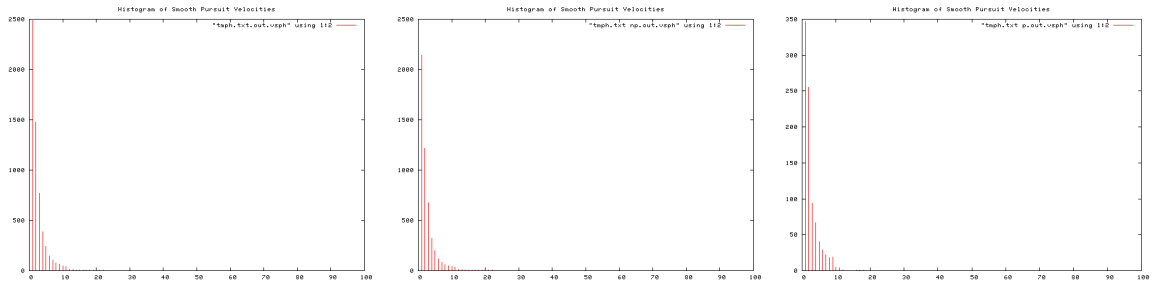



Figure B.143: Histogram of smooth pursuit velocities, Trial 7. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

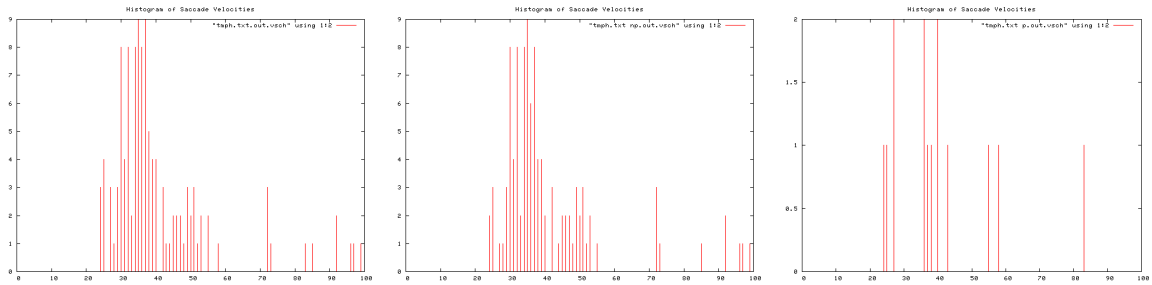



Figure B.144: Histogram of Saccade velocities, Trial 7. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

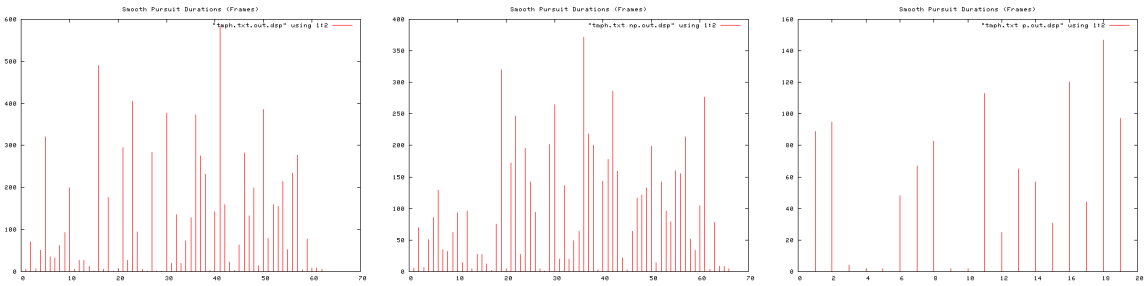



Figure B.145: Smooth pursuit durations, Trial 7. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

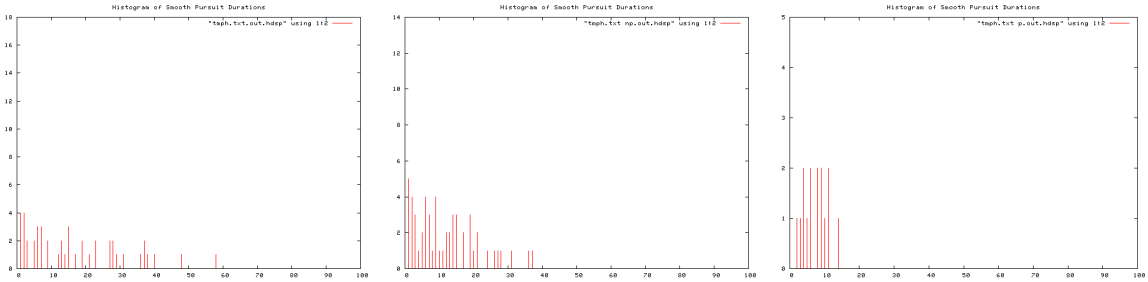



Figure B.146: Histogram of Smooth pursuit durations, Trial 7. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

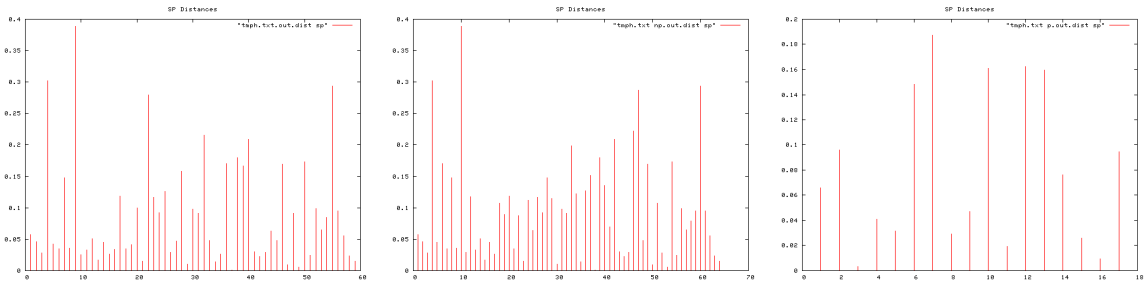



Figure B.147: Smooth pursuit distances, Trial 7. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

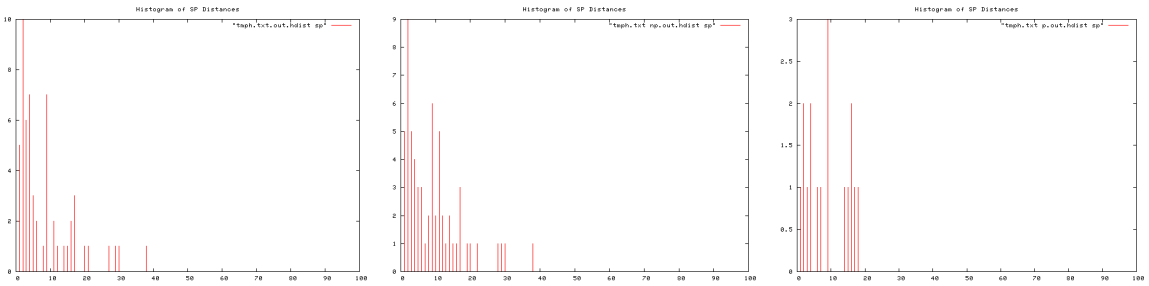



Figure B.148: Histogram of smooth pursuit distances, Trial 7. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

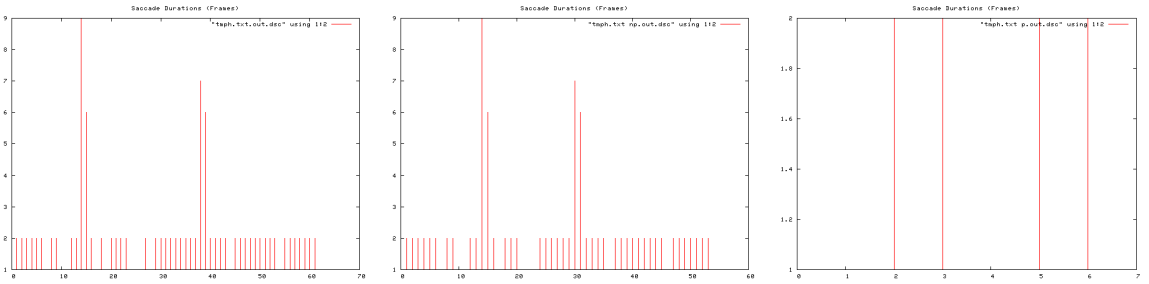



Figure B.149: Saccade durations, Trial 7. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

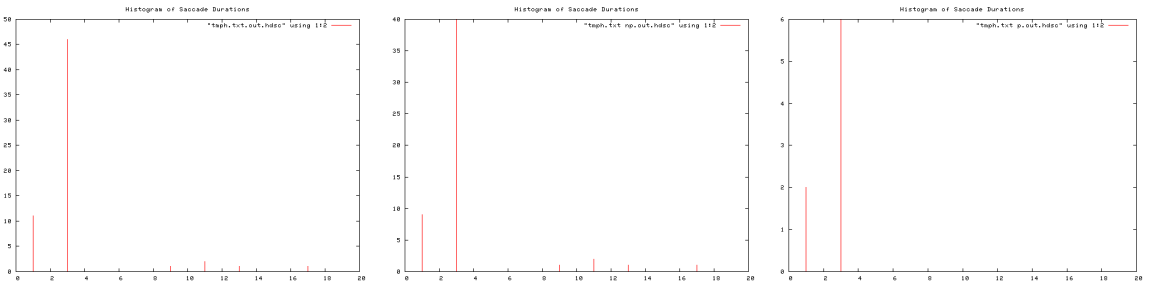



Figure B.150: Histogram of saccade durations, Trial 7. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

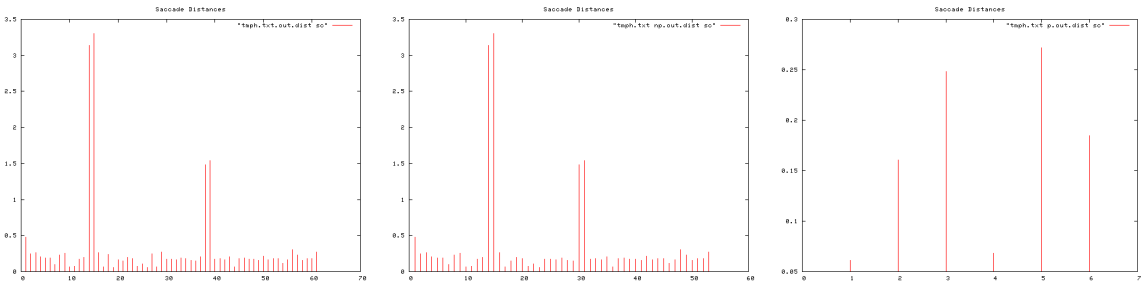



Figure B.151: Saccade distances, Trial 7. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

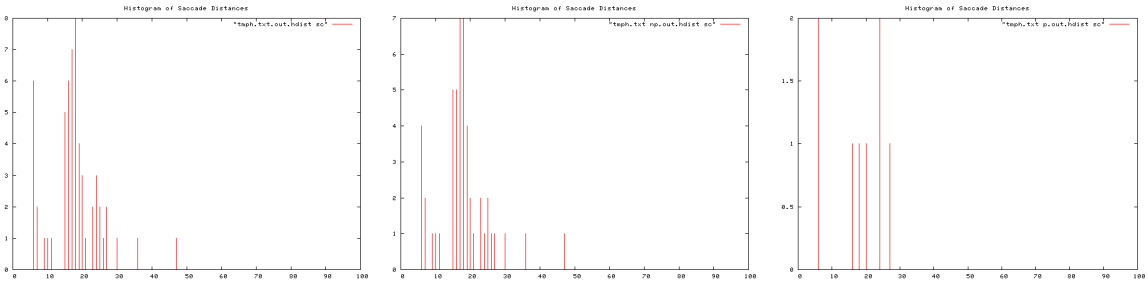



Figure B.152: Histogram of saccade distances, Trial 7. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

| Ti | To | T | I/O Event | Pear | Apple | Peach | Orange | |
|---|---|---|---|---|---|---|---|---|
| 12 | 19 | 7 | Orange In | | | | 2 | |
| 20 | 24 | 4 | Pear In | 1 | | | 1 | |
| 26 | 35 | 9 | | 1 | | | 1 | |
| 36 | 44 | 8 | Peach In | 2 | | 2 | 2 | |
| 46 | 0:53 | 7 | | 2 | | 2 | 1 | |
| 0:55 | 1:04 | 9 | | 2 | | 3 | 3 | |
| 1:05 | 1:16 | 11 | | 3 | | 3 | 4 | |
| 1:18 | 1:26 | 8 | Apple In, Peach Out | 1 | 1 | 1 | 1 | |
| 1:29 | 1:40 | 11 | Orange Out | 3 | 3 | | 3 | |
| 1:41 | 1:50 | 9 | Pear Out | 3 | 2 | | | |
| 1:54 | 2:05 | 11 | Apple Out | | 3 | | | |
| | | | TOTAL Rets | 18 | 9 | 11 | 18 | |
| | | | TOTAL T | 76 | 39 | 43 | 74 | **SD** |
| | | | **Av. Re-attention Period** | **4.2** | **4.3** | **3.9** | **4.1** | 0.17078251 |

Figure B.153: Re-attention period statistics, Trial 7.

### B.1.1.10 Trial 8

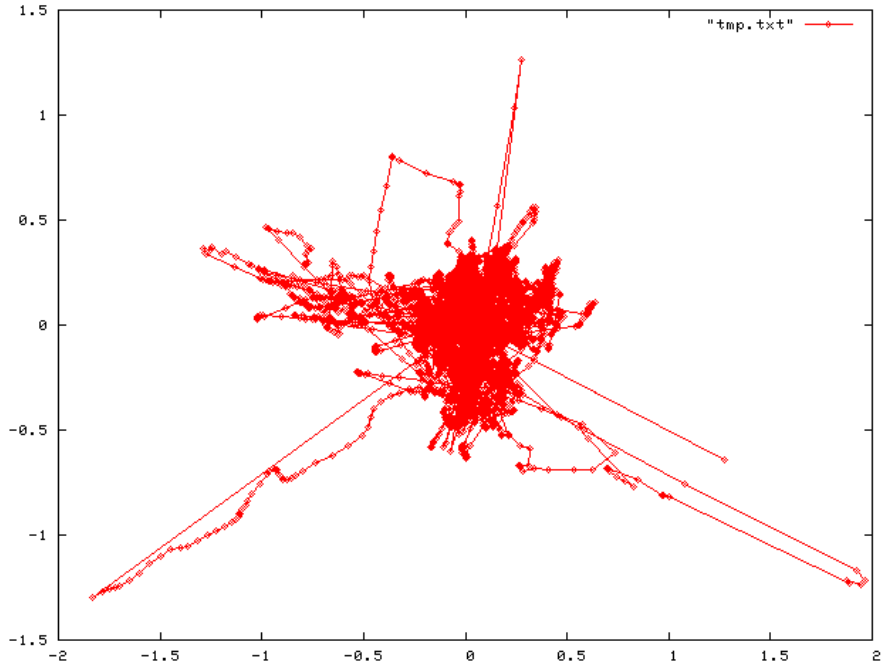



Figure B.154: Complete scan path, Trial 8.

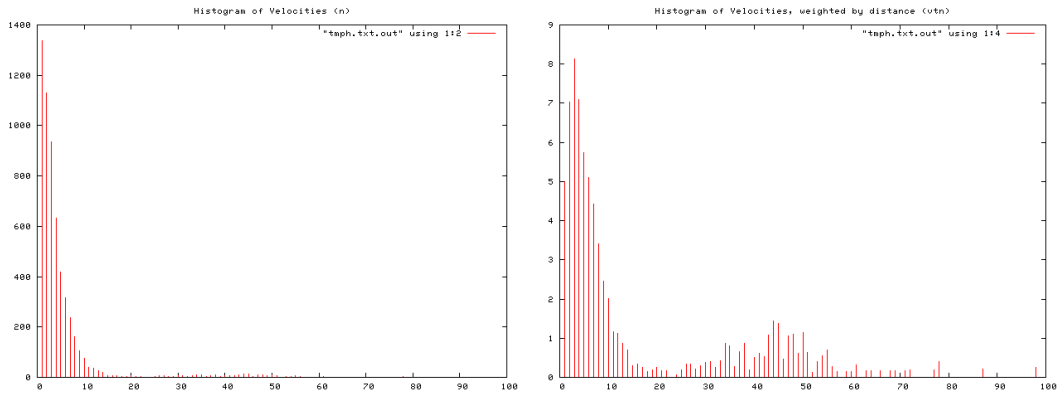



Figure B.155: Histogram of velocity magnitudes, Trial 8 (left). Histogram of distance weighted velocities, Trial 5 (right).

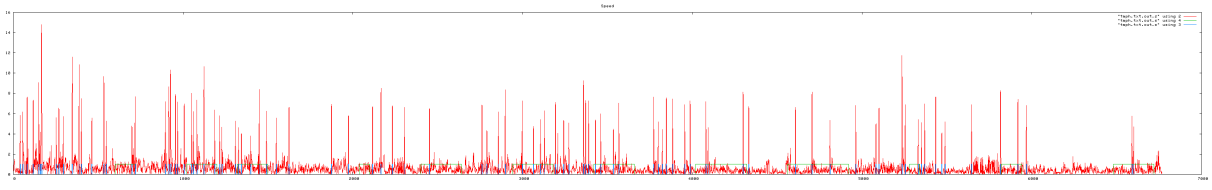


Figure B.156: Velocity profile. Velocity magnitude of each frame, Trial 8.

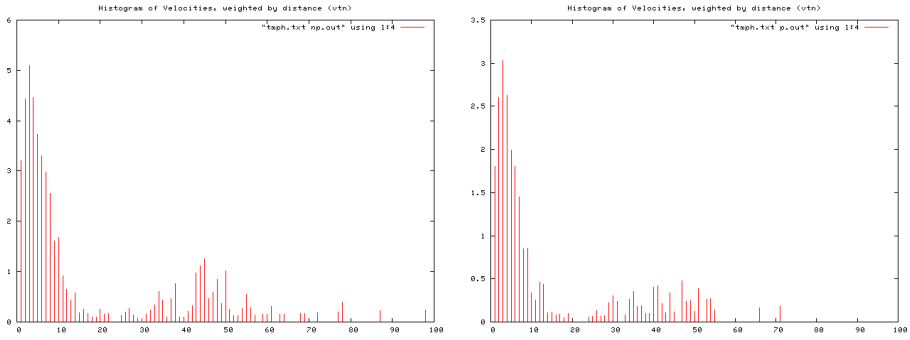



Figure B.157: Histogram of velocities during periods of no perturbation (left), and perturbation (right), Trial 8.

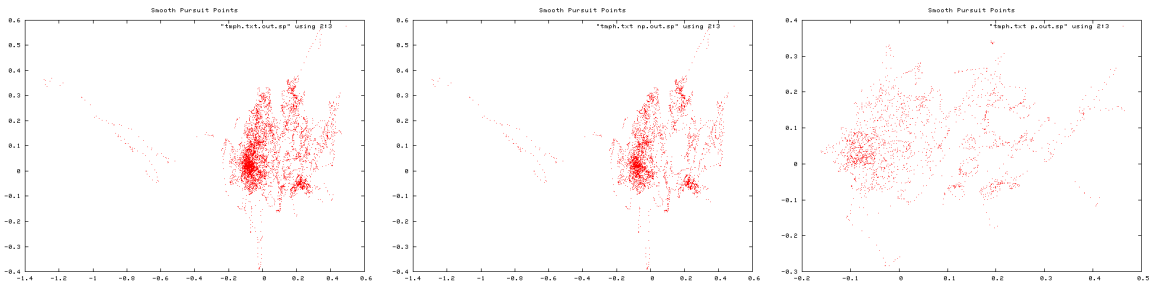



Figure B.158: Smooth pursuit gaze locations, Trial 8. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

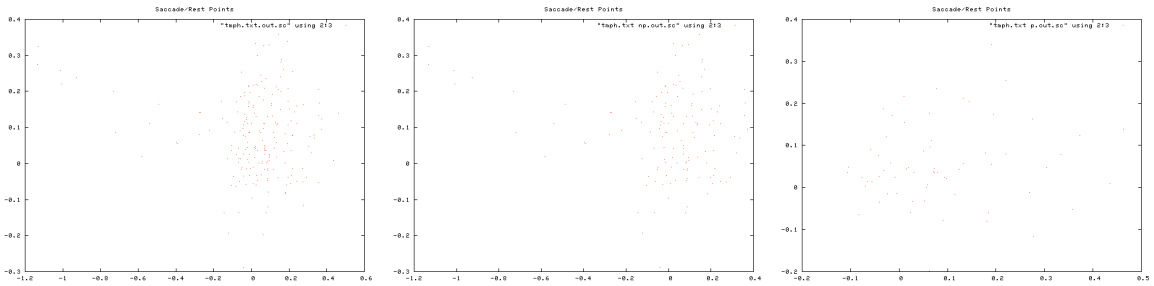



Figure B.159: Saccade gaze locations, Trial 8. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

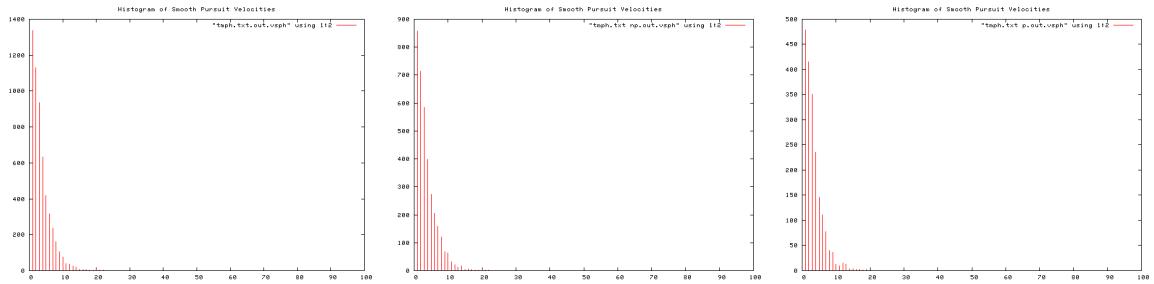



Figure B.160: Histogram of smooth pursuit velocities, Trial 8. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

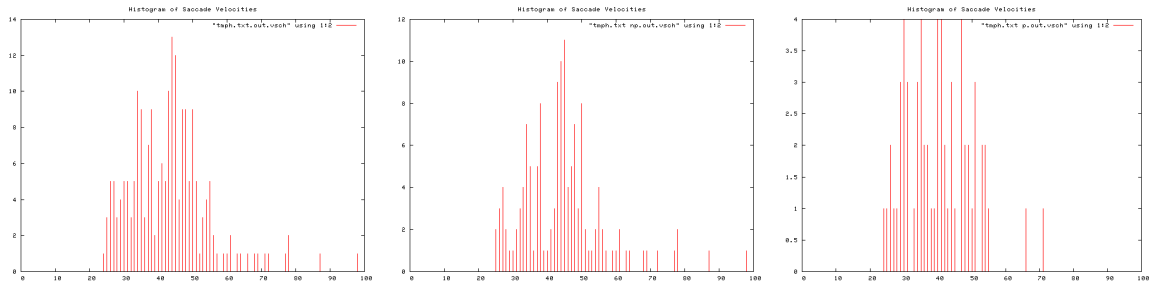



Figure B.161: Histogram of Saccade velocities, Trial 8. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

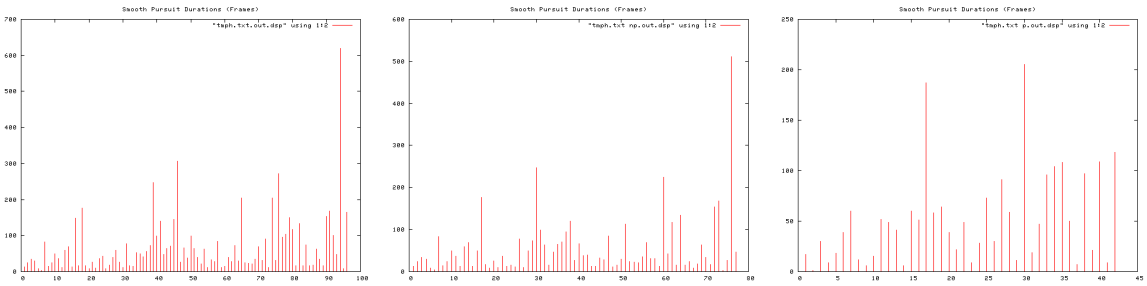



Figure B.162: Smooth pursuit durations, Trial 8. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

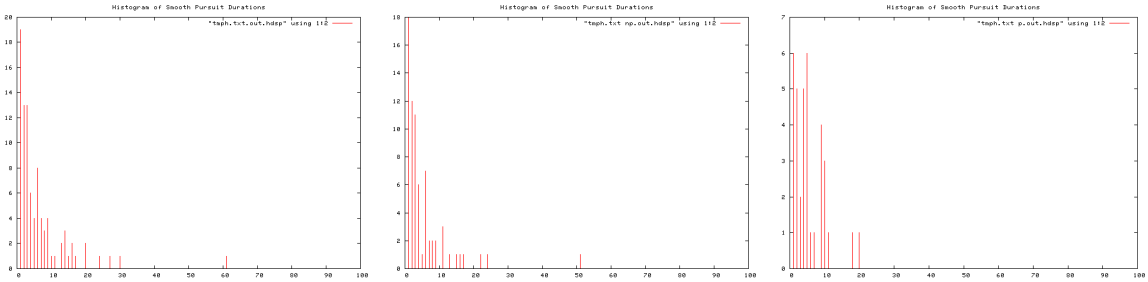



Figure B.163: Histogram of Smooth pursuit durations, Trial 8. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

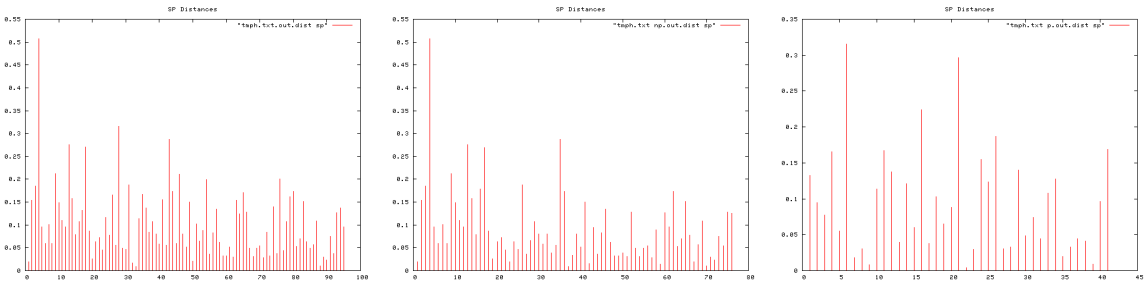



Figure B.164: Smooth pursuit distances, Trial 8. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

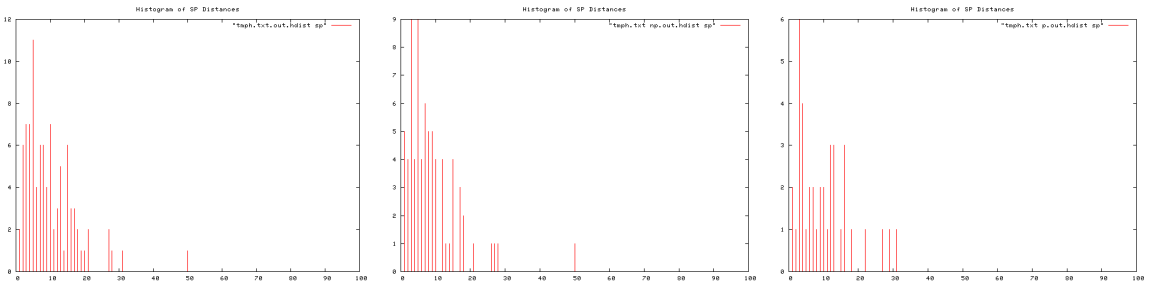



Figure B.165: Histogram of smooth pursuit distances, Trial 8. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

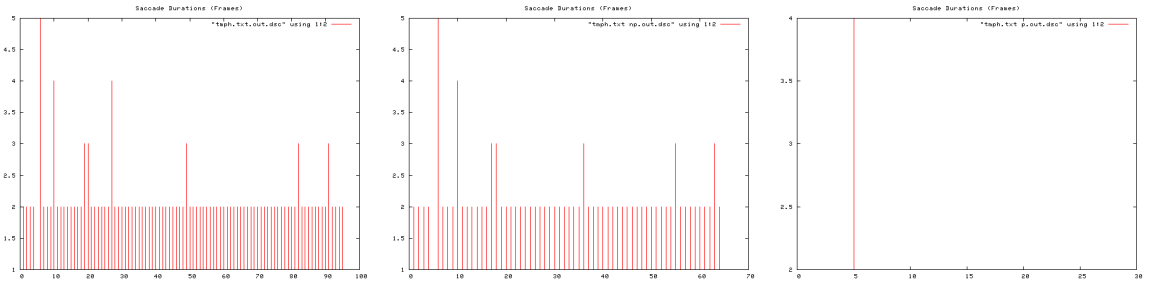



Figure B.166: Saccade durations, Trial 8. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

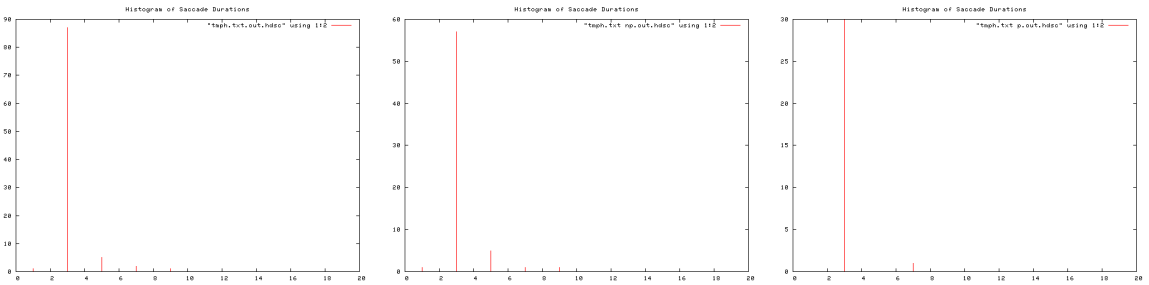



Figure B.167: Histogram of saccade durations, Trial 8. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

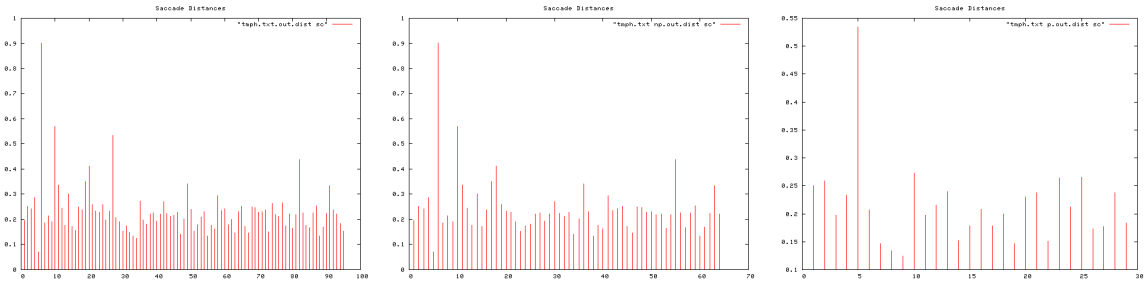



Figure B.168: Saccade distances, Trial 8. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

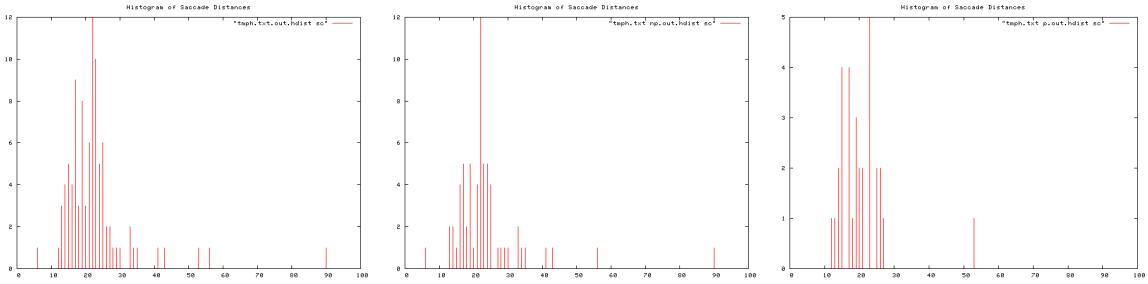



Figure B.169: Histogram of saccade distances, Trial 8. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

| Ti | To | T | I/O Event | Pear | Apple | Peach | Orange | |
|---|---|---|---|---|---|---|---|---|
| 1:07 | 1:12 | 5 | Orange In | | | | 1 | |
| 1:15 | 1:18 | 3 | Pear In | 1 | | | 1 | |
| 1:20 | 1:30 | 10 | | 2 | | | 2 | |
| 1:31 | 1:36 | 5 | Peach In | 2 | | 1 | 1 | |
| 1:38 | 1:45 | 7 | | 2 | | 2 | 2 | |
| 1:47 | 1:53 | 6 | | 2 | | 2 | 2 | |
| 1:55 | 2:03 | 8 | | 1 | | 3 | 3 | |
| 2:04 | 2:13 | 9 | Apple In, Peach Out | 1 | 2 | 1 | 1 | |
| 2:17 | 2:23 | 6 | Orange Out | 2 | 2 | | 3 | |
| 2:25 | 2:34 | 9 | Pear Out | 2 | 2 | | | |
| 2:36 | 2:44 | 8 | Apple Out | | 1 | | | |
| | | | TOTAL Rets | 15 | 7 | 9 | 16 | |
| | | | TOTAL T | 63 | 32 | 35 | 59 | **SD** |
| | | | **Av. Re-attention Period** | **4.2** | **4.6** | **3.9** | **3.7** | 0.391578 |

Figure B.170: Re-attention period statistics, Trial 8.

### B.1.1.11 Trial 9

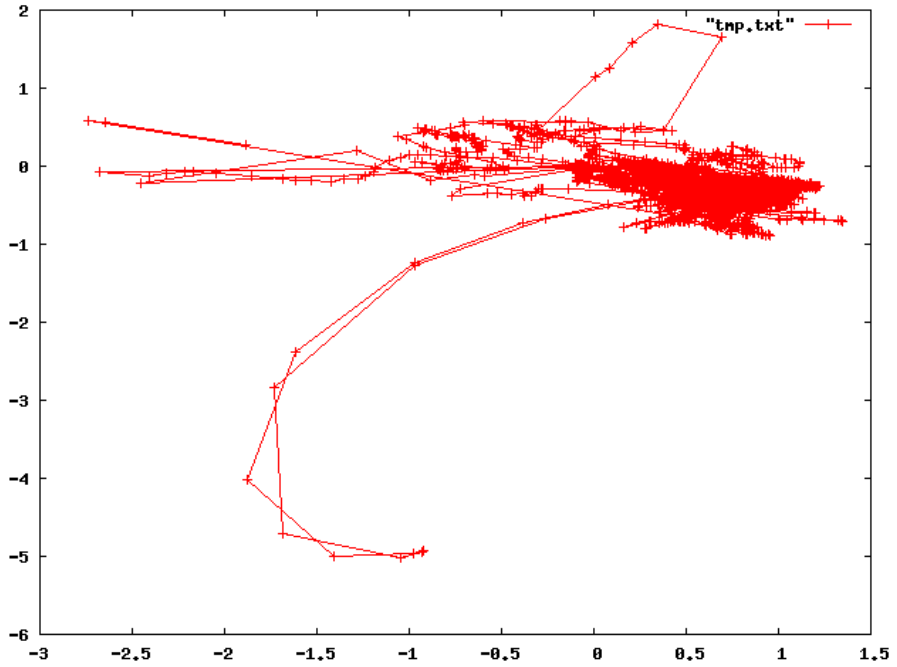



Figure B.171: Complete scan path, Trial 9.

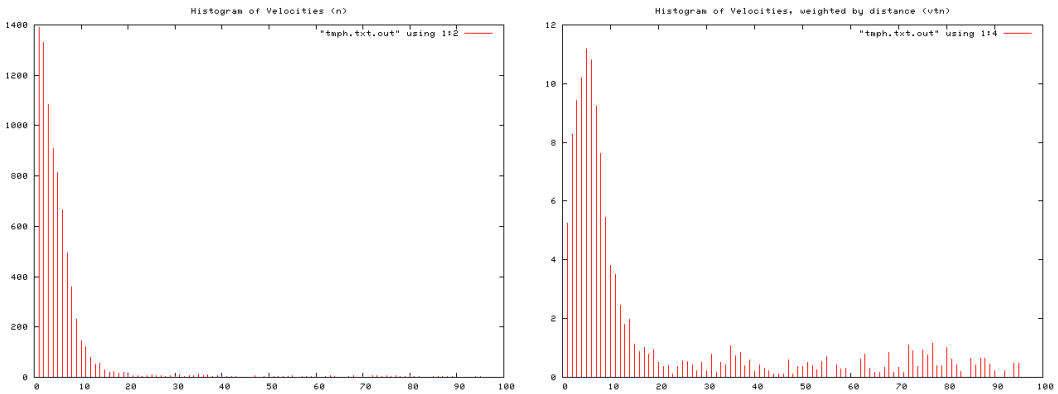



Figure B.172: Histogram of velocity magnitudes, Trial 9 (left). Histogram of distance weighted velocities, Trial 9 (right).

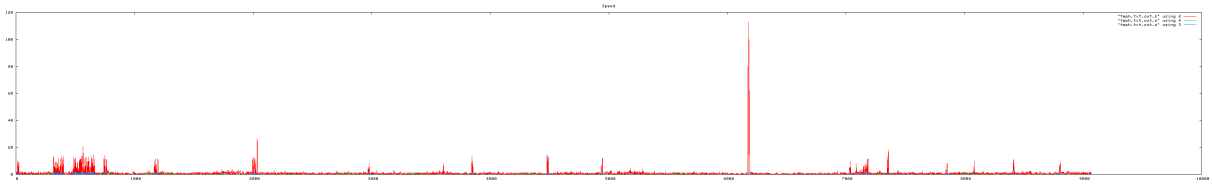


Figure B.173: Velocity profile. Velocity magnitude of each frame, Trial 9.

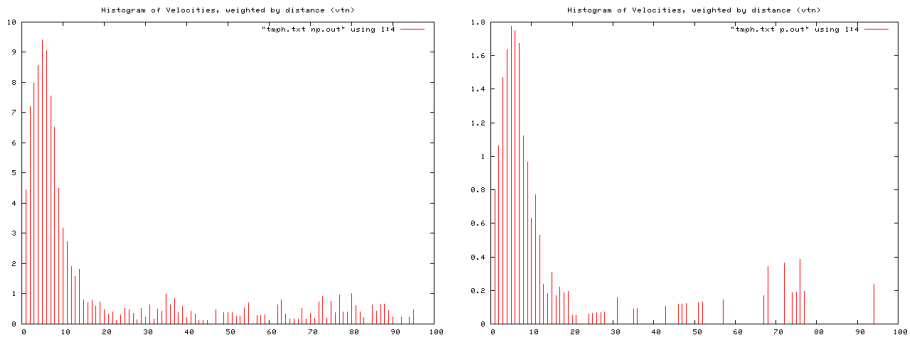



Figure B.174: Histogram of velocities during periods of no perturbation (left), and perturbation (right), Trial 9.

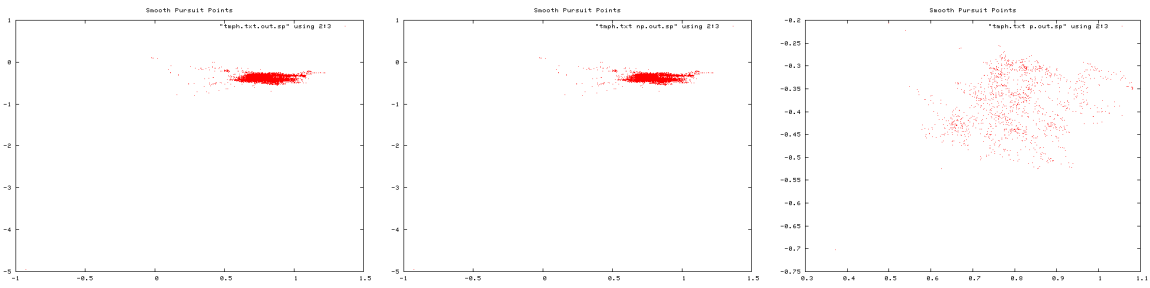



Figure B.175: Smooth pursuit gaze locations, Trial 9. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

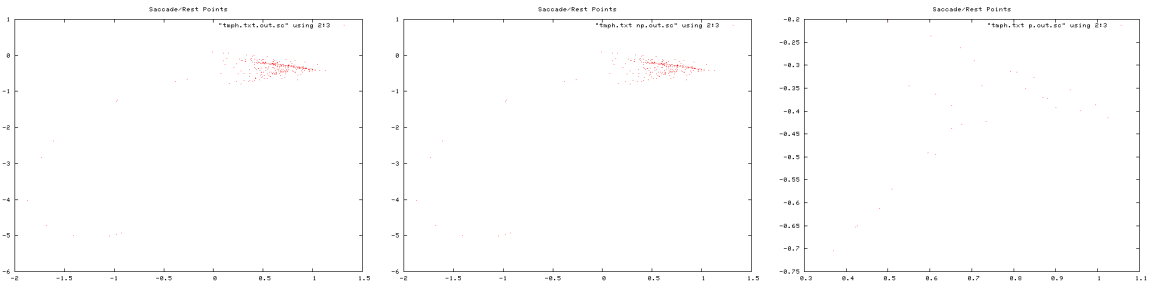



Figure B.176: Saccade gaze locations, Trial 9. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

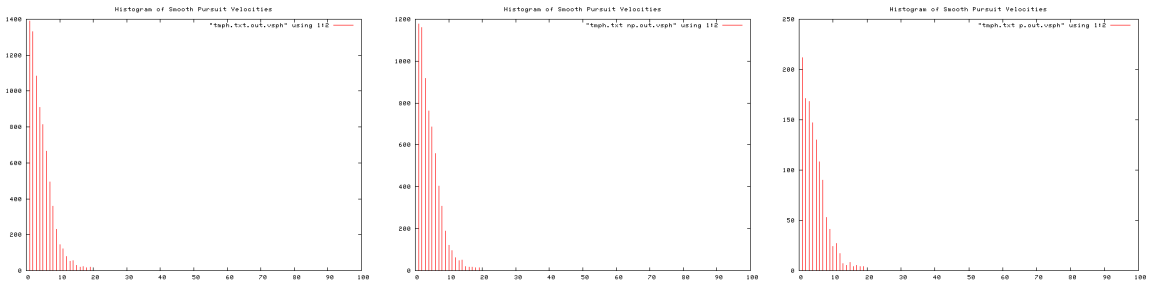



Figure B.177: Histogram of smooth pursuit velocities, Trial 9. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

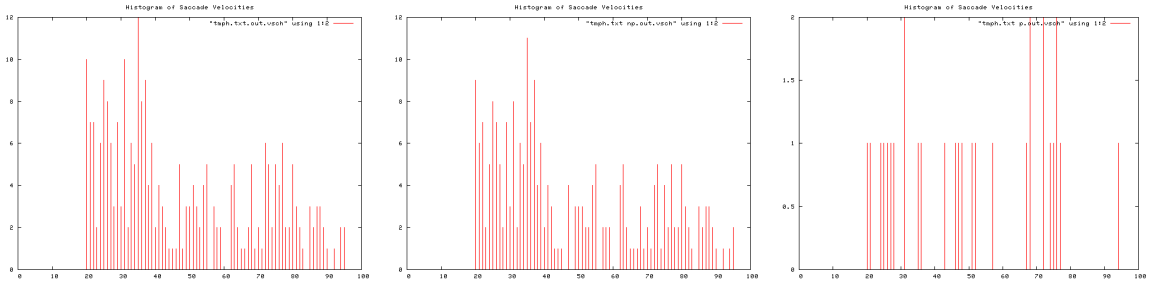



Figure B.178: Histogram of Saccade velocities, Trial 9. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

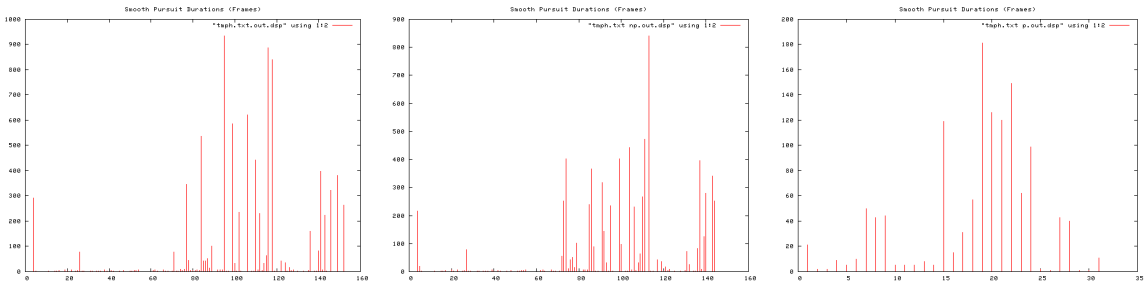



Figure B.179: Smooth pursuit durations, Trial 9. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

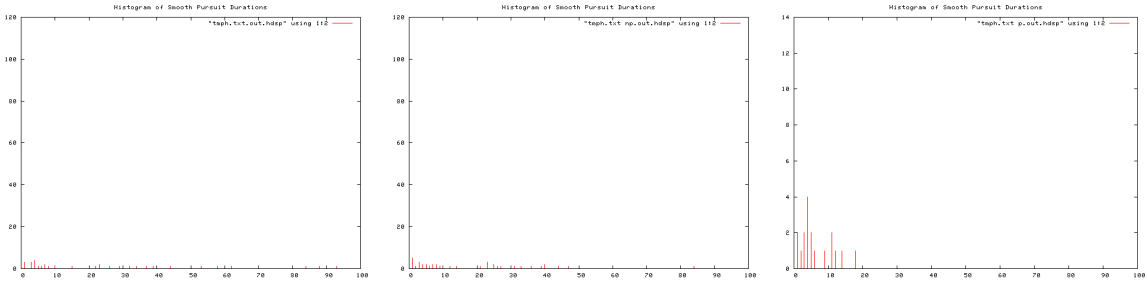



Figure B.180: Histogram of Smooth pursuit durations, Trial 9. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

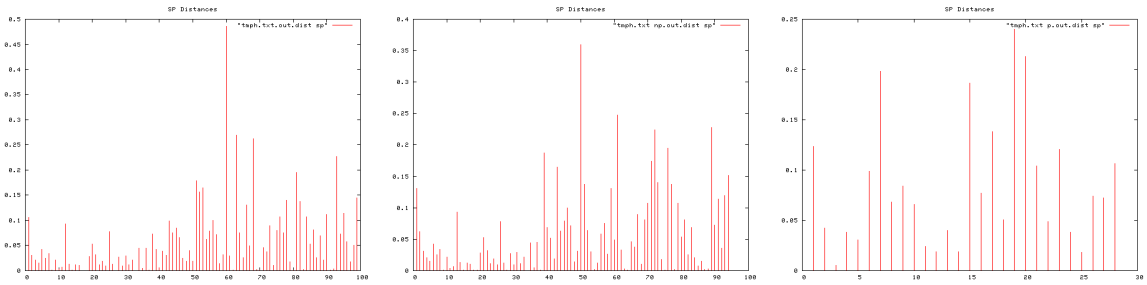



Figure B.181: Smooth pursuit distances, Trial 9. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

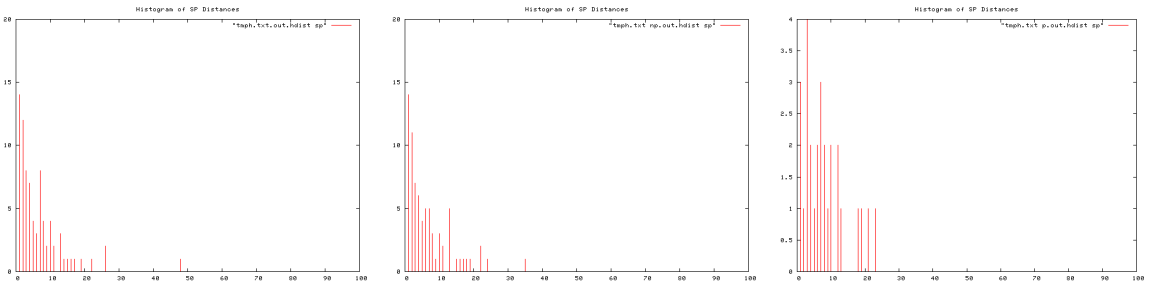



Figure B.182: Histogram of smooth pursuit distances, Trial 9. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

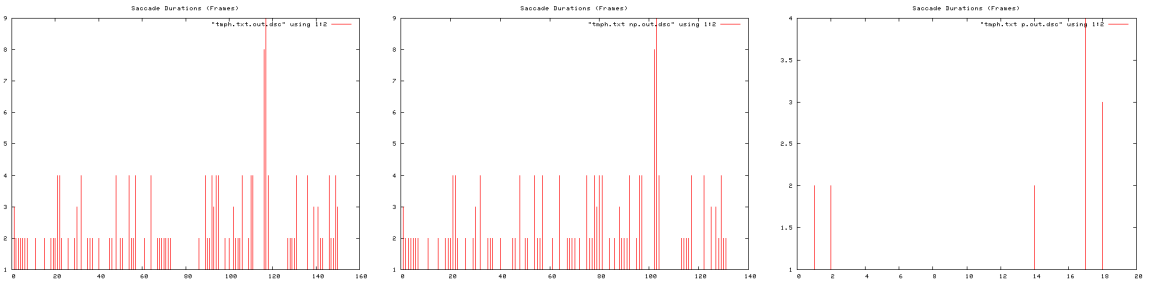



Figure B.183: Saccade durations, Trial 9. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

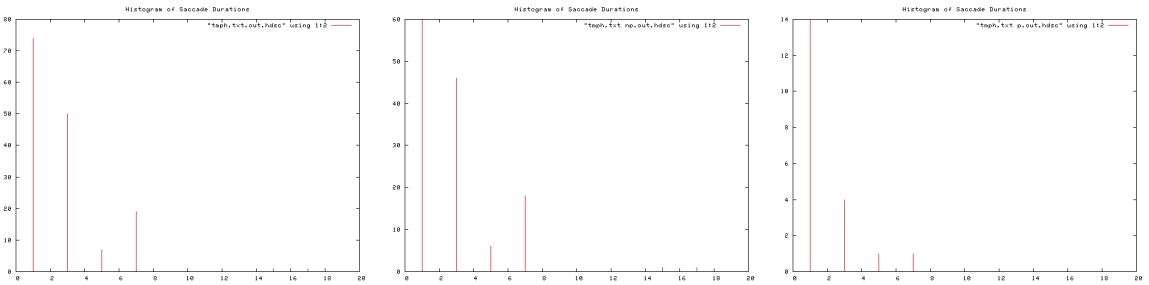



Figure B.184: Histogram of saccade durations, Trial 9. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

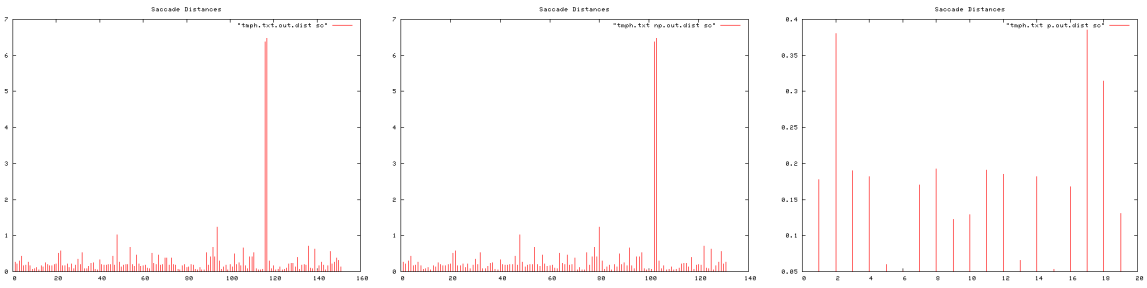



Figure B.185: Saccade distances, Trial 9. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

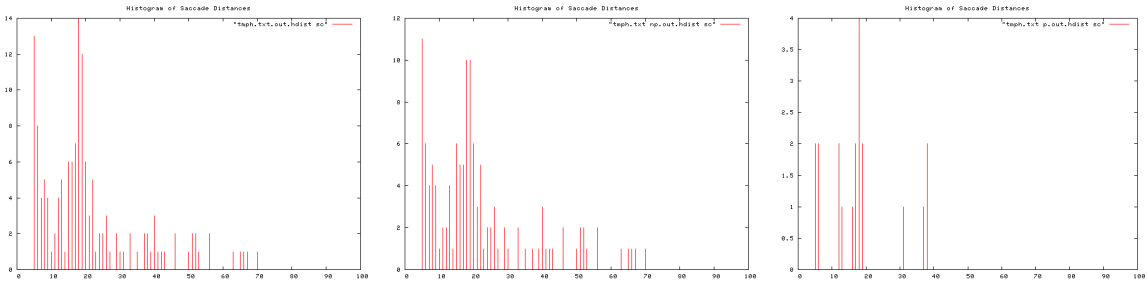



Figure B.186: Histogram of saccade distances, Trial 9. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

| Ti | To | T | I/O Event | Pear | Apple | Peach | Orange | |
|---|---|---|---|---|---|---|---|---|
| 6 | 12 | 6 | Orange In | | | | 1 | |
| 14 | 19 | 5 | Pear In | 1 | | | 0 | |
| 22 | 28 | 6 | | 1 | | | 1 | |
| 30 | 38 | 8 | Peach In | 1 | | 1 | 1 | |
| 40 | 0:46 | 6 | | 1 | | 1 | 0 | |
| 0:48 | 0:56 | 8 | | 1 | | 1 | 2 | |
| 0:58 | 1:11 | 13 | | 1 | | 2 | 1 | |
| 1:13 | 1:34 | 21 | Apple In, Peach Out | 1 | 1 | 1 | 2 | |
| 1:36 | 2:02 | 26 | Orange Out | 2 | 2 | | 1 | |
| 2:04 | 2:13 | 9 | Pear Out | 1 | 2 | | | |
| 2:18 | 2:27 | 9 | Apple Out | | 1 | | | |
| | | | TOTAL Rets | 10 | 6 | 6 | 9 | |
| | | | TOTAL T | 104 | 65 | 56 | 99 | **SD** |
| | | | **Av. Re-attention Period** | **10.4** | **10.8** | **9.3** | **11** | 0.75883683 |

Figure B.187: Re-attention period statistics, Trial 9.

### B.1.1.12 Trial 10

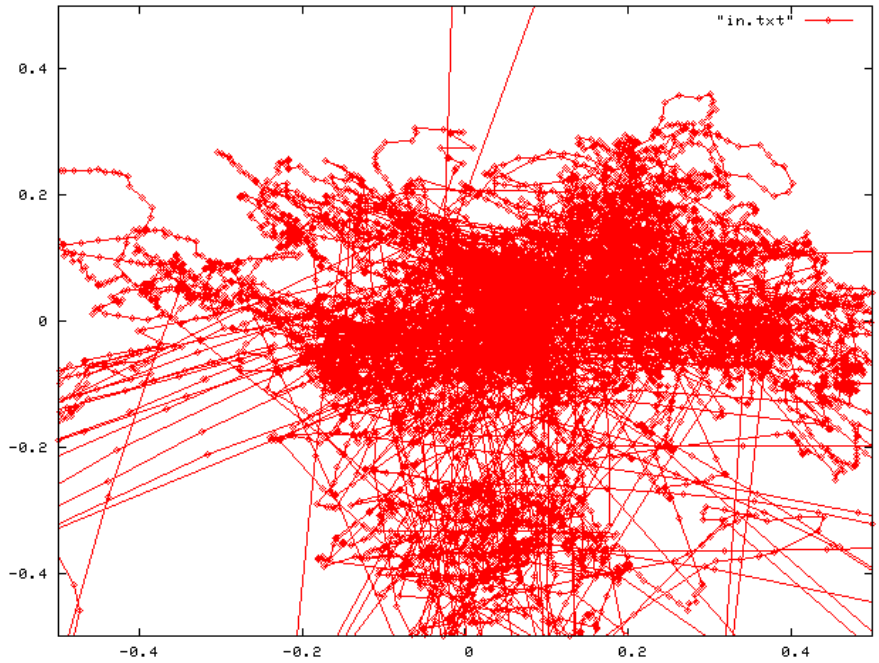



Figure B.188: Complete scan path, Trial 10.

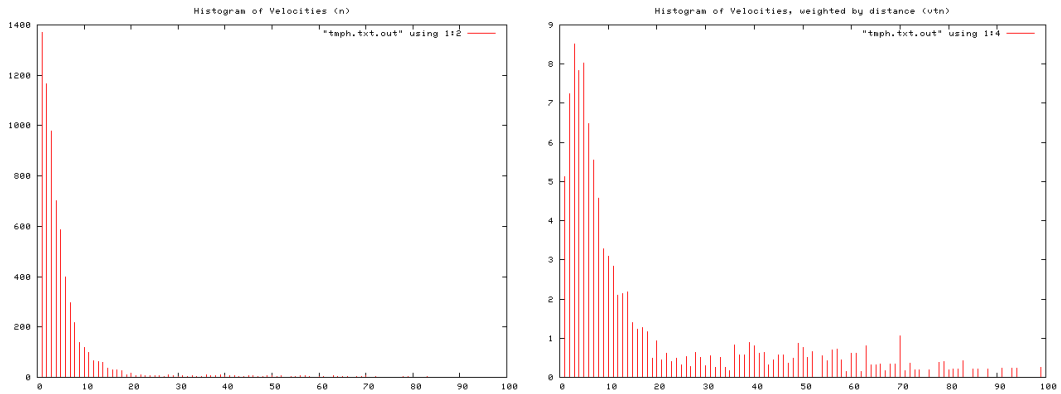



Figure B.189: Histogram of velocity magnitudes, Trial 10 (left). Histogram of distance weighted velocities, Trial 10 (right).

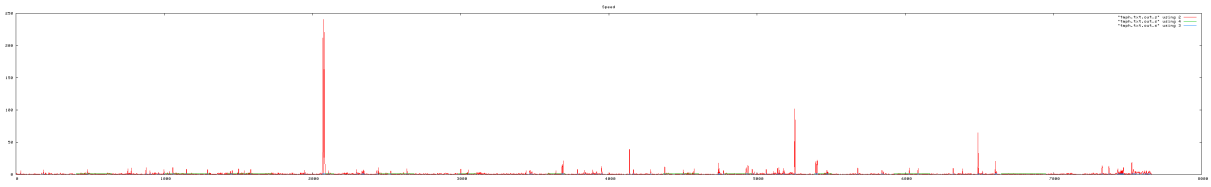


Figure B.190: Velocity profile. Velocity magnitude of each frame, Trial 10.

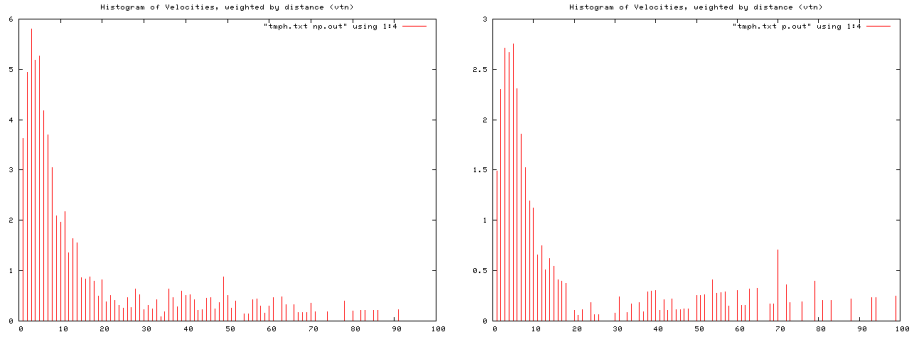



Figure B.191: Histogram of velocities during periods of no perturbation (left), and perturbation (right), Trial 10.

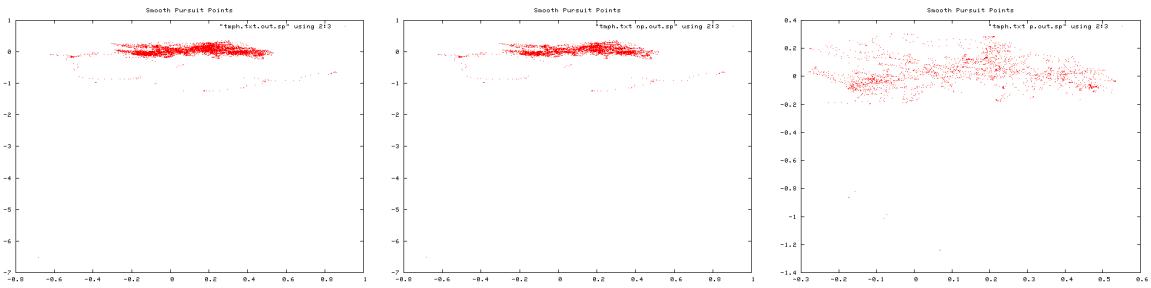



Figure B.192: Smooth pursuit gaze locations, Trial 10. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

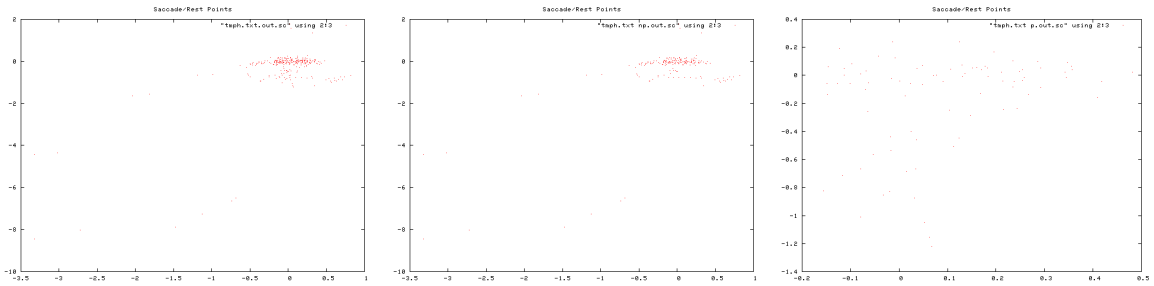



Figure B.193: Saccade gaze locations, Trial 10. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

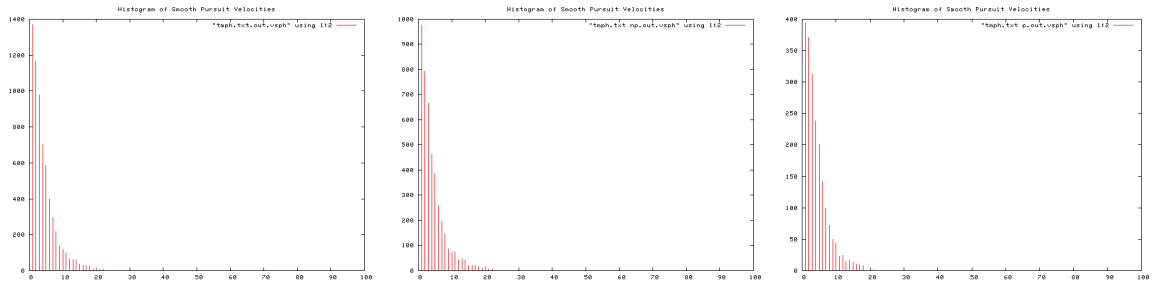



Figure B.194: Histogram of smooth pursuit velocities, Trial 10. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

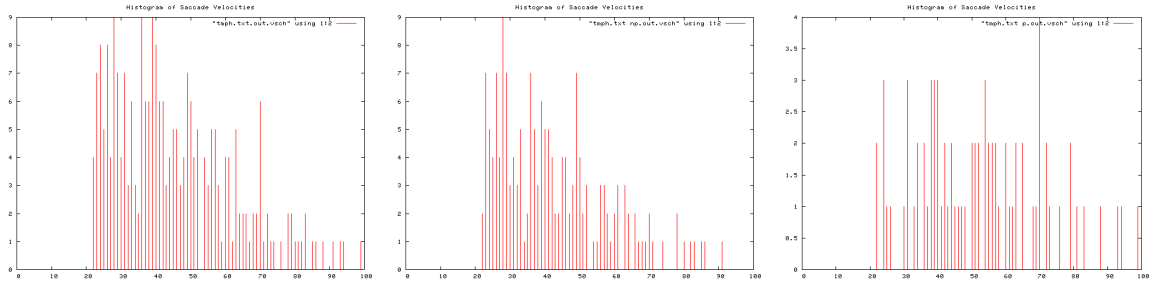



Figure B.195: Histogram of Saccade velocities, Trial 10. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

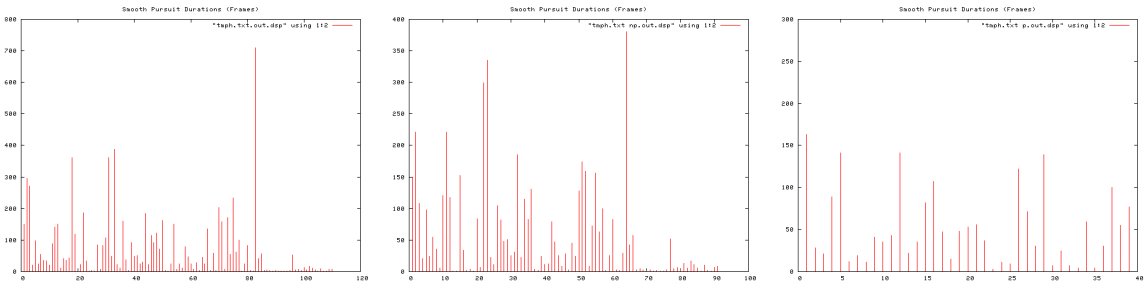



Figure B.196: Smooth pursuit durations, Trial 10. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

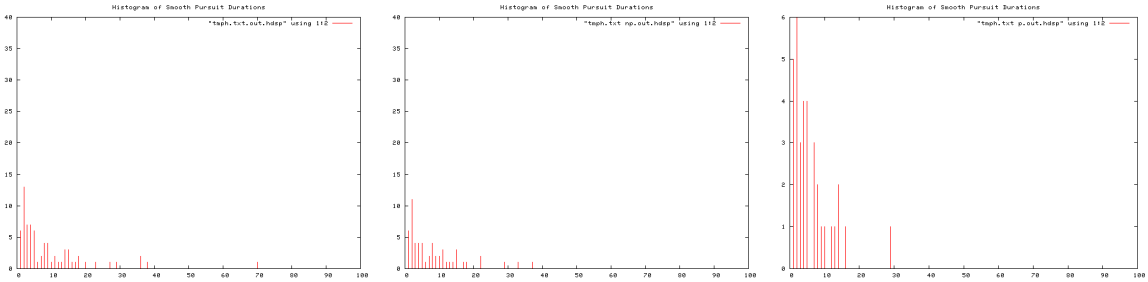



Figure B.197: Histogram of Smooth pursuit durations, Trial 10. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

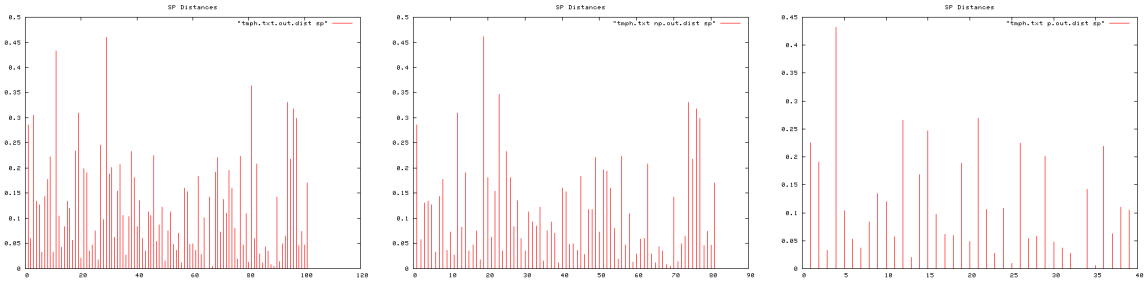



Figure B.198: Smooth pursuit distances, Trial 10. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

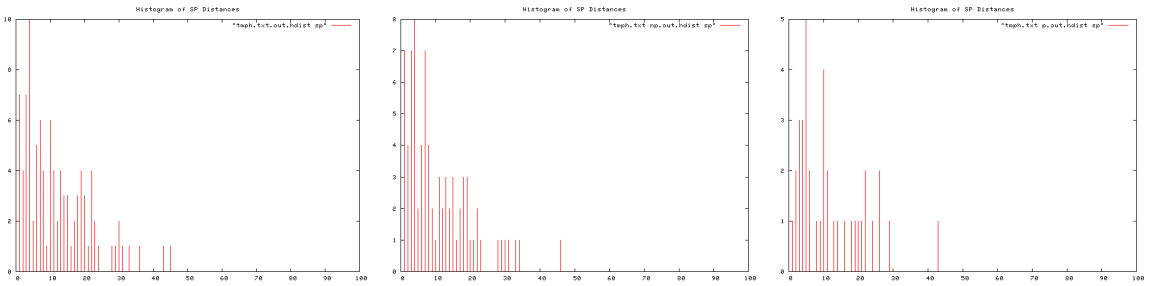



Figure B.199: Histogram of smooth pursuit distances, Trial 10. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

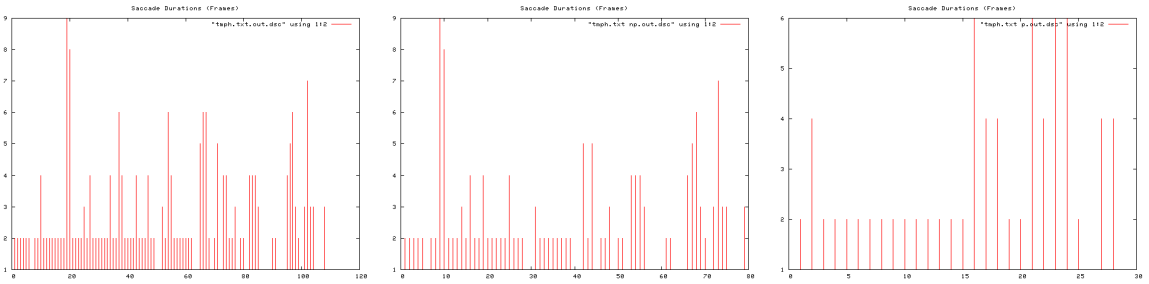



Figure B.200: Saccade durations, Trial 10. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

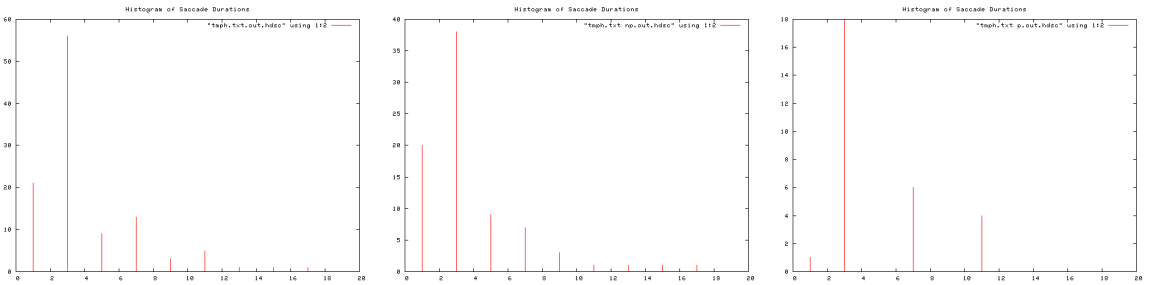



Figure B.201: Histogram of saccade durations, Trial 10. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

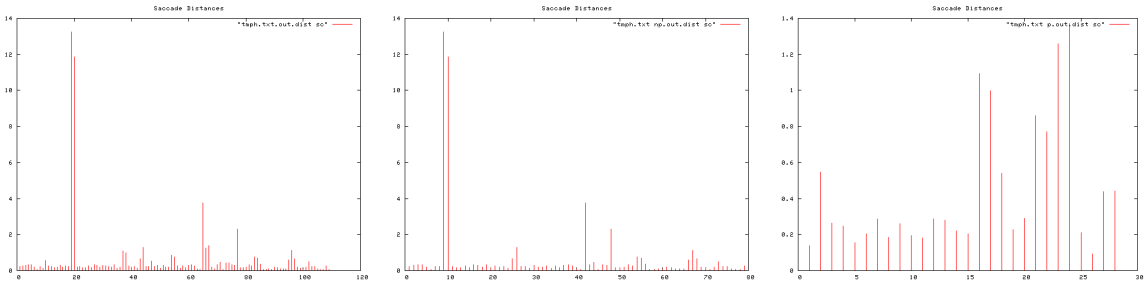



Figure B.202: Saccade distances, Trial 10. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

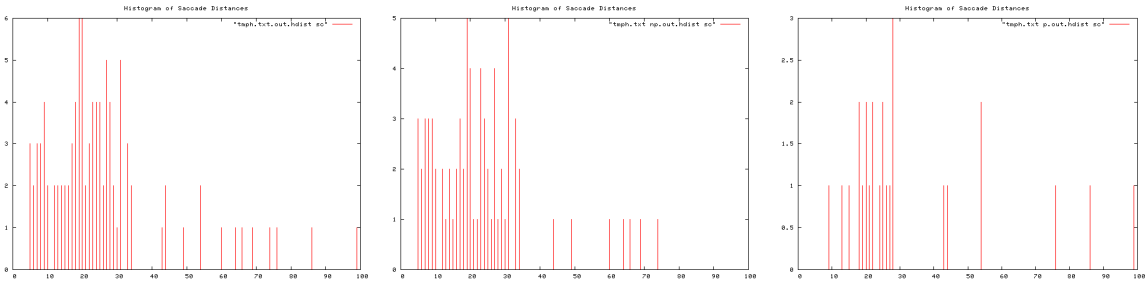



Figure B.203: Histogram of saccade distances, Trial 10. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

| Ti | To | T | I/O Event | Pear | Apple | Peach | Orange | |
|---|---|---|---|---|---|---|---|---|
| 43 | 51 | 8 | Orange In | | | | 2 | |
| 52 | 59 | 7 | Pear In | 1 | | | 1 | |
| 1:01 | 1:09 | 8 | | 3 | | | 2 | |
| 1:10 | 1:16 | 6 | Peach In | 1 | | 1 | 1 | |
| 1:18 | 1:23 | 5 | | 2 | | 2 | 2 | |
| 1:25 | 1:34 | 9 | | 1 | | 1 | 1 | |
| 1:35 | 1:46 | 11 | | 3 | | 2 | 1 | |
| 1:49 | 1:54 | 5 | Apple In, Peach Out | 1 | 2 | 1 | 1 | |
| 1:56 | 2:03 | 7 | Orange Out | 0 | 2 | | 3 | |
| 2:04 | 2:13 | 9 | Pear Out | 1 | 2 | | | |
| 2:16 | 2:25 | 9 | Apple Out | | 3 | | | |
| | | | TOTAL Rets | 13 | 9 | 7 | 14 | |
| | | | TOTAL T | 51 | 36 | 27 | 47 | **SD** |
| | | | **Av. Re-attention Period** | **3.9** | **4** | **3.9** | **3.6** | 0.17320508 |

Figure B.204: Re-attention period statistics, Trial 10.

### B.1.1.13 Trial 11

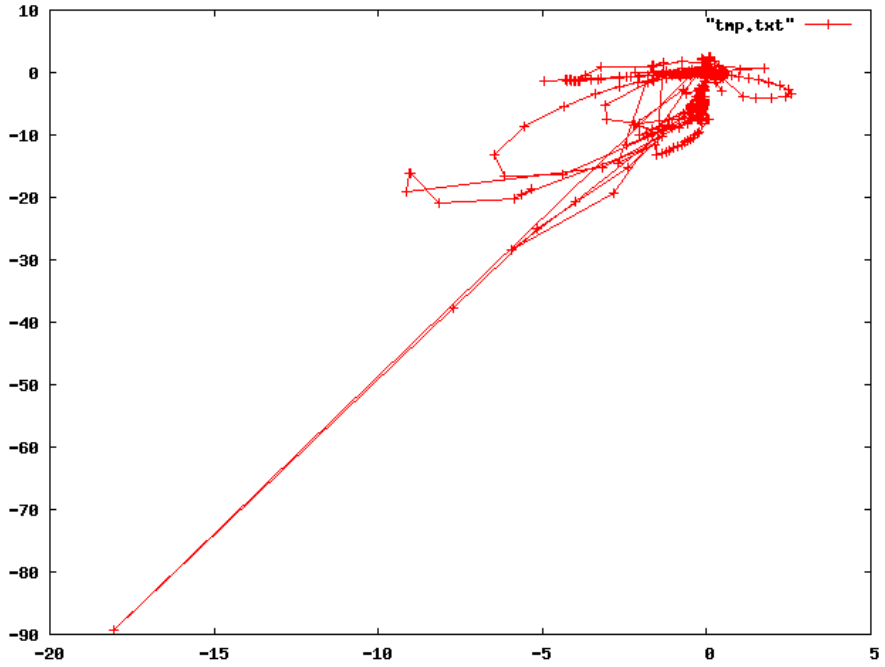



Figure B.205: Complete scan path, Trial 11.

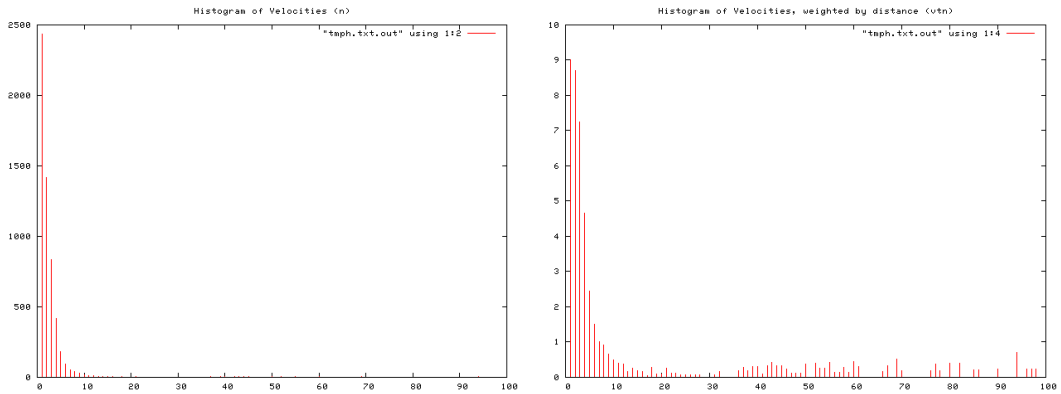



Figure B.206: Histogram of velocity magnitudes, Trial 11 (left). Histogram of distance weighted velocities, Trial 11 (right).

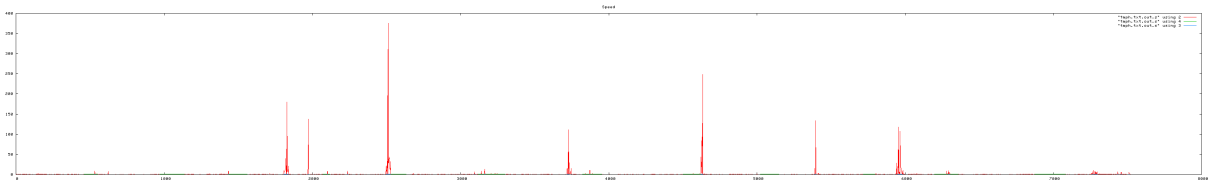


Figure B.207: Velocity profile. Velocity magnitude of each frame, Trial 11.

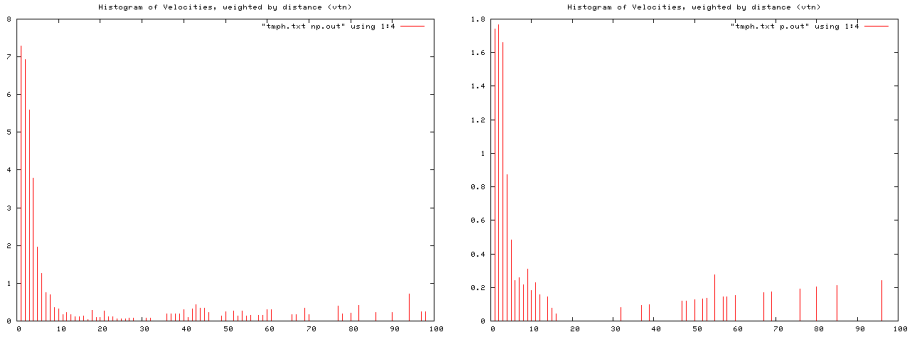



Figure B.208: Histogram of velocities during periods of no perturbation (left), and perturbation (right), Trial 11.

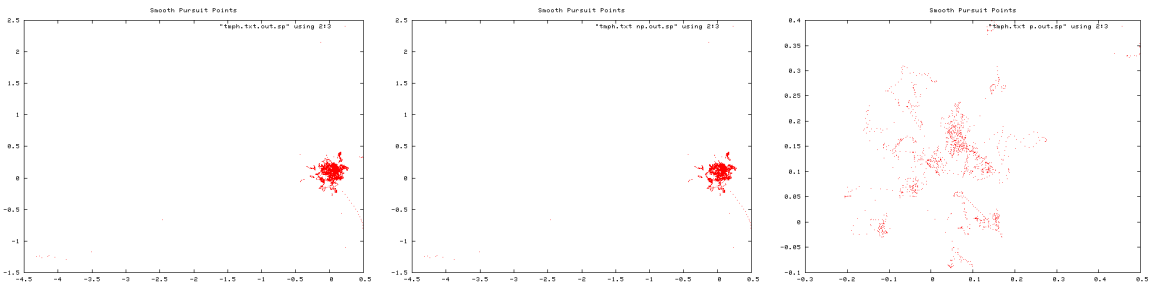



Figure B.209: Smooth pursuit gaze locations, Trial 11. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

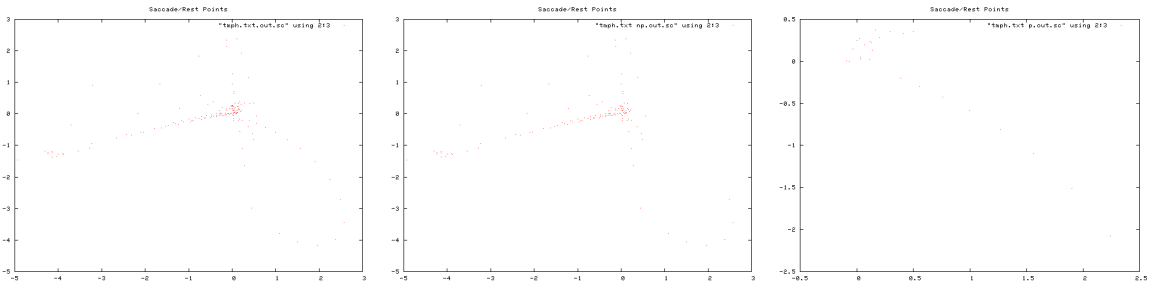



Figure B.210: Saccade gaze locations, Trial 11. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

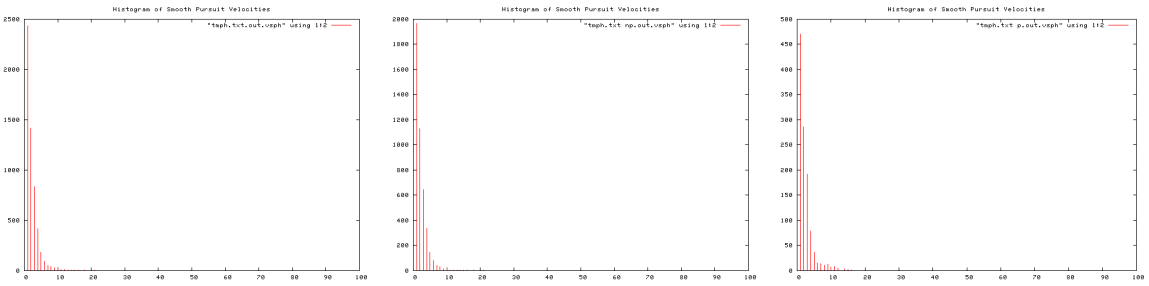



Figure B.211: Histogram of smooth pursuit velocities, Trial 11. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

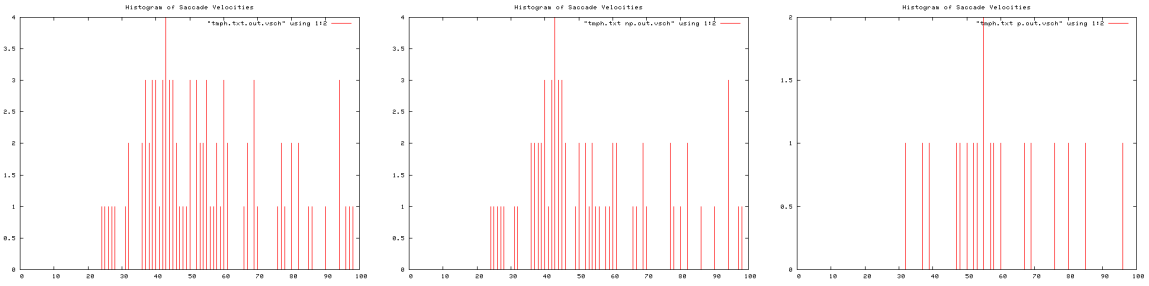



Figure B.212: Histogram of Saccade velocities, Trial 11. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

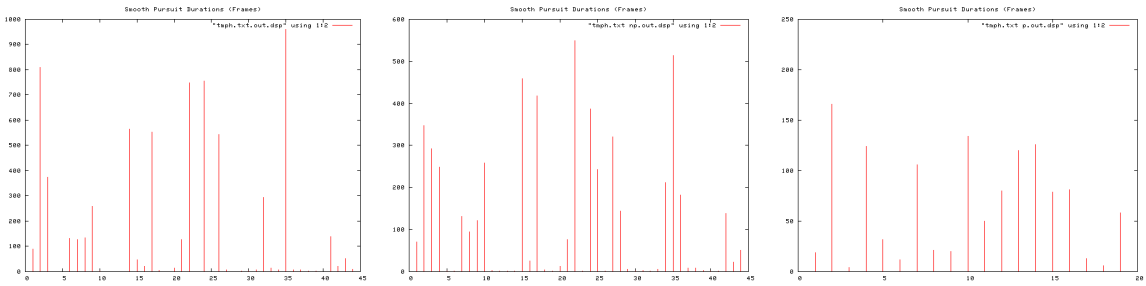



Figure B.213: Smooth pursuit durations, Trial 11. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

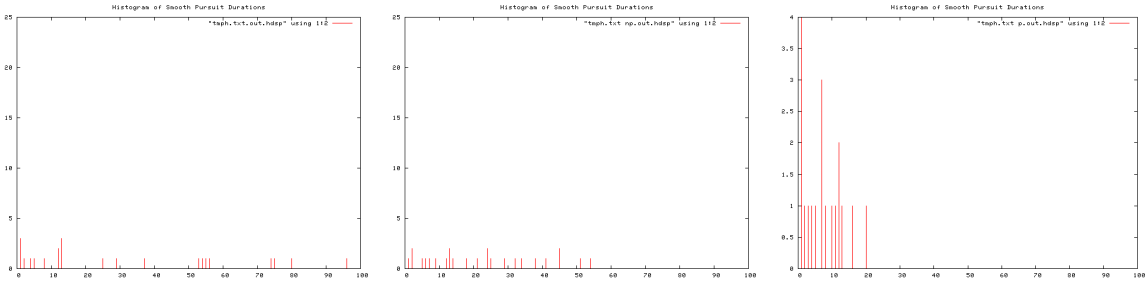



Figure B.214: Histogram of Smooth pursuit durations, Trial 11. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

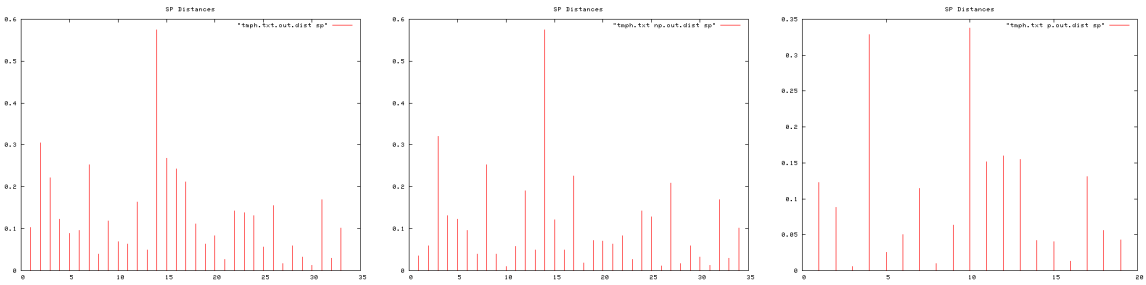



Figure B.215: Smooth pursuit distances, Trial 11. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

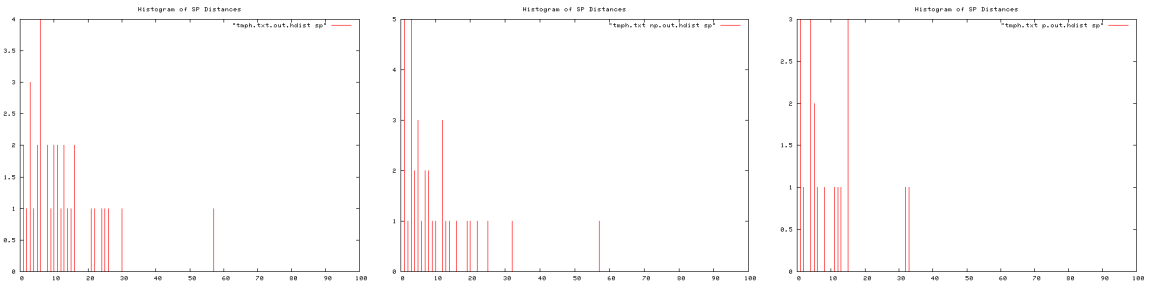



Figure B.216: Histogram of smooth pursuit distances, Trial 11. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

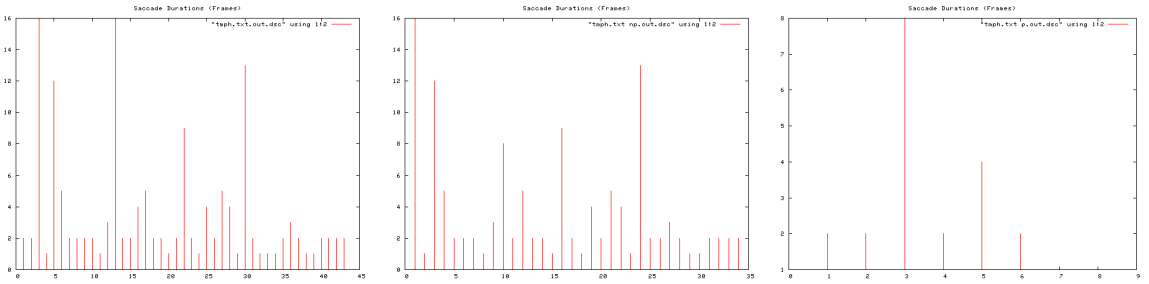



Figure B.217: Saccade durations, Trial 11. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

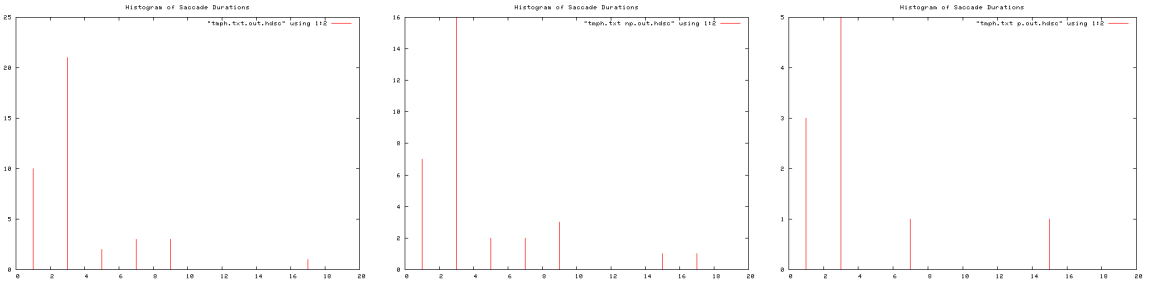



Figure B.218: Histogram of saccade durations, Trial 11. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

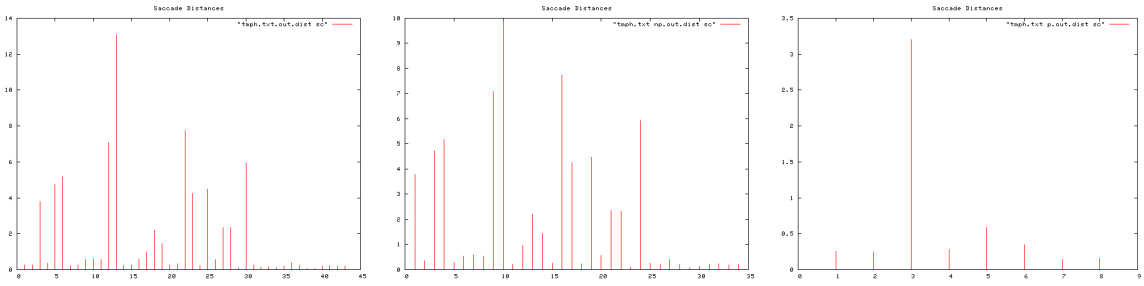



Figure B.219: Saccade distances, Trial 11. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

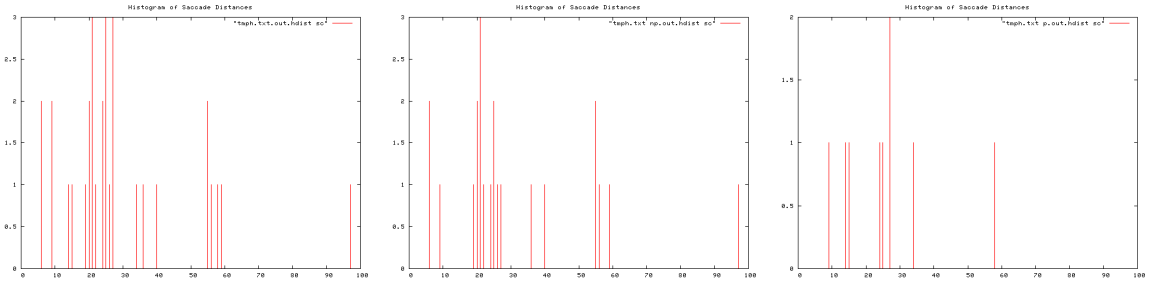



Figure B.220: Histogram of saccade distances, Trial 11. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

| Ti | To | T | I/O Event | Pear | Apple | Peach | Orange | |
|---|---|---|---|---|---|---|---|---|
| 9 | 16 | 7 | Orange In | | | | 1 | |
| 17 | 24 | 7 | Pear In | 0 | | | 1 | |
| 27 | 35 | 8 | | 1 | | | 1 | |
| 35 | 43 | 8 | Peach In | 1 | | 1 | 1 | |
| 44 | 0:53 | 9 | | 1 | | 1 | 1 | |
| 0:55 | 1:04 | 9 | | 1 | | 1 | 0 | |
| 1:05 | 1:16 | 11 | | 1 | | 1 | 1 | |
| 1:17 | 1:24 | 7 | Apple In, Peach Out | 1 | 1 | 1 | 0 | |
| 1:26 | 1:35 | 9 | Orange Out | 0 | 1 | | 1 | |
| 1:36 | 1:44 | 8 | Pear Out | 1 | 1 | | | |
| 1:46 | 1:57 | 11 | Apple Out | | 1 | | | |
| | | | TOTAL Rets | 7 | 4 | 5 | 7 | |
| | | | TOTAL T | 76 | 35 | 44 | 75 | **SD** |
| | | | **Av. Re-attention Period** | **10.8** | **8.8** | **8.8** | **10.7** | 1.12657297 |

Figure B.221: Re-attention period statistics, Trial 11.

### B.1.1.14 Trial 12

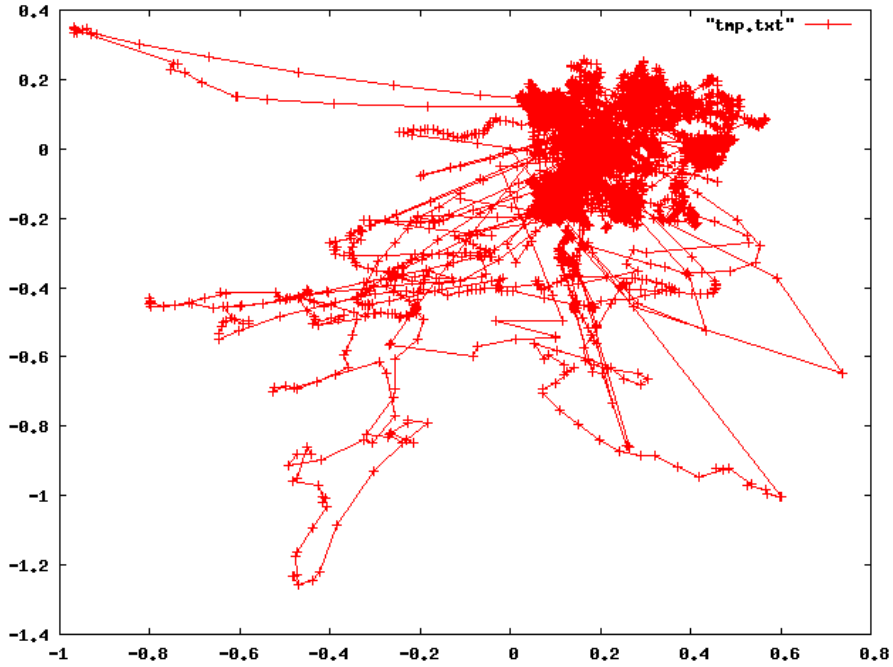



Figure B.222: Complete scan path, Trial 12.

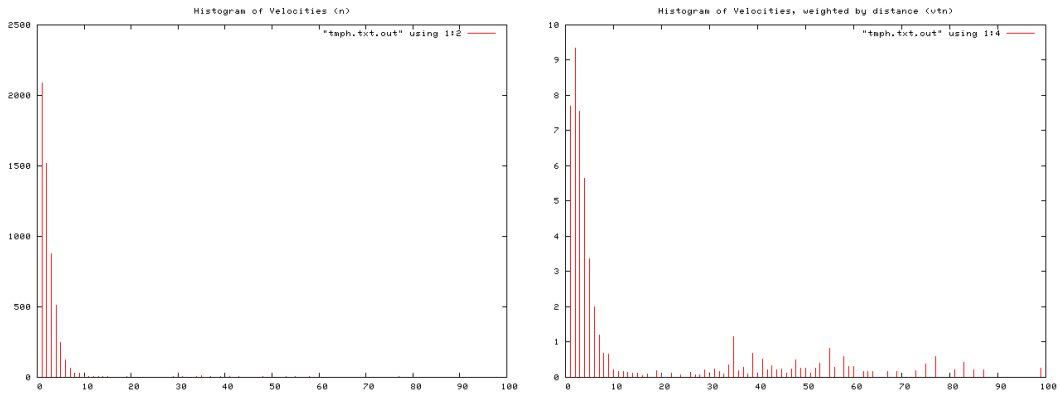



Figure B.223: Histogram of velocity magnitudes, Trial 12 (left). Histogram of distance weighted velocities, Trial 12 (right).

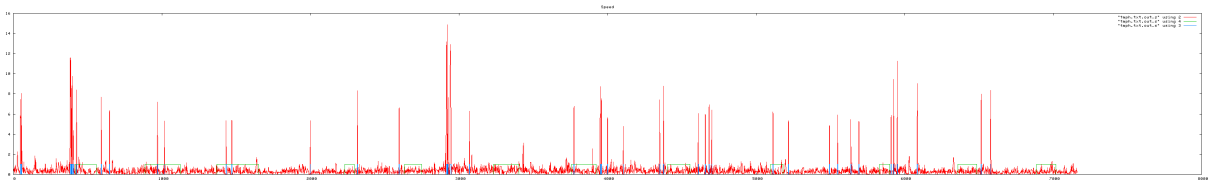


Figure B.224: Velocity profile. Velocity magnitude of each frame, Trial 12.

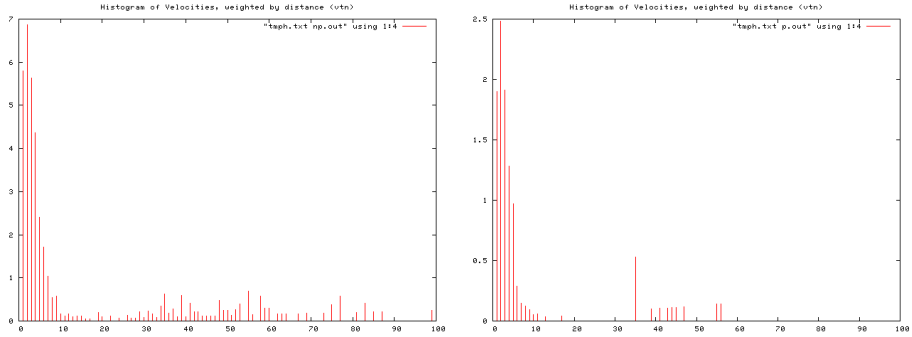



Figure B.225: Histogram of velocities during periods of no perturbation (left), and perturbation (right), Trial 12.

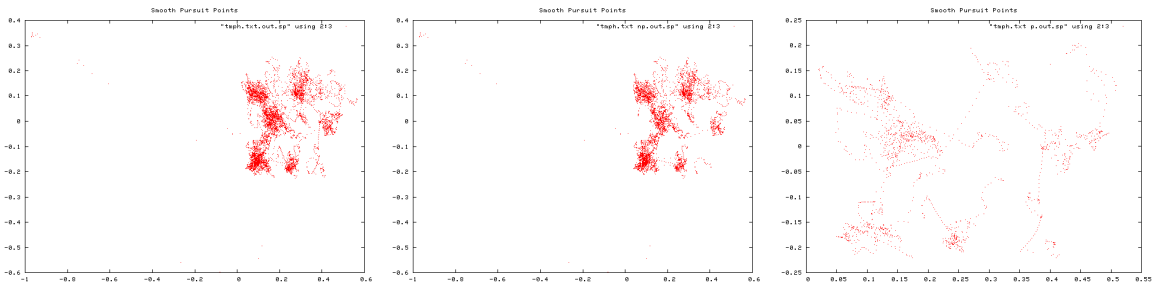



Figure B.226: Smooth pursuit gaze locations, Trial 12. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

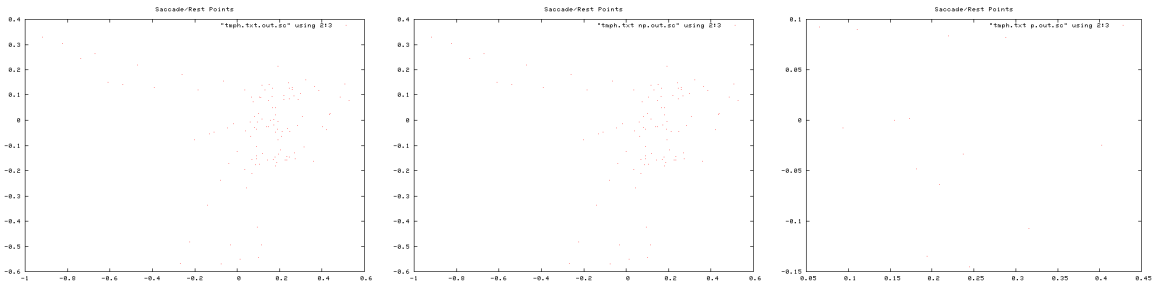



Figure B.227: Saccade gaze locations, Trial 12. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

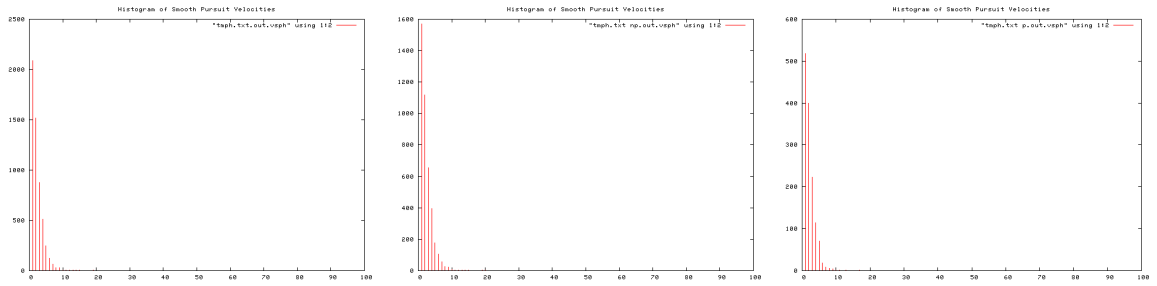



Figure B.228: Histogram of smooth pursuit velocities, Trial 12. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

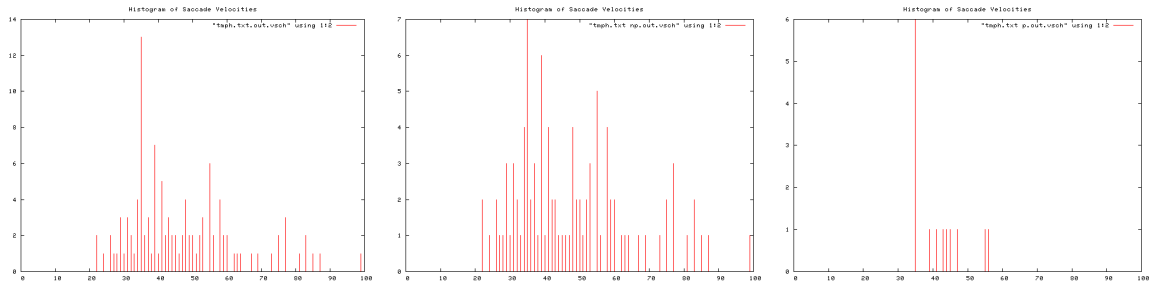



Figure B.229: Histogram of Saccade velocities, Trial 12. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

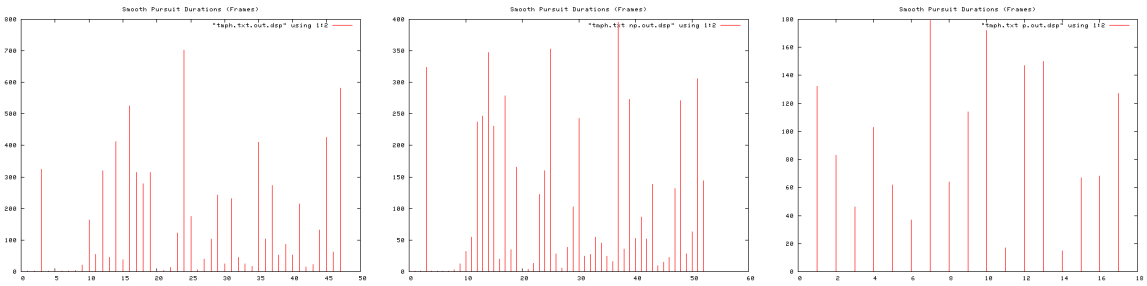



Figure B.230: Smooth pursuit durations, Trial 12. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

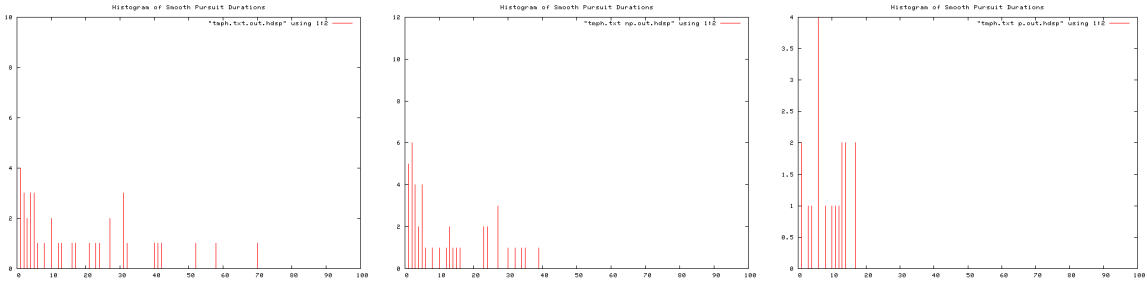



Figure B.231: Histogram of Smooth pursuit durations, Trial 12. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

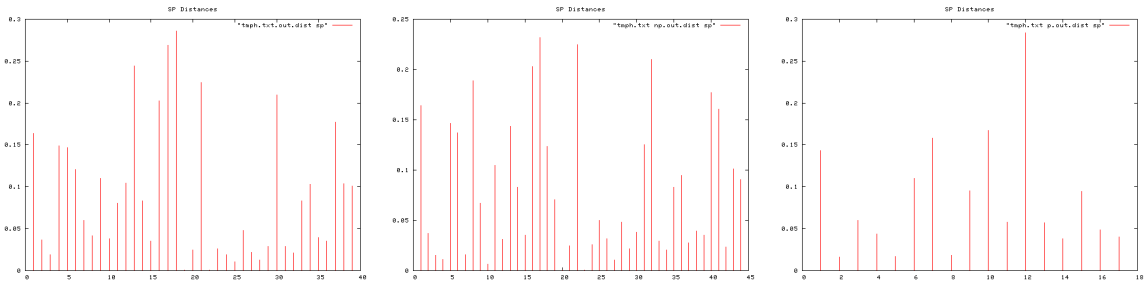



Figure B.232: Smooth pursuit distances, Trial 12. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

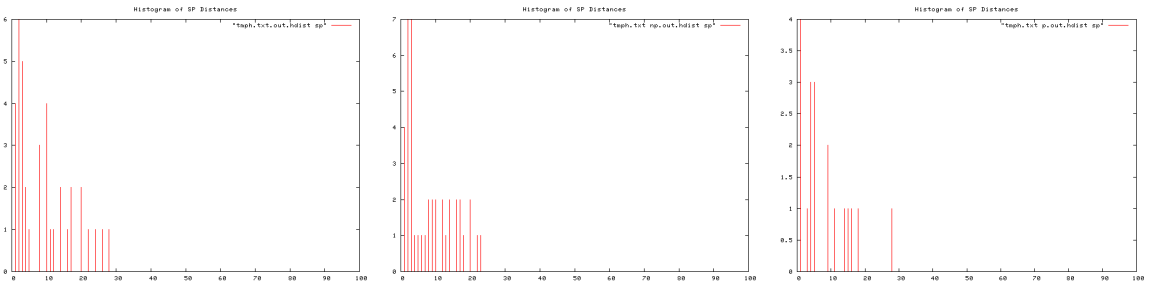



Figure B.233: Histogram of smooth pursuit distances, Trial 12. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

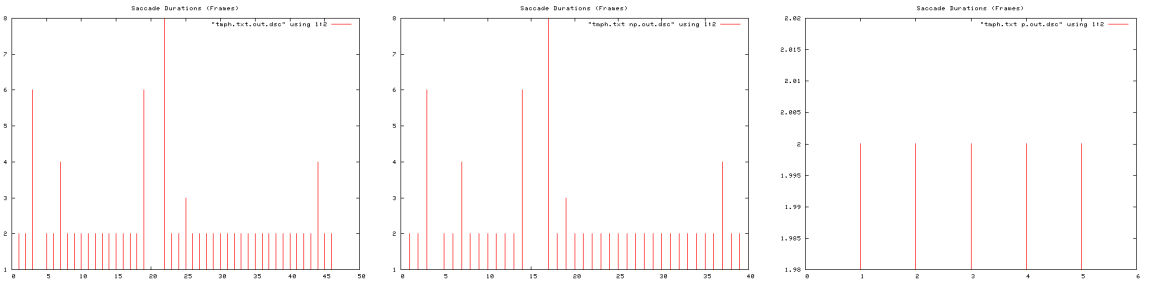



Figure B.234: Saccade durations, Trial 12. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

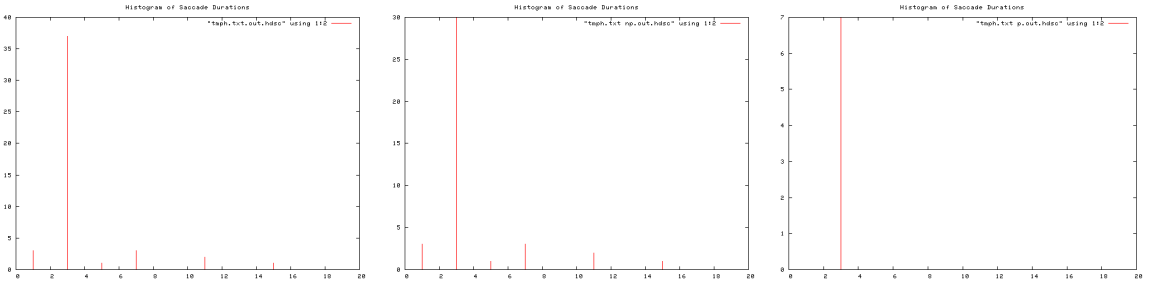



Figure B.235: Histogram of saccade durations, Trial 12. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

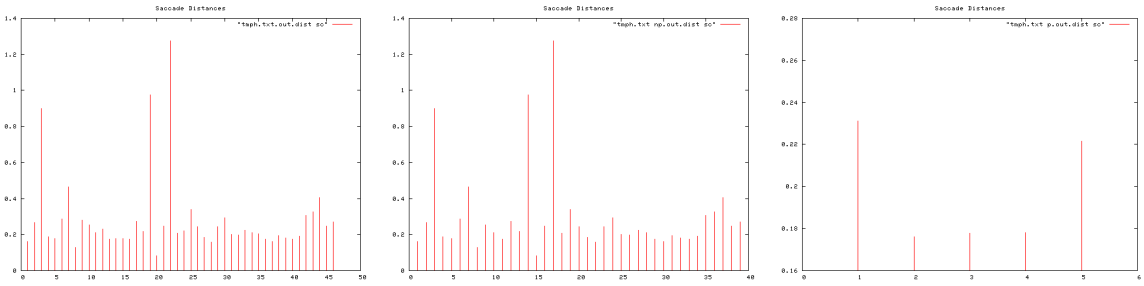



Figure B.236: Saccade distances, Trial 12. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

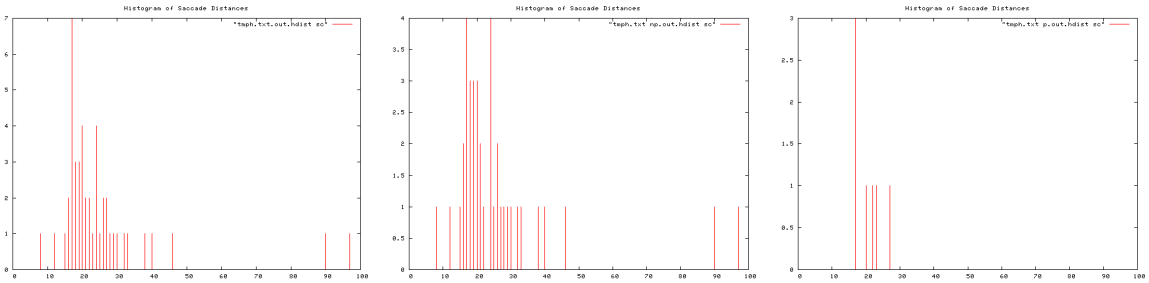



Figure B.237: Histogram of saccade distances, Trial 12. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

| Ti | To | T | I/O Event | Pear | Apple | Peach | Orange | |
|---|---|---|---|---|---|---|---|---|
| 8 | 16 | 8 | Orange In | | | | 2 | |
| 17 | 23 | 6 | Pear In | 1 | | | 0 | |
| 26 | 37 | 11 | | 2 | | | 3 | |
| 38 | 44 | 6 | Peach In | 0 | | 1 | 0 | |
| 45 | 0:55 | 10 | | 1 | | 2 | 1 | |
| 0:56 | 1:04 | 8 | | 2 | | 1 | 2 | |
| 1:04 | 1:14 | 10 | | 2 | | 2 | 3 | |
| 1:16 | 1:25 | 9 | Apple In, Peach Out | 1 | 2 | 1 | 1 | |
| 1:28 | 1:38 | 10 | Orange Out | 2 | 2 | | 1 | |
| 1:38 | 1:47 | 9 | Pear Out | 2 | 1 | | | |
| 1:49 | 1:55 | 6 | Apple Out | | 1 | | | |
| | | | TOTAL Rets | 13 | 6 | 7 | 13 | |
| | | | TOTAL T | 69 | 34 | 43 | 6 | **SD** |
| | | | **Av. Re-attention Period** | **5.3** | **5.7** | **6.1** | **6.6** | 0.55602758 |

Figure B.238: Re-attention period statistics, Trial 12.

### B.1.1.15 Trial 13

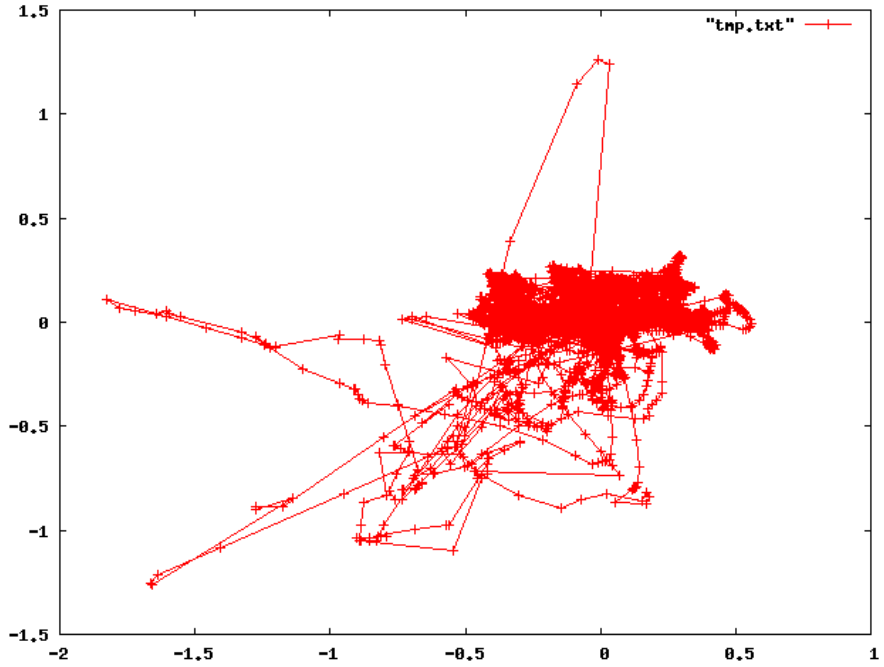



Figure B.239: Complete scan path, Trial 13.

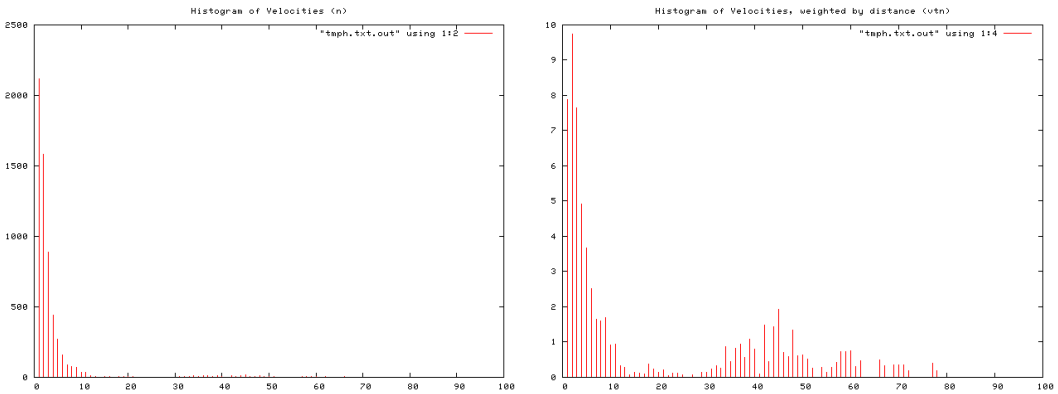



Figure B.240: Histogram of velocity magnitudes, Trial 13 (left). Histogram of distance weighted velocities, Trial 13 (right).

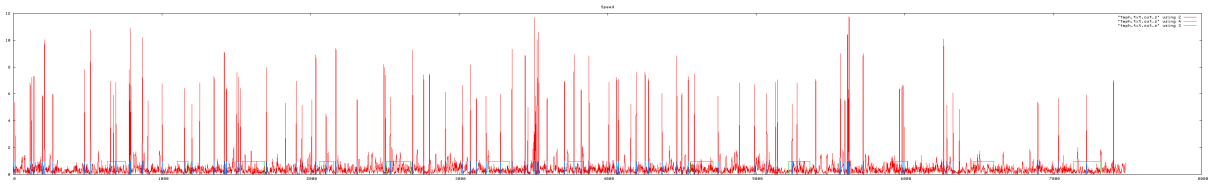


Figure B.241: Velocity profile. Velocity magnitude of each frame, Trial 13.

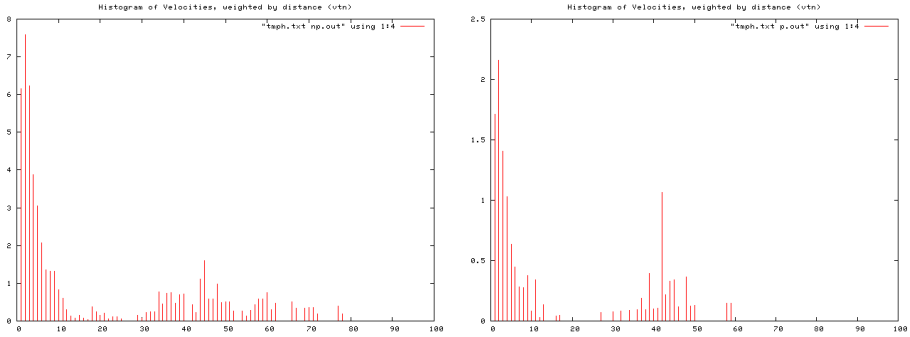



Figure B.242: Histogram of velocities during periods of no perturbation (left), and perturbation (right), Trial 13.

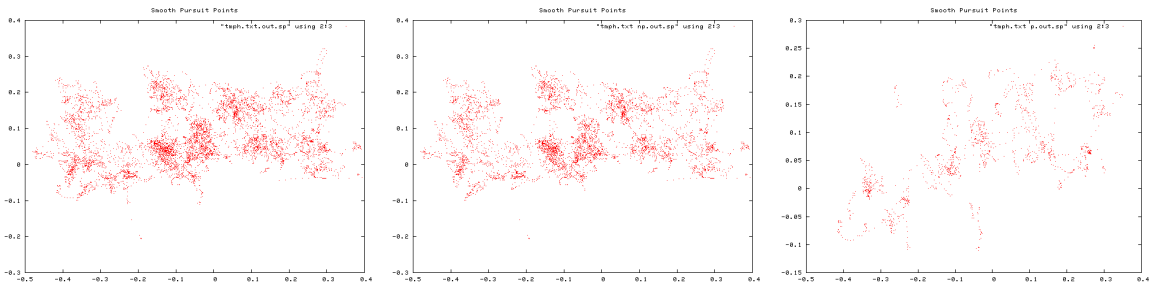



Figure B.243: Smooth pursuit gaze locations, Trial 13. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

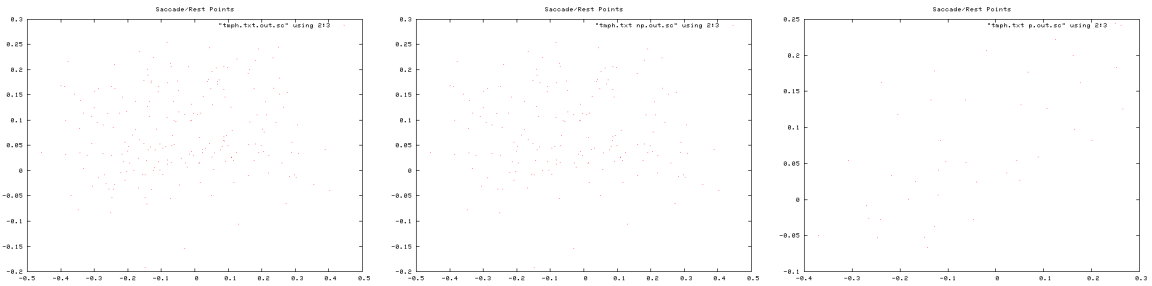



Figure B.244: Saccade gaze locations, Trial 13. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

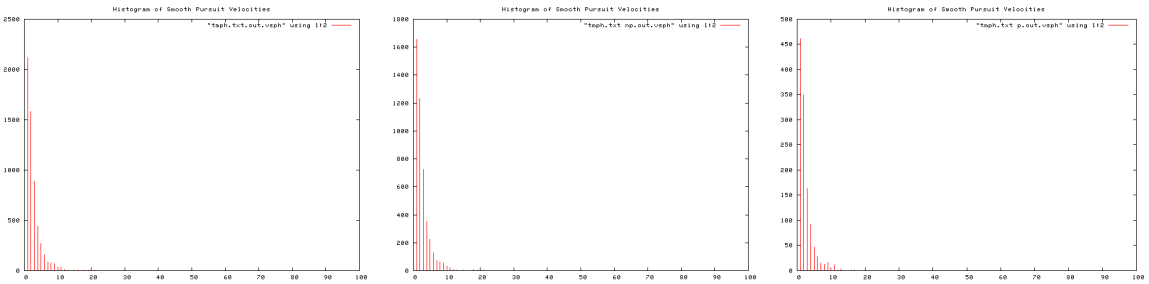



Figure B.245: Histogram of smooth pursuit velocities, Trial 13. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

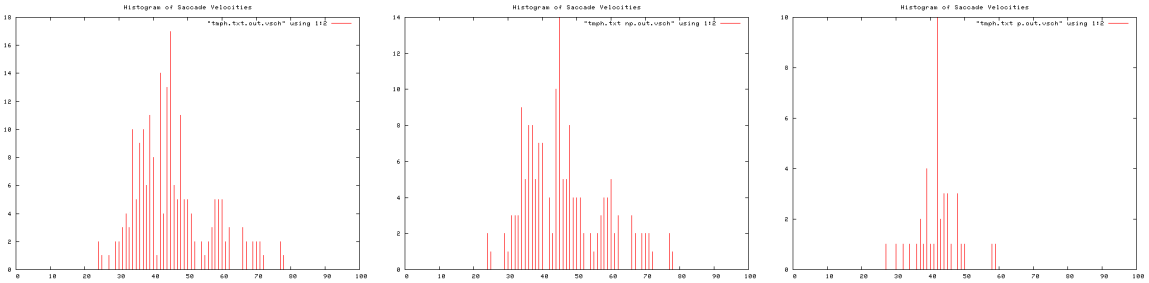



Figure B.246: Histogram of Saccade velocities, Trial 13. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

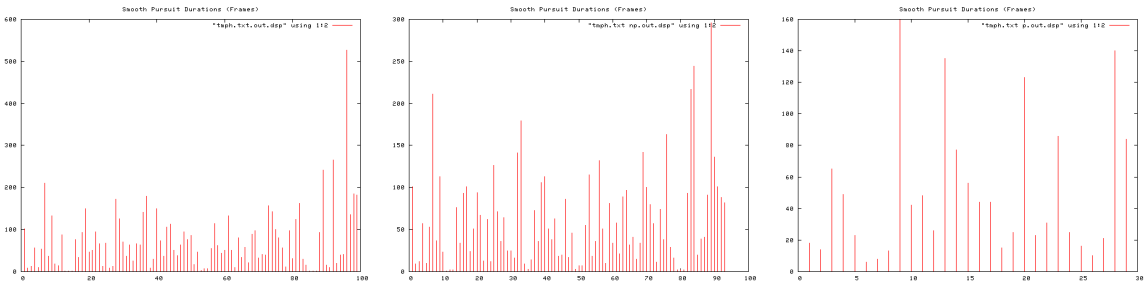



Figure B.247: Smooth pursuit durations, Trial 13. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

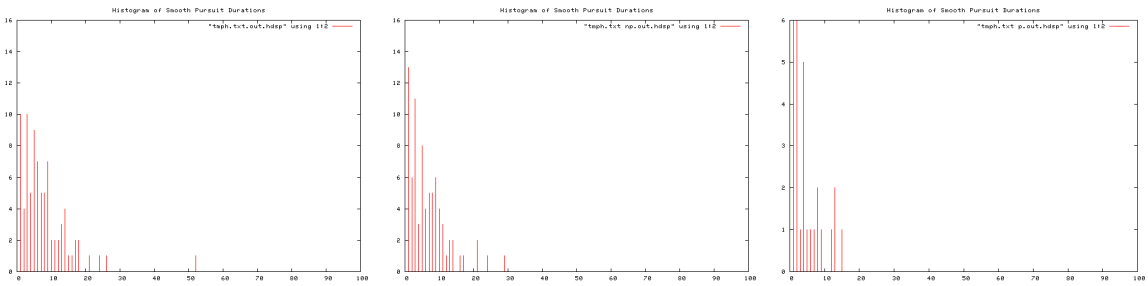



Figure B.248: Histogram of Smooth pursuit durations, Trial 13. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

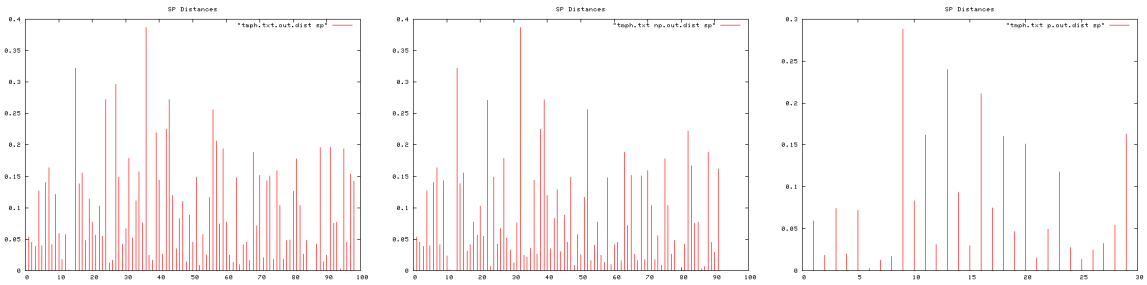



Figure B.249: Smooth pursuit distances, Trial 13. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

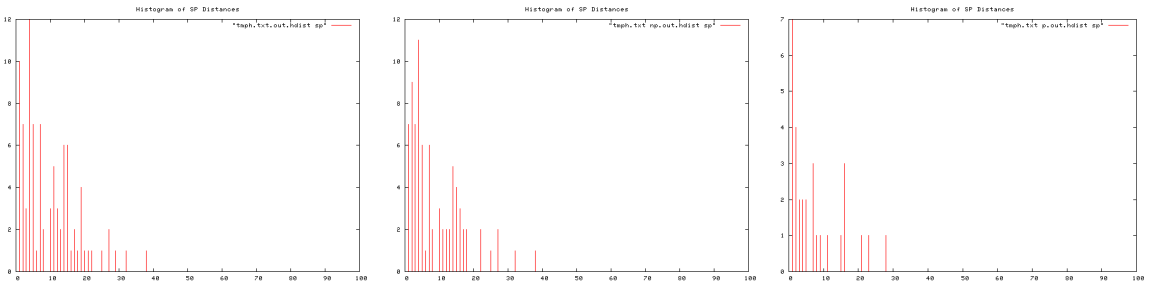



Figure B.250: Histogram of smooth pursuit distances, Trial 13. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

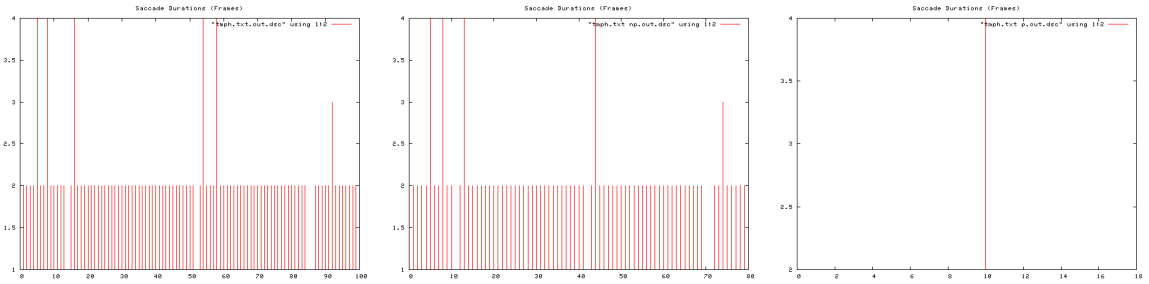



Figure B.251: Saccade durations, Trial 13. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

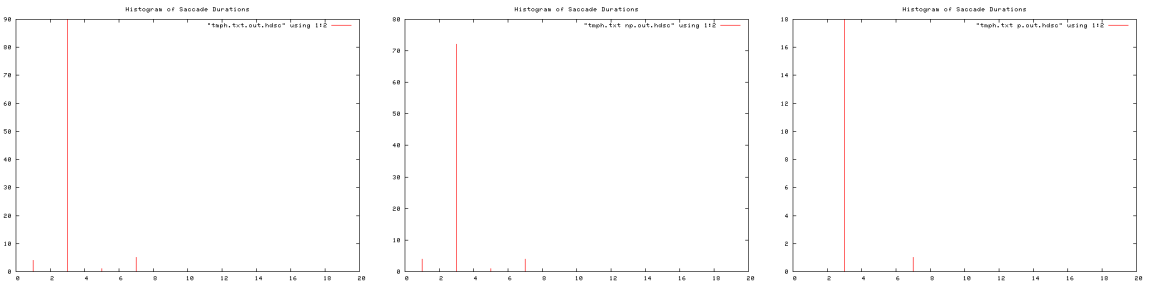



Figure B.252: Histogram of saccade durations, Trial 13. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

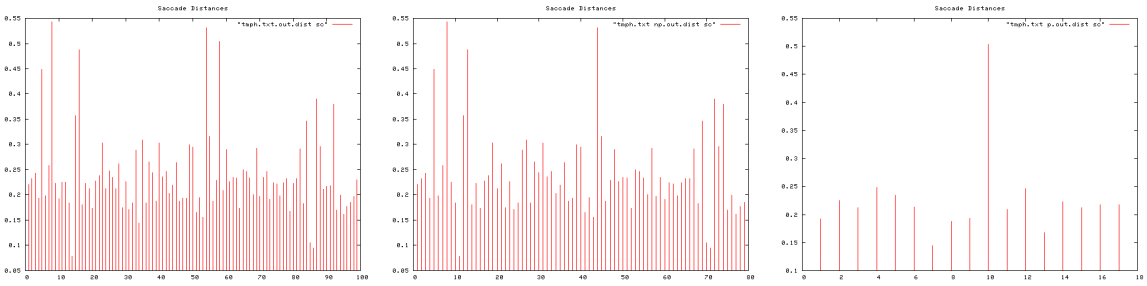



Figure B.253: Saccade distances, Trial 13. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

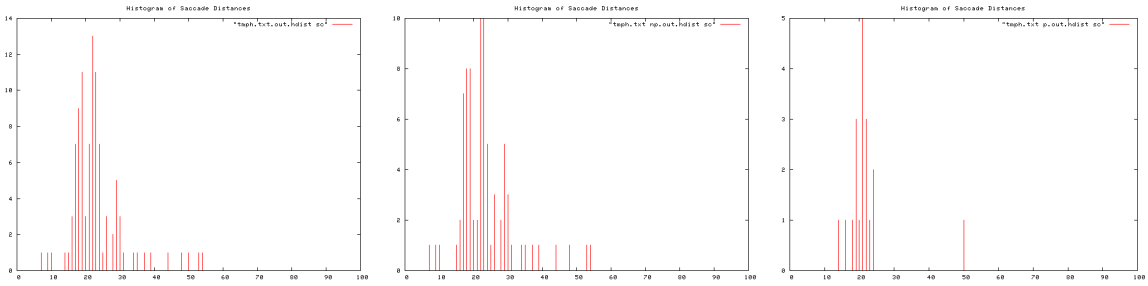



Figure B.254: Histogram of saccade distances, Trial 13. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

| Ti | To | T | I/O Event | Pear | Apple | Peach | Orange | |
|---|---|---|---|---|---|---|---|---|
| 11 | 18 | 7 | Orange In | | | | 2 | |
| 20 | 26 | 6 | Pear In | 1 | | | 1 | |
| 28 | 36 | 8 | | 2 | | | 2 | |
| 36 | 42 | 6 | Peach In | 1 | | 1 | 1 | |
| 43 | 0:53 | 10 | | 2 | | 2 | 2 | |
| 0:55 | 1:03 | 8 | | 1 | | 0 | 1 | |
| 1:04 | 1:17 | 13 | | 2 | | 2 | 2 | |
| 1:18 | 1:27 | 9 | Apple In, Peach Out | 1 | 2 | 1 | 1 | |
| 1:29 | 1:39 | 10 | Orange Out | 2 | 1 | | 2 | |
| 1:41 | 1:48 | 7 | Pear Out | 1 | 1 | | | |
| 1:50 | 2:00 | 10 | Apple Out | | 2 | | | |
| | | | TOTAL Rets | 13 | 6 | 6 | 14 | |
| | | | TOTAL T | 77 | 36 | 37 | 77 | **SD** |
| | | | **Av. Re-attention Period** | **5.9** | **6** | **6.1** | **5.5** | 0.26299556 |

Figure B.255: Re-attention period statistics, Trial 13.

### B.1.1.16 Trial 14

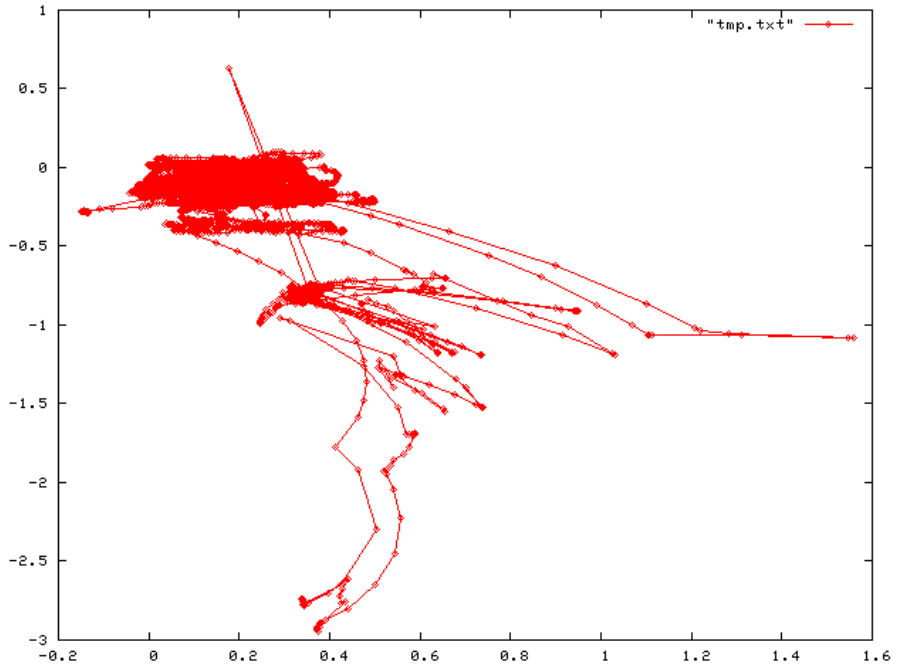



Figure B.256: Complete scan path, Trial 14.

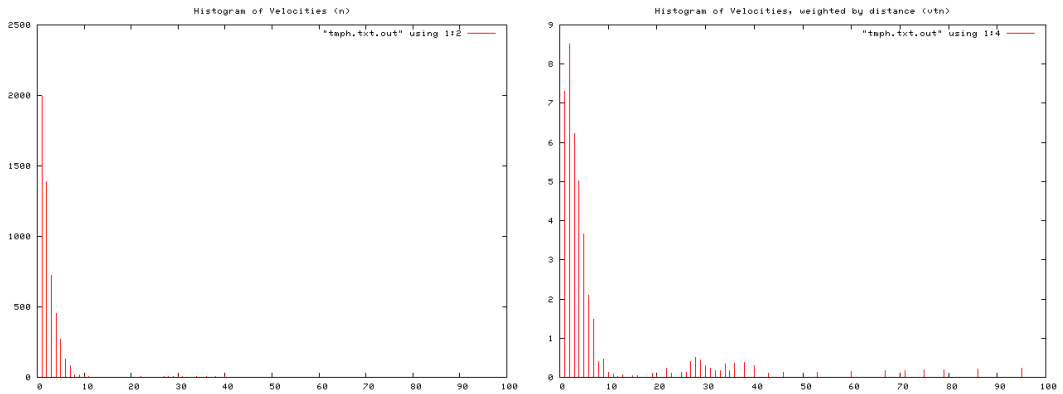



Figure B.257: Histogram of velocity magnitudes, Trial 14 (left). Histogram of distance weighted velocities, Trial 14 (right).

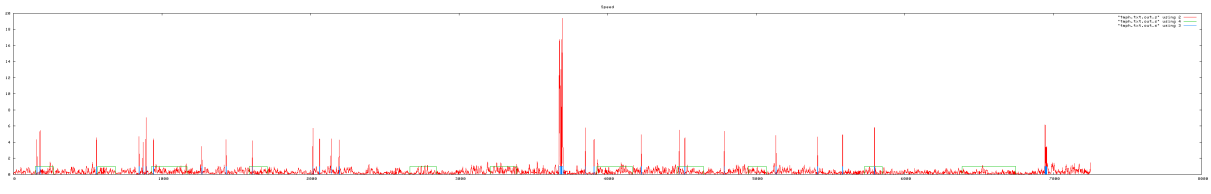


Figure B.258: Velocity profile. Velocity magnitude of each frame, Trial 14.

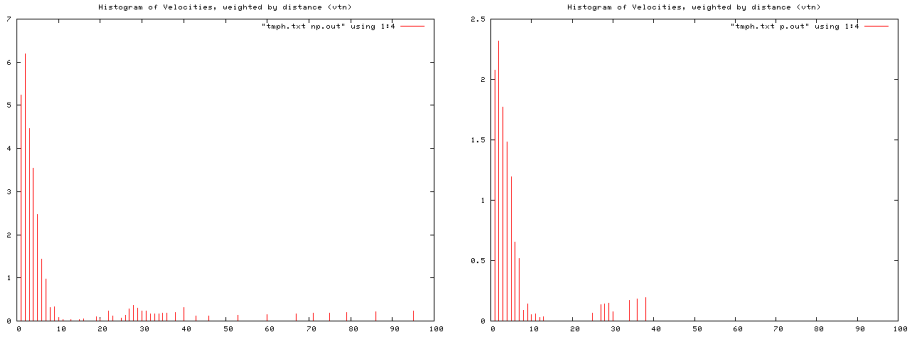



Figure B.259: Histogram of velocities during periods of no perturbation (left), and perturbation (right), Trial 14.

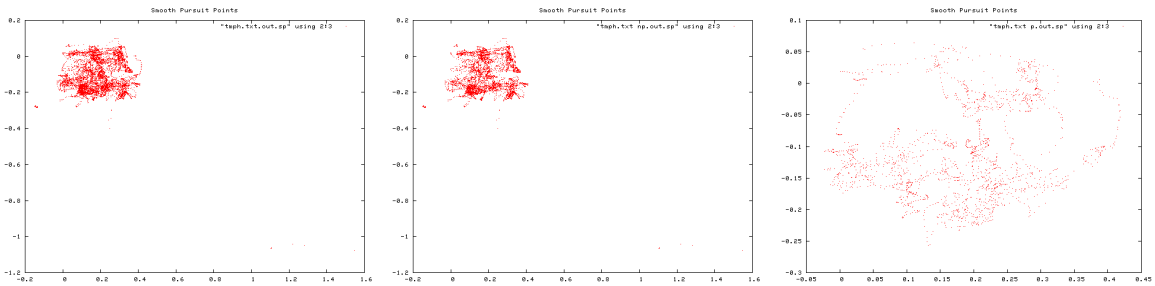



Figure B.260: Smooth pursuit gaze locations, Trial 14. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

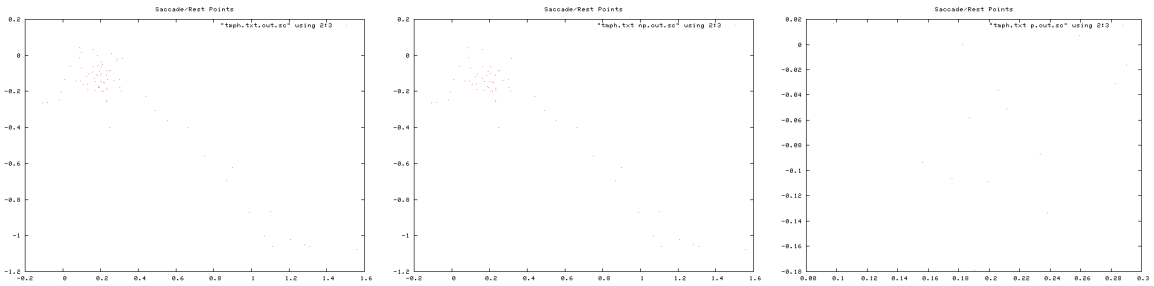



Figure B.261: Saccade gaze locations, Trial 14. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

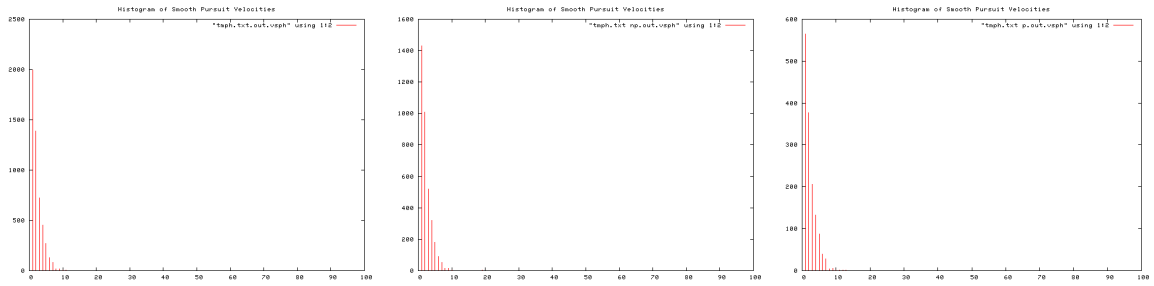



Figure B.262: Histogram of smooth pursuit velocities, Trial 14. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

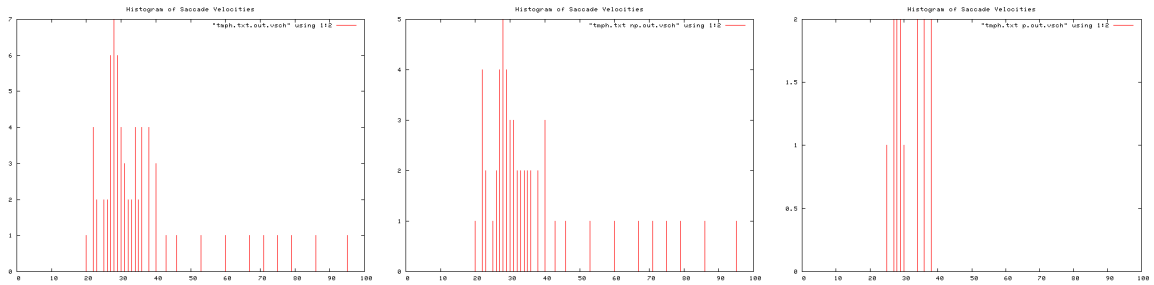



Figure B.263: Histogram of Saccade velocities, Trial 14. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

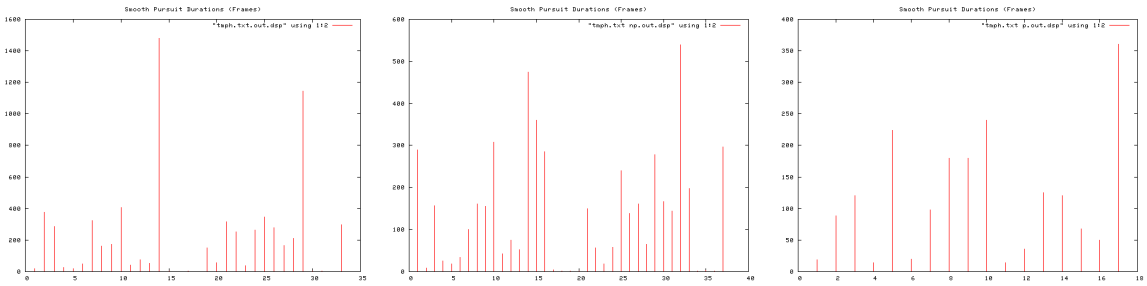



Figure B.264: Smooth pursuit durations, Trial 14. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

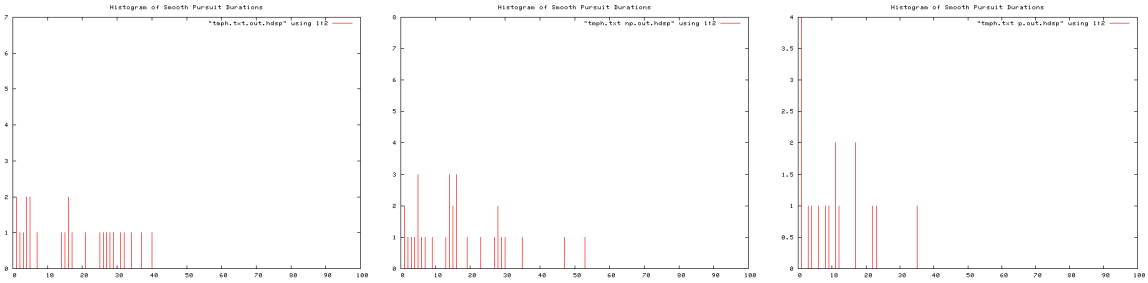



Figure B.265: Histogram of Smooth pursuit durations, Trial 14. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

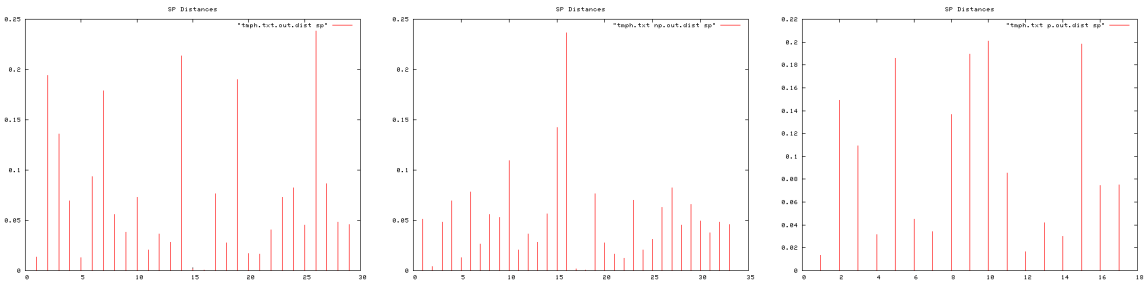



Figure B.266: Smooth pursuit distances, Trial 14. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

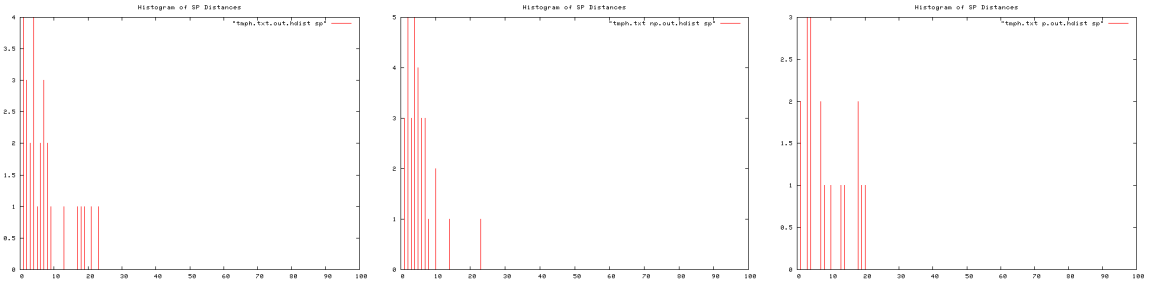



Figure B.267: Histogram of smooth pursuit distances, Trial 14. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

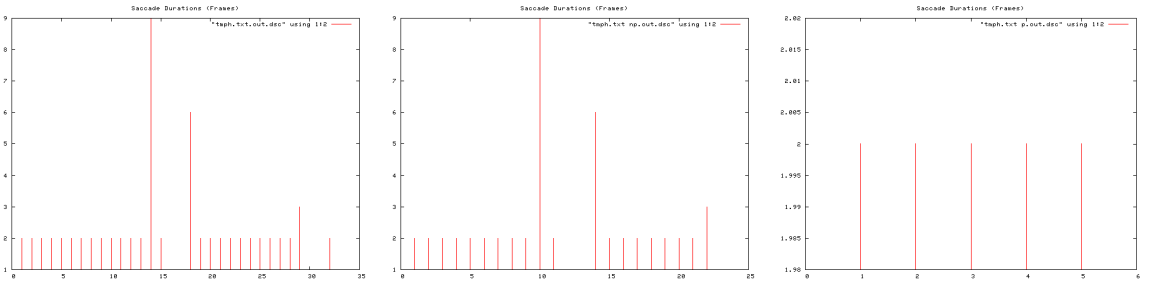



Figure B.268: Saccade durations, Trial 14. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

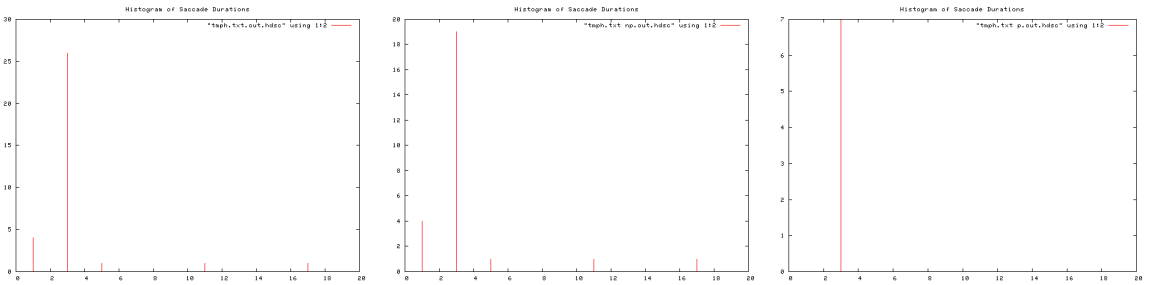



Figure B.269: Histogram of saccade durations, Trial 14. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

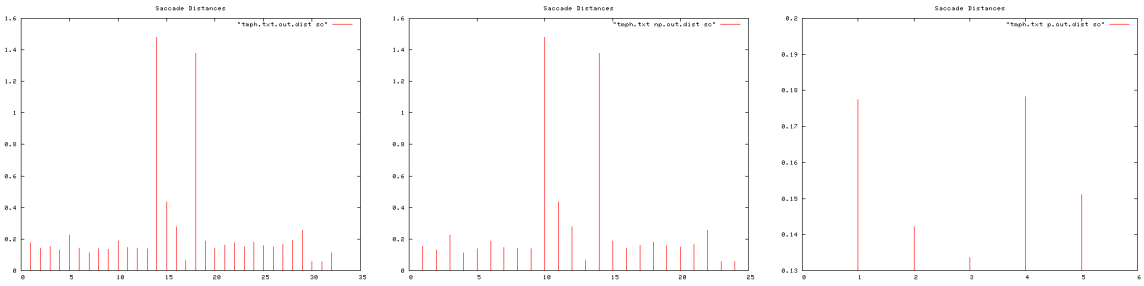



Figure B.270: Saccade distances, Trial 14. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

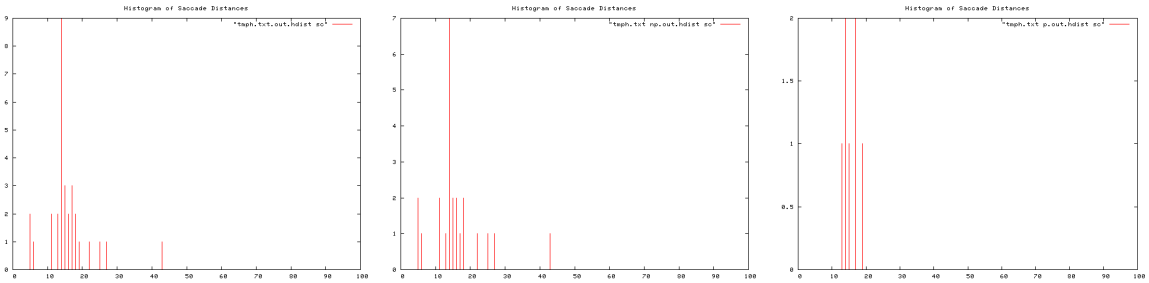



Figure B.271: Histogram of saccade distances, Trial 14. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

| Ti | To | T | I/O Event | Pear | Apple | Peach | Orange | |
|---|---|---|---|---|---|---|---|---|
| 27 | 33 | 6 | Orange In | | | | 1 | |
| 34 | 38 | 4 | Pear In | 1 | | | 0 | |
| 41 | 49 | 8 | | 1 | | | 2 | |
| 50 | 57 | 7 | Peach In | 2 | | 1 | 1 | |
| 59 | 1:08 | 9 | | 2 | | 2 | 0 | |
| 1:10 | 1:16 | 5 | | 0 | | 1 | 1 | |
| 1:18 | 1:28 | 10 | | 2 | | 2 | 2 | |
| 1:30 | 1:38 | 8 | Apple In, Peach Out | 1 | 2 | 1 | 2 | |
| 1:40 | 1:47 | 7 | Orange Out | 1 | 1 | | 1 | |
| 1:49 | 1:58 | 9 | Pear Out | 2 | 2 | | | |
| 2:00 | 2:11 | 11 | Apple Out | | 1 | | | |
| | | | TOTAL Rets | 12 | 6 | 7 | 10 | |
| | | | TOTAL T | 67 | 35 | 39 | 64 | **SD** |
| | | | **Av. Re-attention Period** | **5.6** | **5.8** | **5.6** | **6.4** | 0.37859389 |

Figure B.272: Re-attention period statistics, Trial 14.

### B.1.1.17 Trial 15

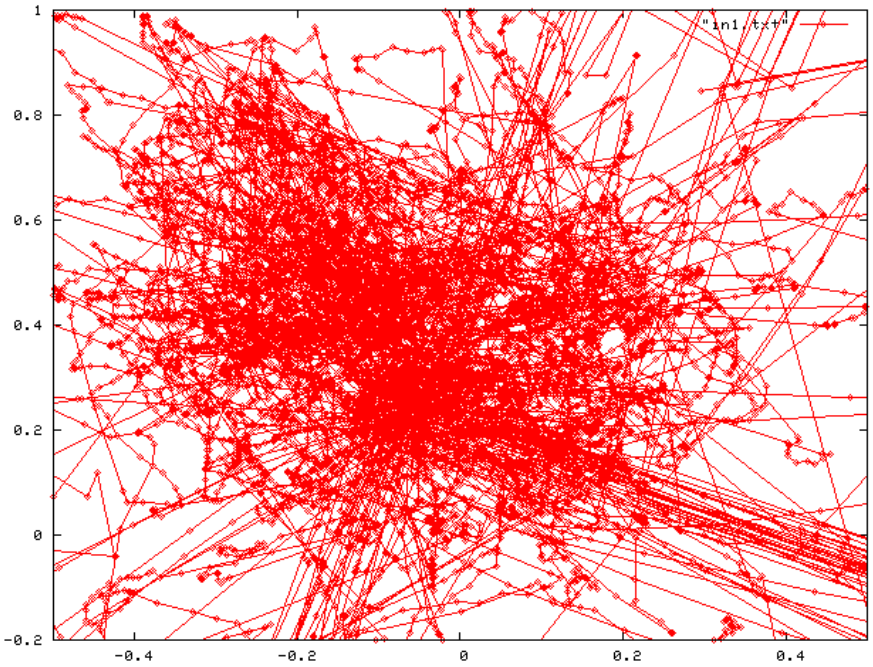



Figure B.273: Complete scan path, Trial 15.

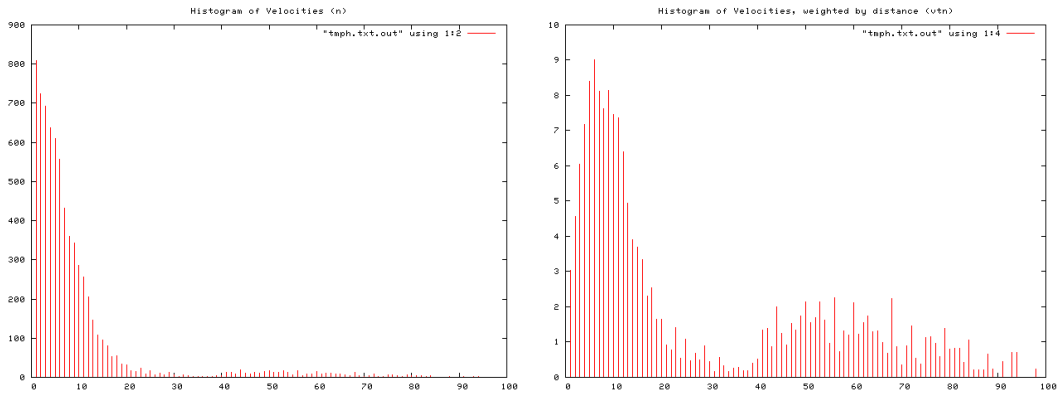



Figure B.274: Histogram of velocity magnitudes, Trial 15 (left). Histogram of distance weighted velocities, Trial 15 (right).

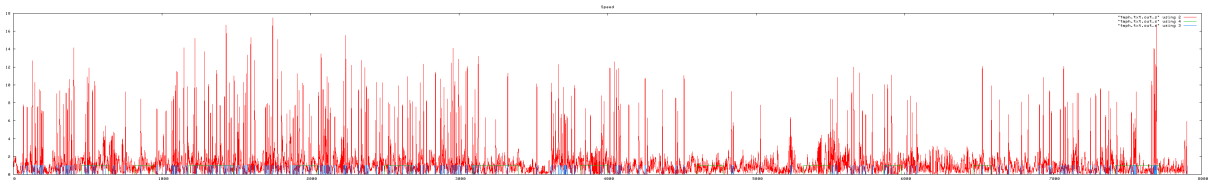


Figure B.275: Velocity profile. Velocity magnitude of each frame, Trial 15.

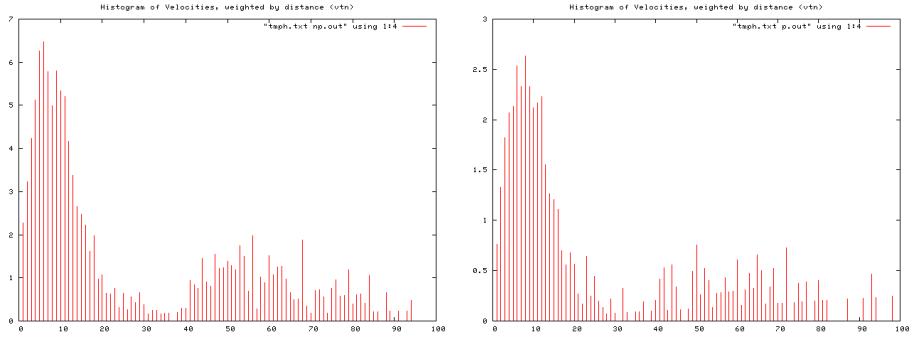



Figure B.276: Histogram of velocities during periods of no perturbation (left), and perturbation (right), Trial 15.

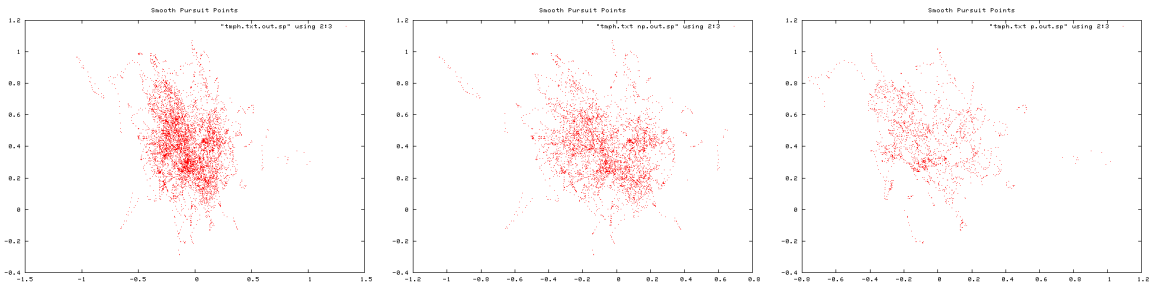



Figure B.277: Smooth pursuit gaze locations, Trial 15. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

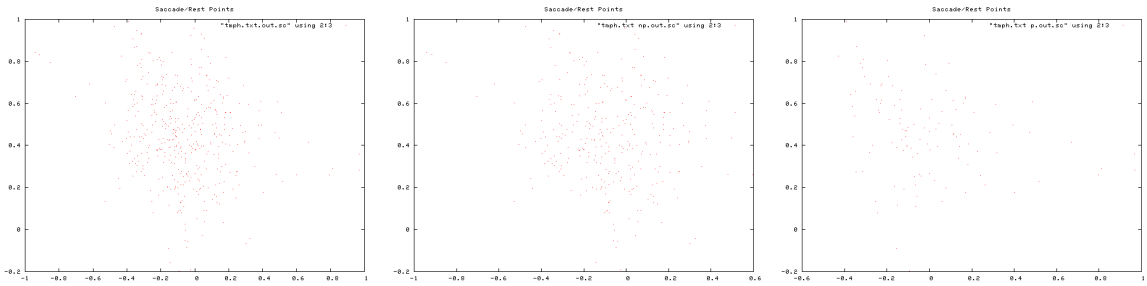



Figure B.278: Saccade gaze locations, Trial 15. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

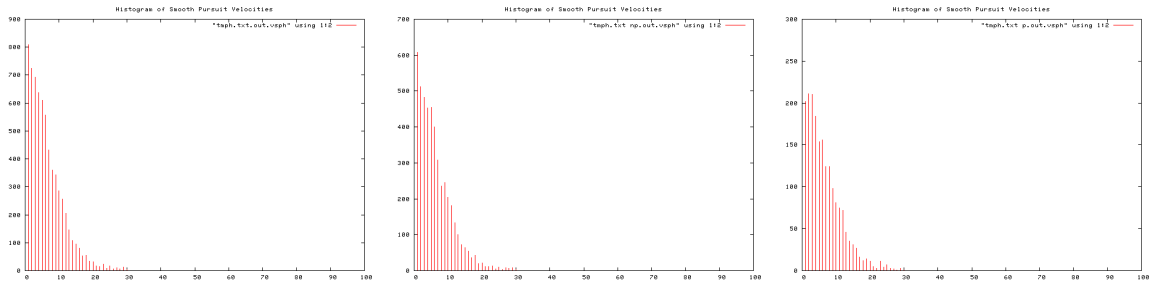



Figure B.279: Histogram of smooth pursuit velocities, Trial 15. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

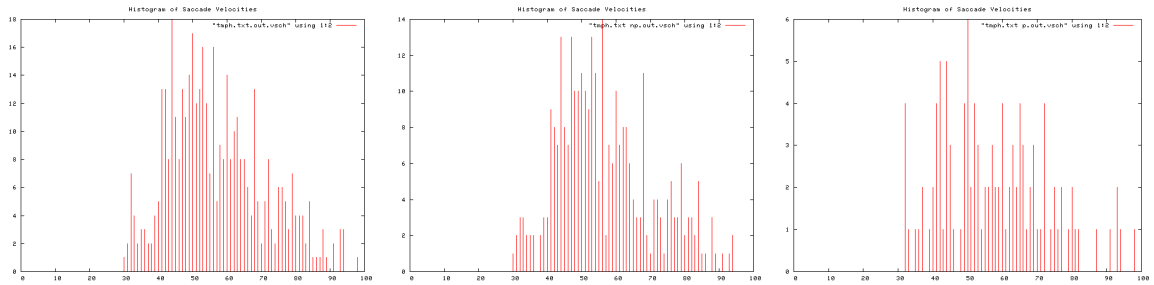



Figure B.280: Histogram of Saccade velocities, Trial 15. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

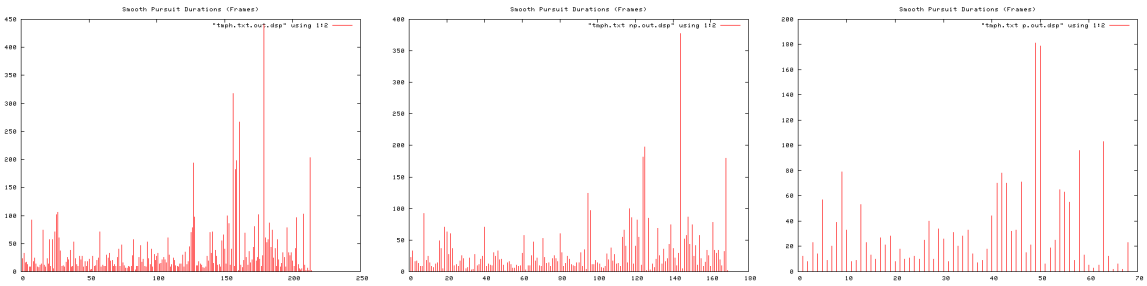



Figure B.281: Smooth pursuit durations, Trial 15. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

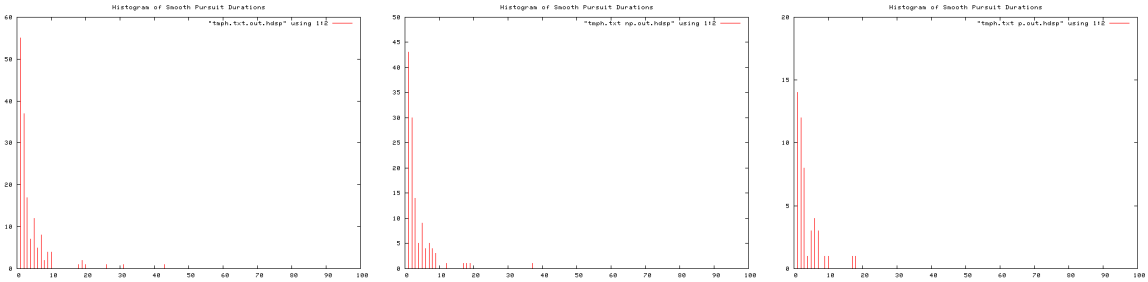



Figure B.282: Histogram of Smooth pursuit durations, Trial 15. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

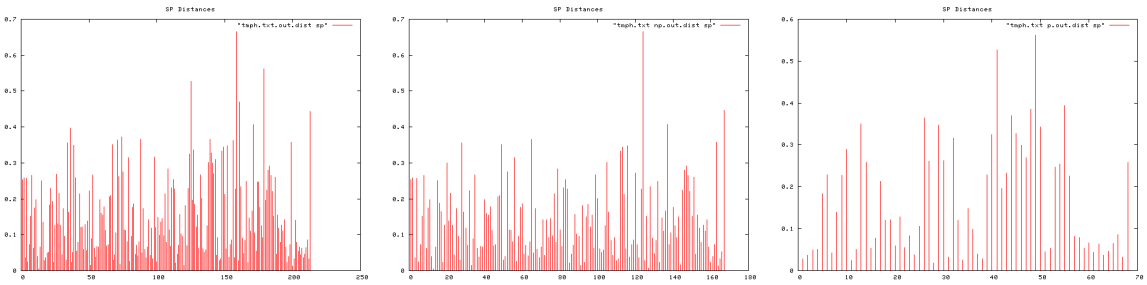



Figure B.283: Smooth pursuit distances, Trial 15. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

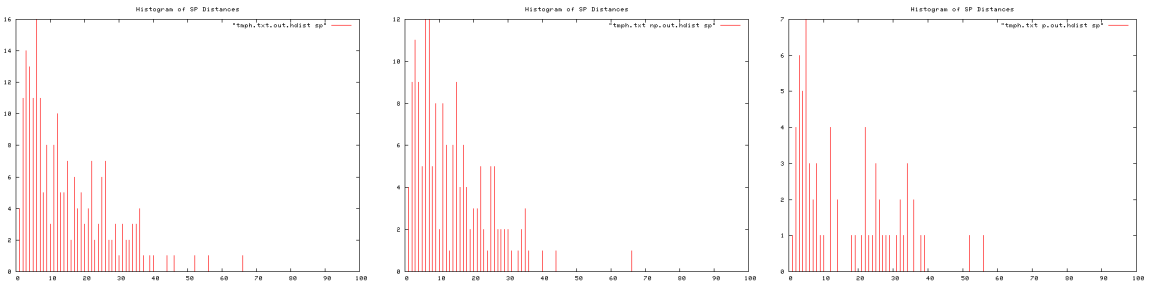



Figure B.284: Histogram of smooth pursuit distances, Trial 15. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

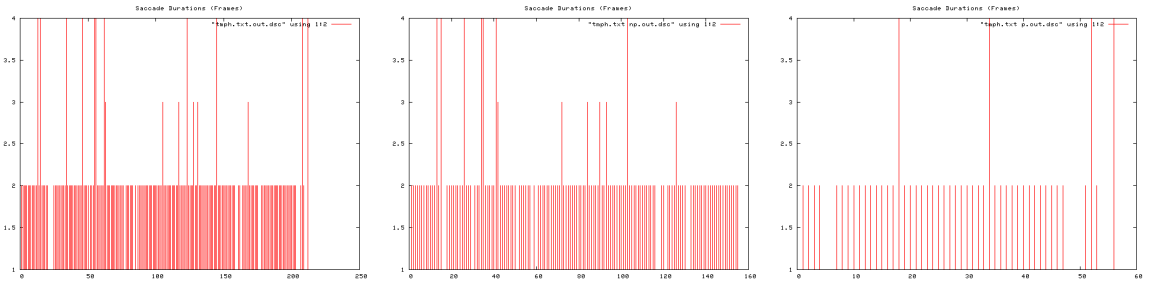



Figure B.285: Saccade durations, Trial 15. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

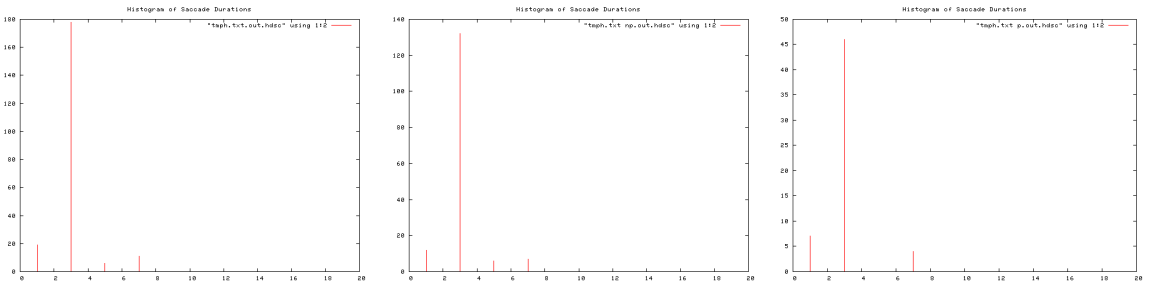



Figure B.286: Histogram of saccade durations, Trial 15. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

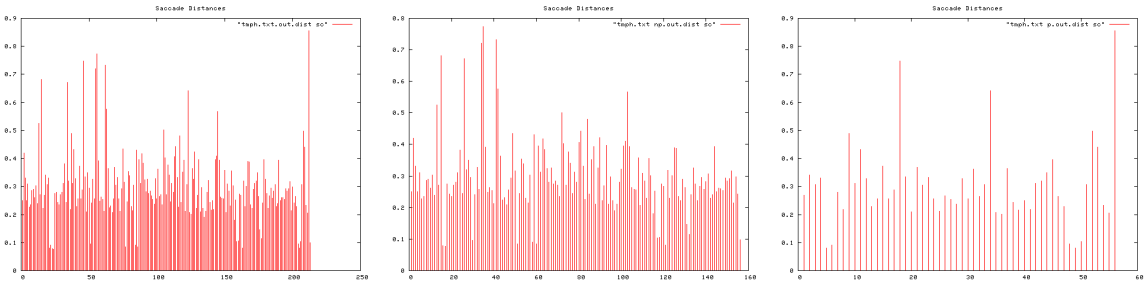



Figure B.287: Saccade distances, Trial 15. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

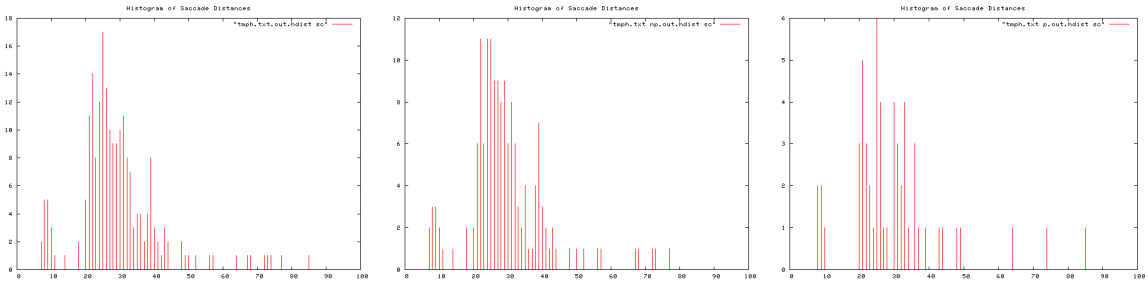



Figure B.288: Histogram of saccade distances, Trial 15. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

| Ti | To | T | I/O Event | Pear | Apple | Peach | Orange | |
|---|---|---|---|---|---|---|---|---|
| 18 | 22 | 4 | Orange In | | | | 1 | |
| 24 | 28 | 4 | Pear In | 2 | | | 2 | |
| 33 | 40 | 7 | | 1 | | | 1 | |
| 41 | 49 | 8 | Peach In | 2 | | 1 | 2 | |
| 52 | 1:02 | 10 | | 2 | | 2 | 1 | |
| 1:04 | 1:13 | 9 | | 1 | | 2 | 2 | |
| 1:14 | 1:25 | 11 | | 1 | | 2 | 1 | |
| 1:28 | 1:37 | 9 | Apple In, Peach Out | 1 | 1 | 1 | 1 | |
| 1:40 | 1:47 | 7 | Orange Out | 1 | 1 | | 1 | |
| 1:48 | 1:56 | 8 | Pear Out | 1 | 2 | | | |
| 1:59 | 2:12 | 13 | Apple Out | | 2 | | | |
| | | | TOTAL Rets | 12 | 6 | 8 | 12 | |
| | | | TOTAL T | 73 | 37 | 47 | 69 | **SD** |
| | | | **Av. Re-attention Period** | **6.1** | **6.2** | **5.9** | **5.8** | 0.18257419 |

Figure B.289: Re-attention period statistics, Trial 15.

### B.1.1.18 Trial 16

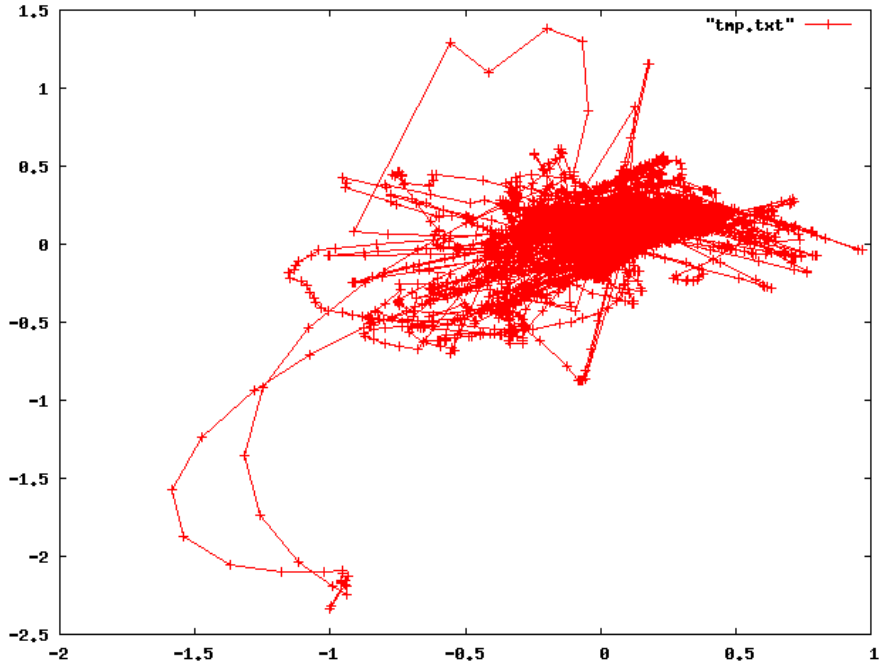



Figure B.290: Complete scan path, Trial 16.

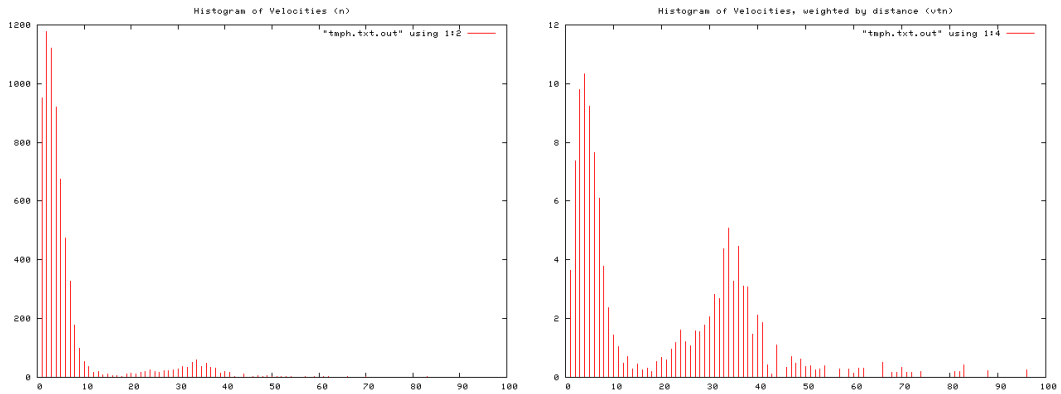



Figure B.291: Histogram of velocity magnitudes, Trial 16 (left). Histogram of distance weighted velocities, Trial 16 (right).

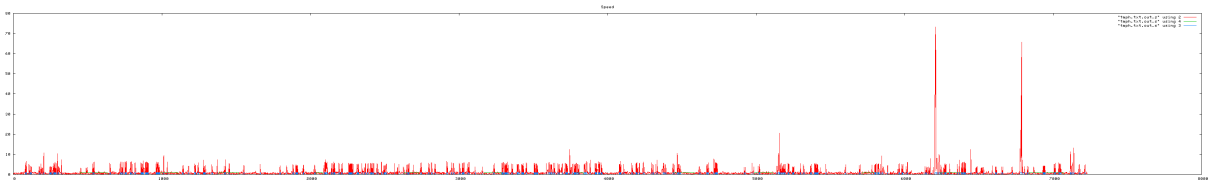


Figure B.292: Velocity profile. Velocity magnitude of each frame, Trial 16.

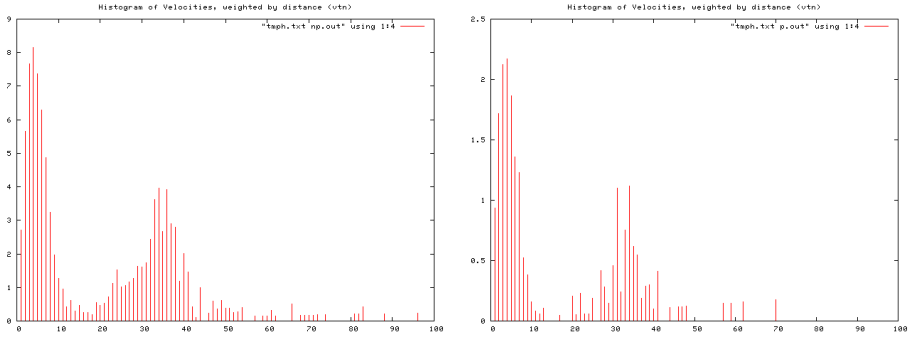



Figure B.293: Histogram of velocities during periods of no perturbation (left), and perturbation (right), Trial 16.

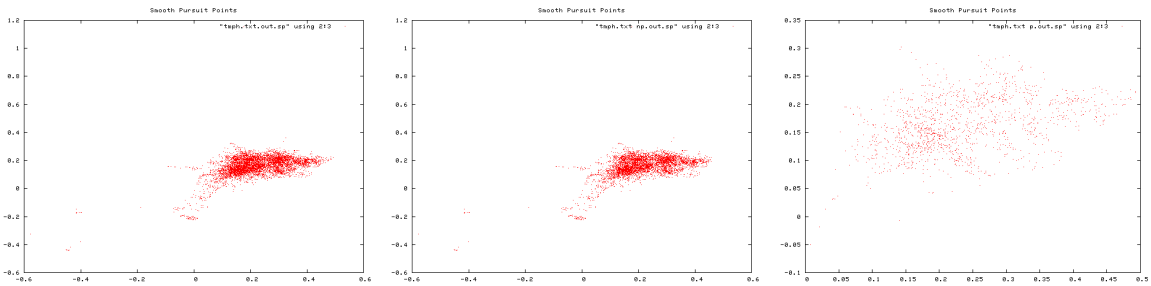



Figure B.294: Smooth pursuit gaze locations, Trial 16. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

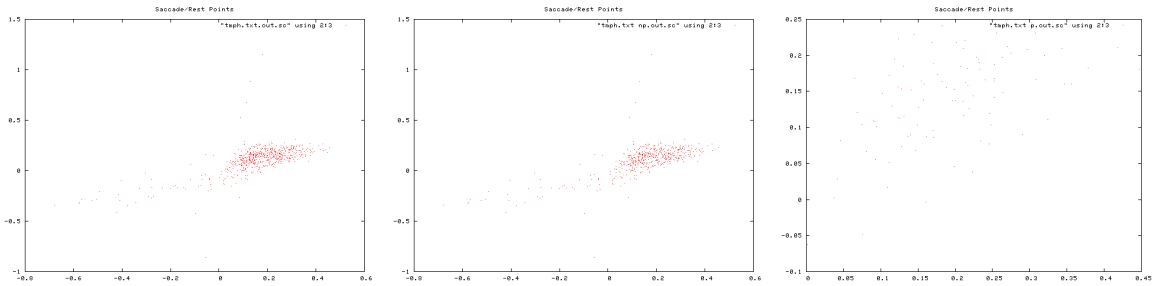



Figure B.295: Saccade gaze locations, Trial 16. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

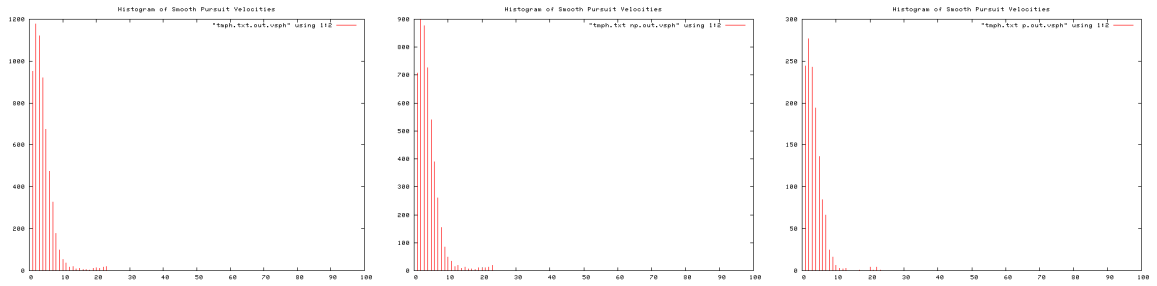



Figure B.296: Histogram of smooth pursuit velocities, Trial 16. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

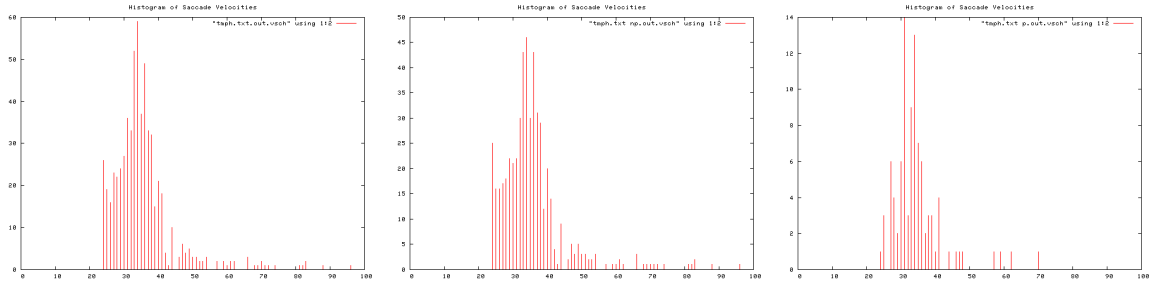



Figure B.297: Histogram of Saccade velocities, Trial 16. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

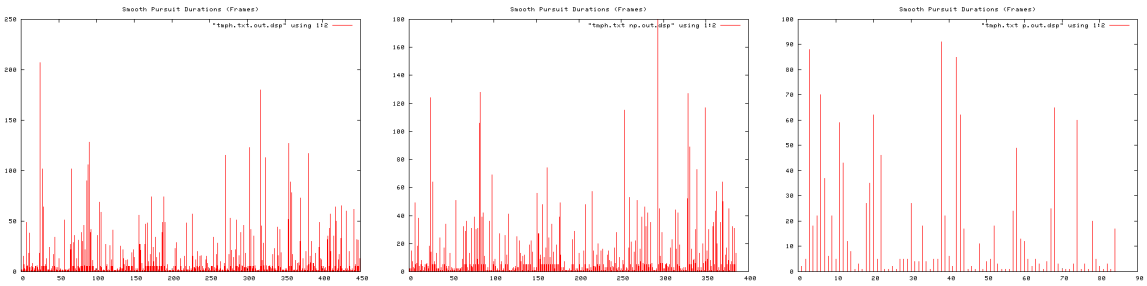



Figure B.298: Smooth pursuit durations, Trial 16. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

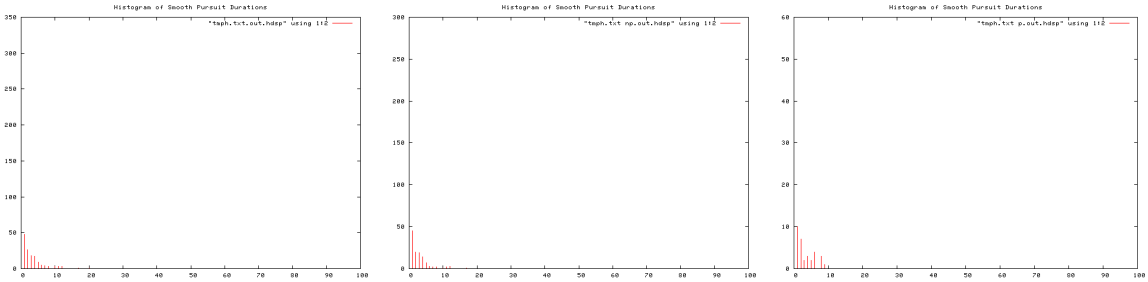



Figure B.299: Histogram of Smooth pursuit durations, Trial 16. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

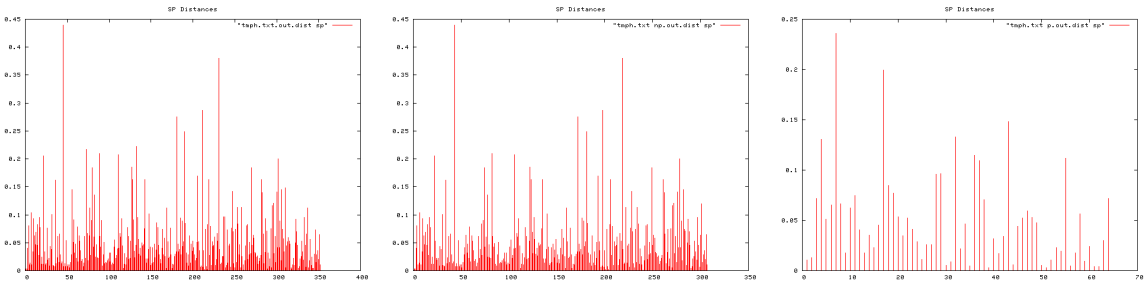



Figure B.300: Smooth pursuit distances, Trial 16. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

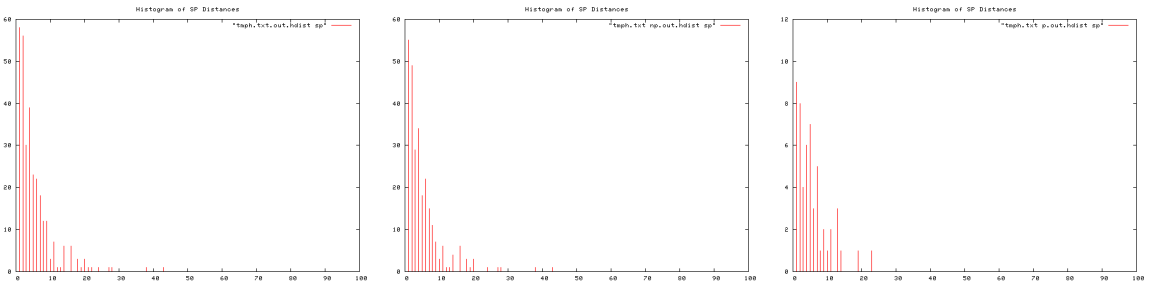



Figure B.301: Histogram of smooth pursuit distances, Trial 16. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

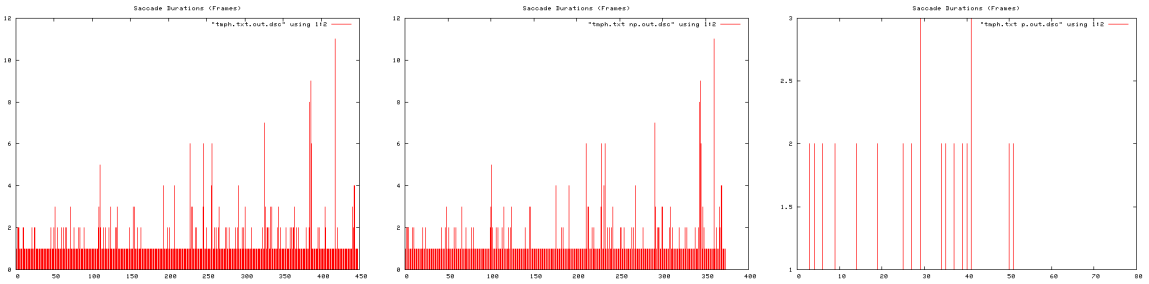



Figure B.302: Saccade durations, Trial 16. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

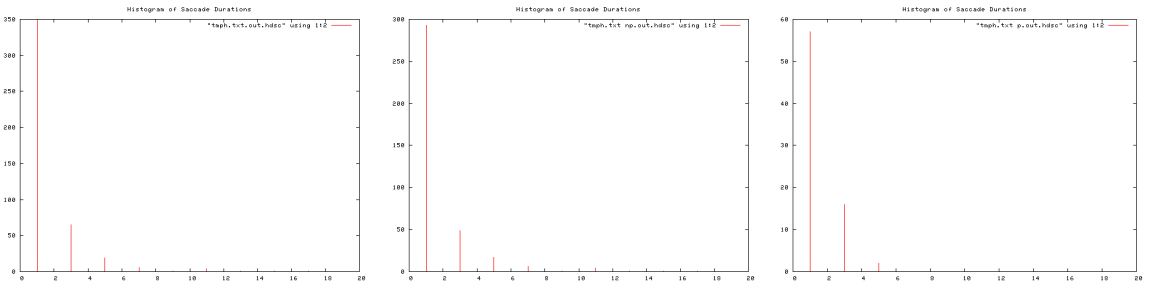



Figure B.303: Histogram of saccade durations, Trial 16. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

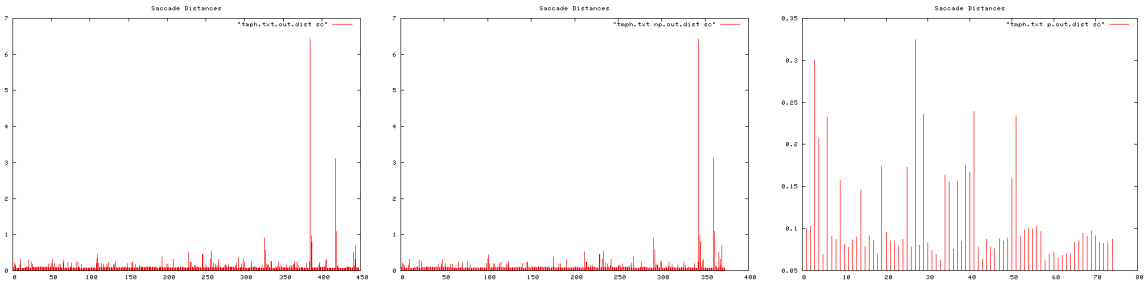



Figure B.304: Saccade distances, Trial 16. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

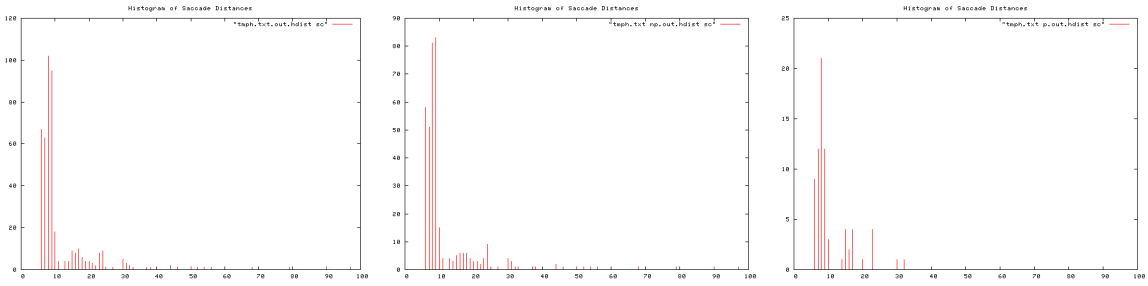



Figure B.305: Histogram of saccade distances, Trial 16. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

| Ti | To | T | I/O Event | Pear | Apple | Peach | Orange | |
|---|---|---|---|---|---|---|---|---|
| 9 | 16 | 7 | Orange In | | | | 1 | |
| 18 | 23 | 5 | Pear In | 2 | | | 2 | |
| 26 | 35 | 9 | | 5 | | | 4 | |
| 35 | 44 | 9 | Peach In | 1 | | 2 | 1 | |
| 45 | 0:53 | 8 | | 1 | | 1 | 1 | |
| 0:54 | 1:03 | 9 | | 2 | | 3 | 3 | |
| 1:04 | 1:14 | 10 | | 1 | | 4 | 4 | |
| 1:15 | 1:23 | 8 | Apple In, Peach Out | 1 | 2 | 1 | 2 | |
| 1:25 | 1:36 | 11 | Orange Out | 1 | 4 | | 4 | |
| 1:36 | 1:45 | 9 | Pear Out | 4 | 3 | | | |
| 1:47 | 1:54 | 7 | Apple Out | | 2 | | | |
| | | | TOTAL Rets | 18 | 11 | 11 | 22 | |
| | | | TOTAL T | 78 | 35 | 44 | 76 | **SD** |
| | | | **Av. Re-attention Period** | **4.3** | **3.2** | **4** | **3.6** | 0.47871355 |

Figure B.306: Re-attention period statistics, Trial 16.

### B.1.1.19 Trial 17

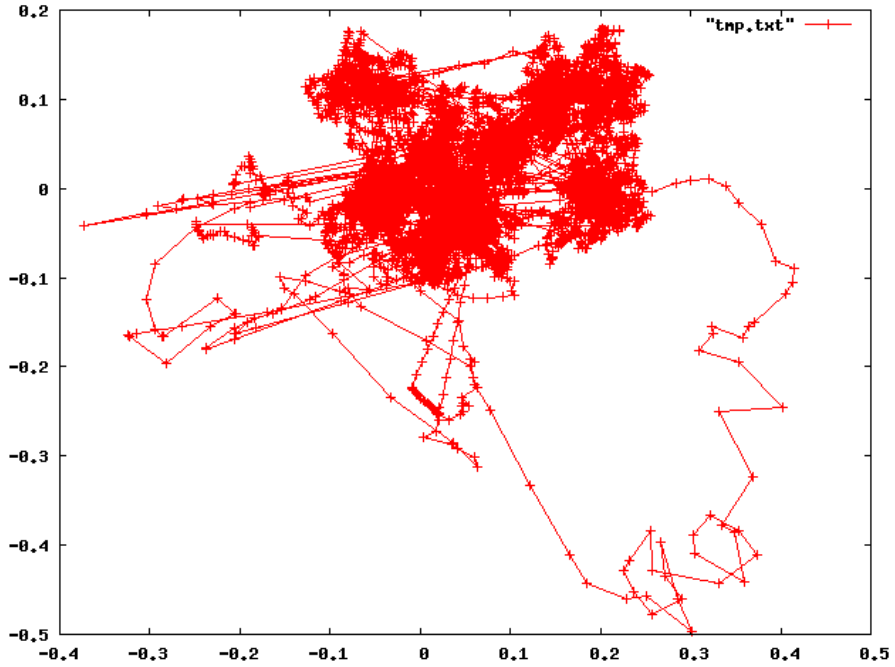



Figure B.307: Complete scan path, Trial 17.

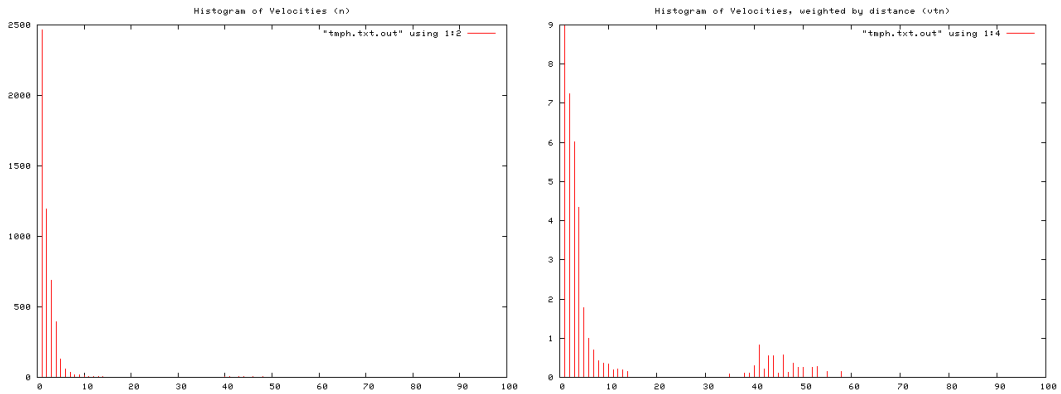



Figure B.308: Histogram of velocity magnitudes, Trial 17 (left). Histogram of distance weighted velocities, Trial 17 (right).

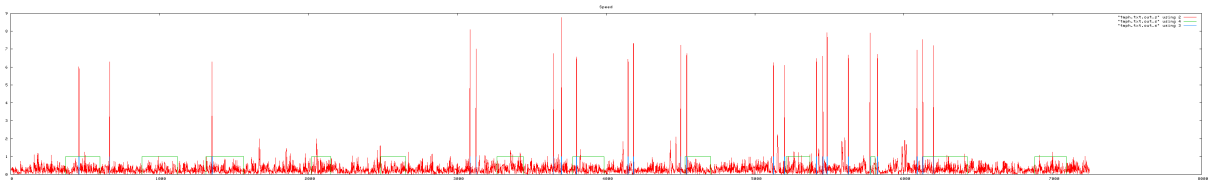


Figure B.309: Velocity profile. Velocity magnitude of each frame, Trial 17.

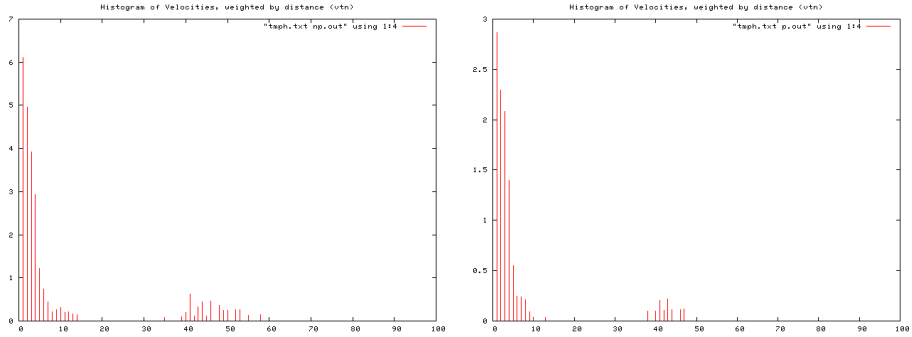



Figure B.310: Histogram of velocities during periods of no perturbation (left), and perturbation (right), Trial 17.

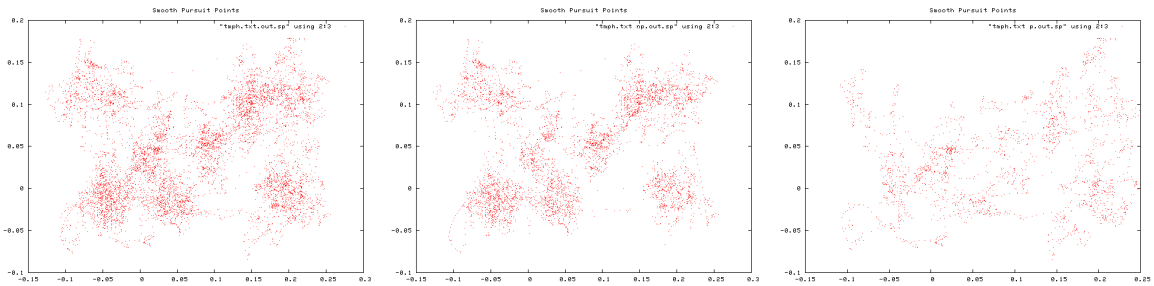



Figure B.311: Smooth pursuit gaze locations, Trial 17. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

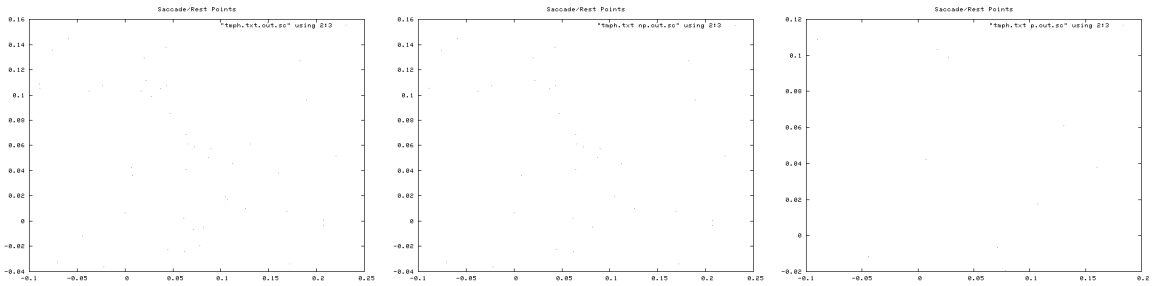



Figure B.312: Saccade gaze locations, Trial 17. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

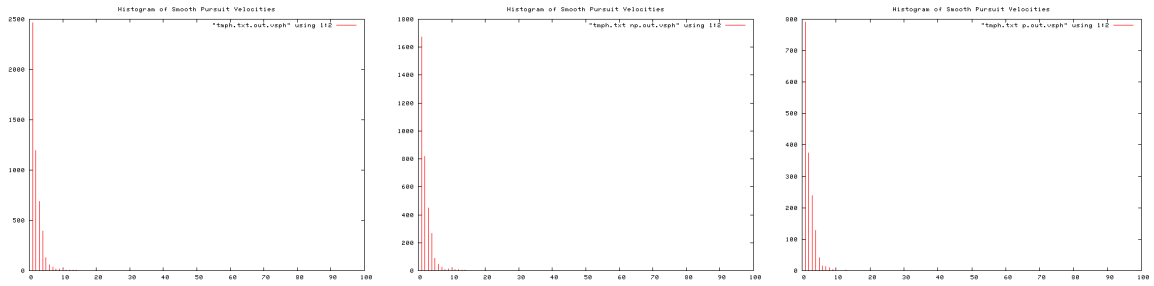



Figure B.313: Histogram of smooth pursuit velocities, Trial 17. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

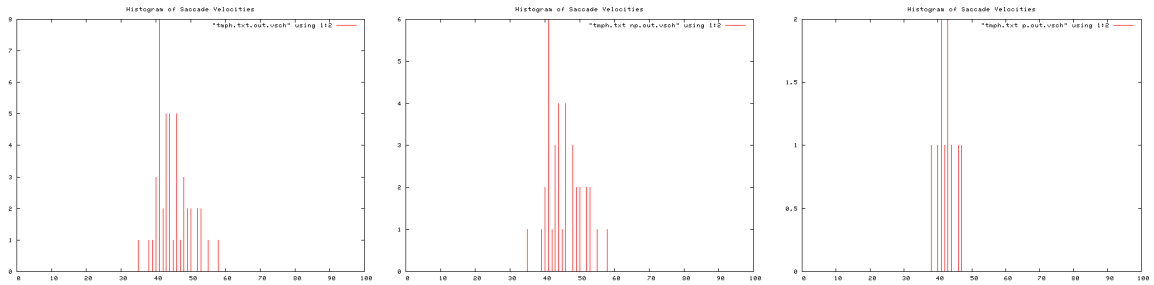



Figure B.314: Histogram of Saccade velocities, Trial 17. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

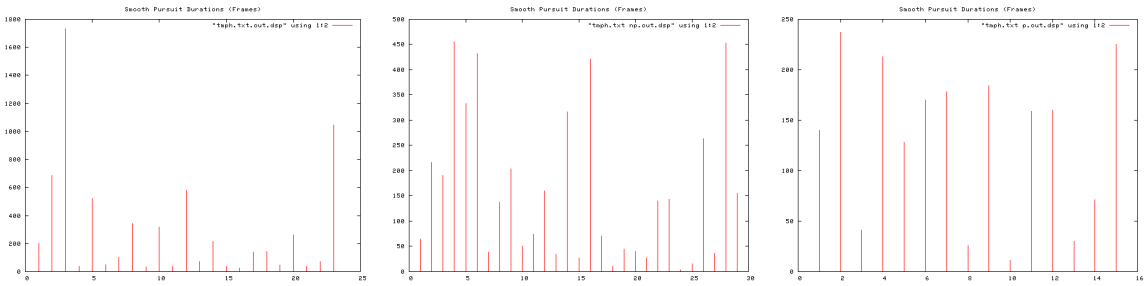



Figure B.315: Smooth pursuit durations, Trial 17. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

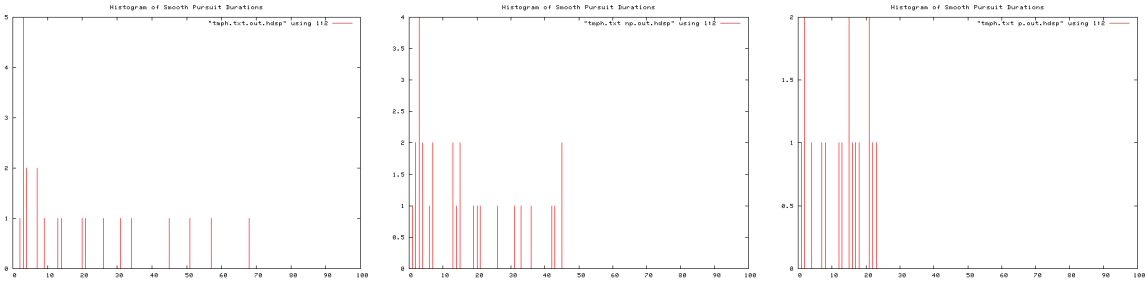



Figure B.316: Histogram of Smooth pursuit durations, Trial 17. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

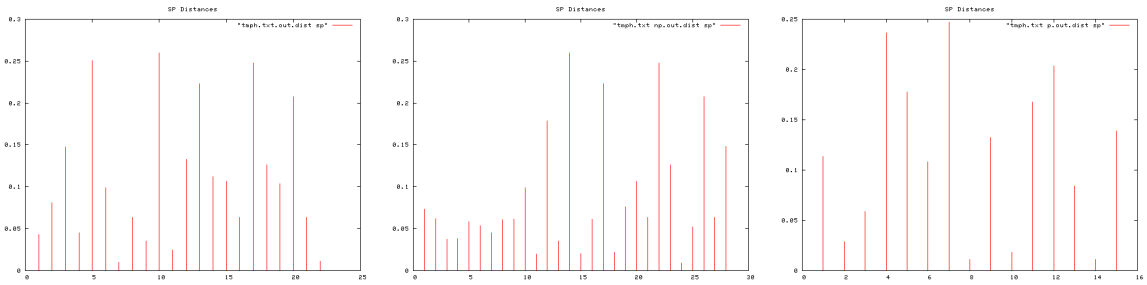



Figure B.317: Smooth pursuit distances, Trial 17. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

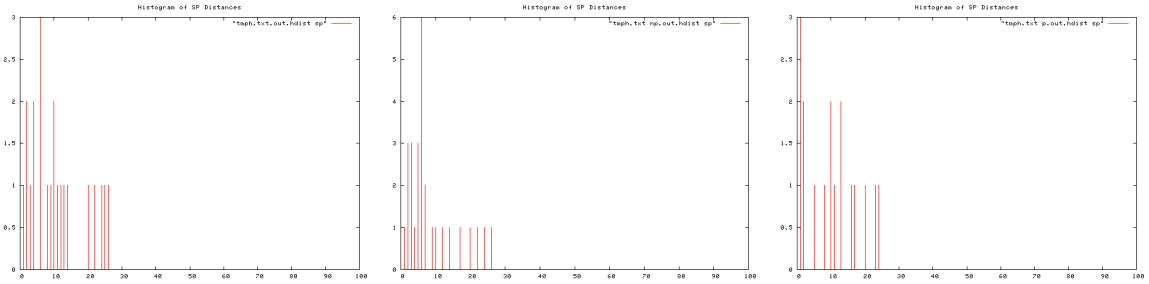



Figure B.318: Histogram of smooth pursuit distances, Trial 17. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

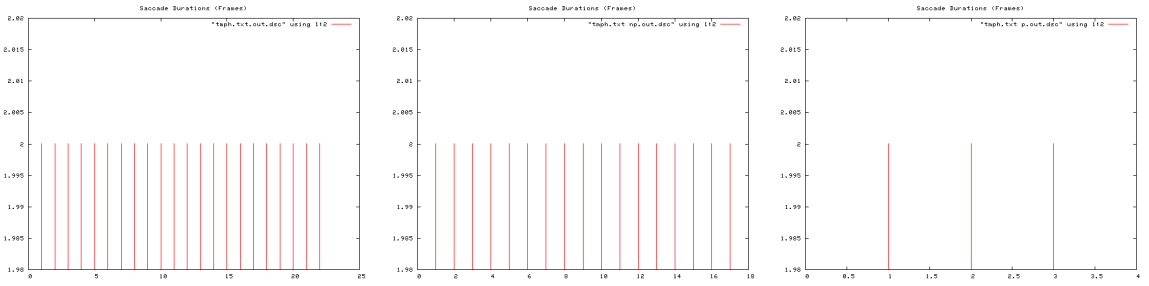



Figure B.319: Saccade durations, Trial 17. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

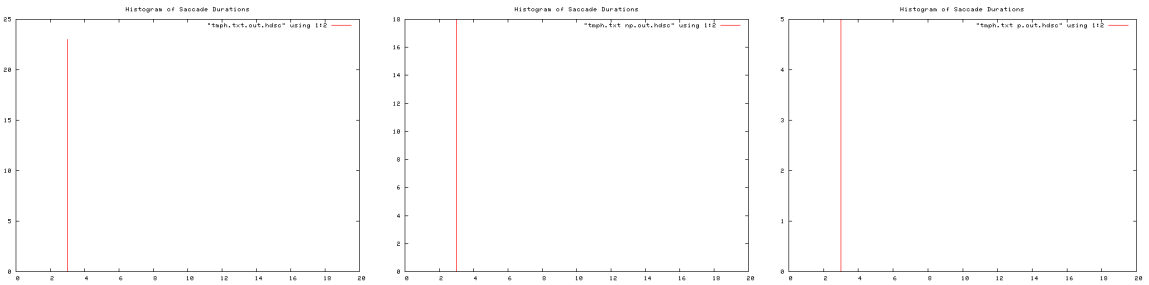



Figure B.320: Histogram of saccade durations, Trial 17. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

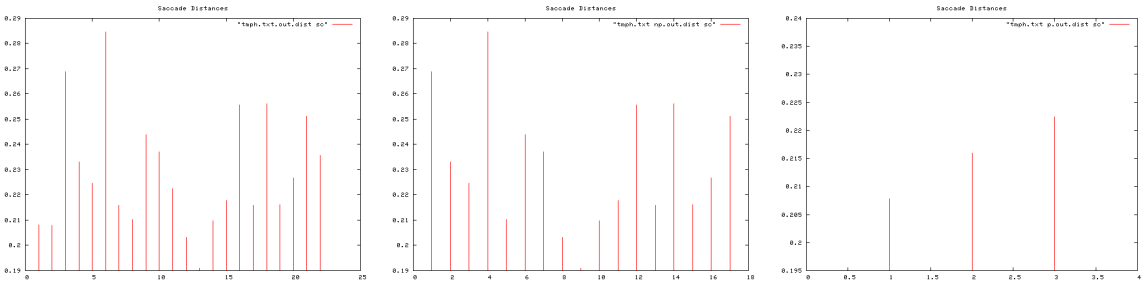



Figure B.321: Saccade distances, Trial 17. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

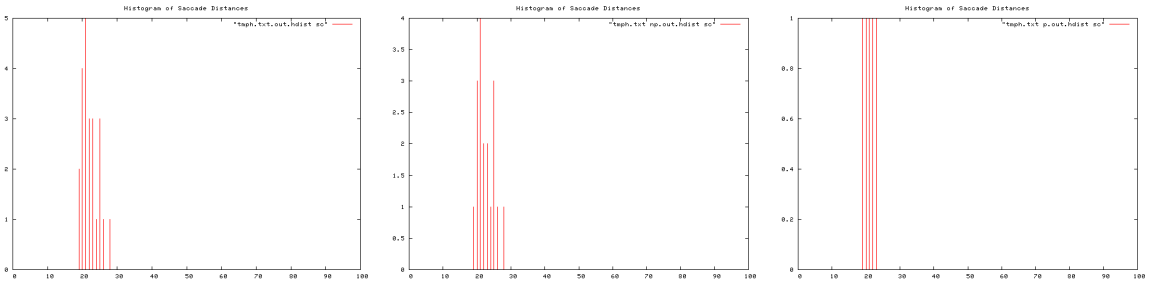



Figure B.322: Histogram of saccade distances, Trial 17. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

| Ti | To | T | I/O Event | Pear | Apple | Peach | Orange | |
|---|---|---|---|---|---|---|---|---|
| 9 | 16 | 7 | Orange In | | | | 1 | |
| 17 | 23 | 6 | Pear In | 1 | | | 0 | |
| 26 | 34 | 8 | | 2 | | | 1 | |
| 35 | 43 | 8 | Peach In | 1 | | 1 | 1 | |
| 44 | 0:55 | 11 | | 2 | | 2 | 2 | |
| 0:57 | 1:04 | 7 | | 2 | | 2 | 2 | |
| 1:05 | 1:16 | 11 | | 2 | | 3 | 2 | |
| 1:18 | 1:26 | 8 | Apple In, Peach Out | 1 | 1 | 1 | 1 | |
| 1:29 | 1:37 | 8 | Orange Out | 3 | 2 | | 4 | |
| 1:38 | 1:45 | 7 | Pear Out | 3 | 3 | | | |
| 1:47 | 1:56 | 9 | Apple Out | | 1 | | | |
| | | | TOTAL Rets | 17 | 7 | 9 | 14 | |
| | | | TOTAL T | 74 | 32 | 45 | 74 | **SD** |
| | | | **Av. Re-attention Period** | **4.4** | **4.6** | **5** | **5.3** | 0.40311289 |

Figure B.323: Re-attention period statistics, Trial 17.

### B.1.1.20 Trial 18

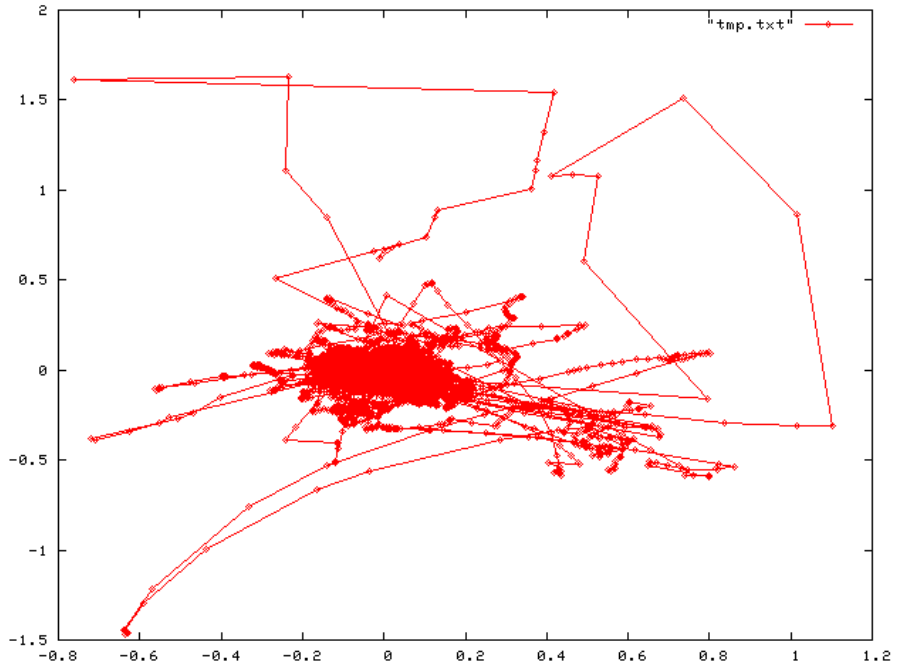



Figure B.324: Complete scan path, Trial 18.

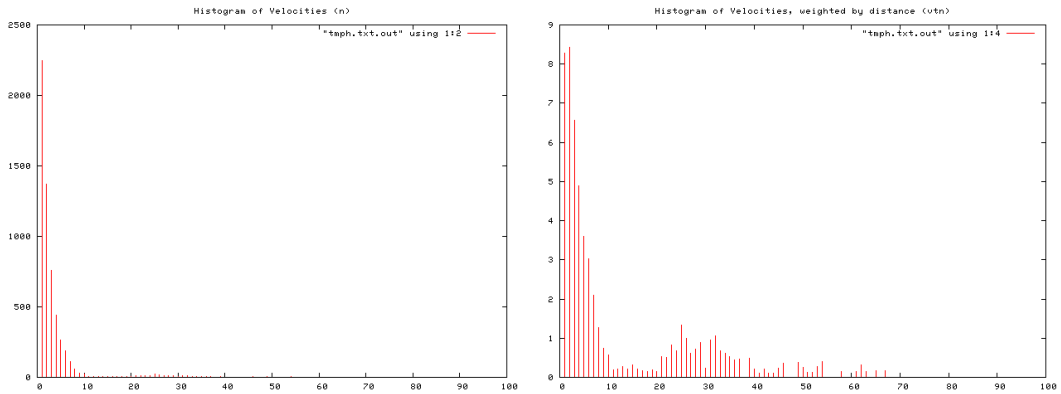



Figure B.325: Histogram of velocity magnitudes, Trial 18 (left). Histogram of distance weighted velocities, Trial 18 (right).

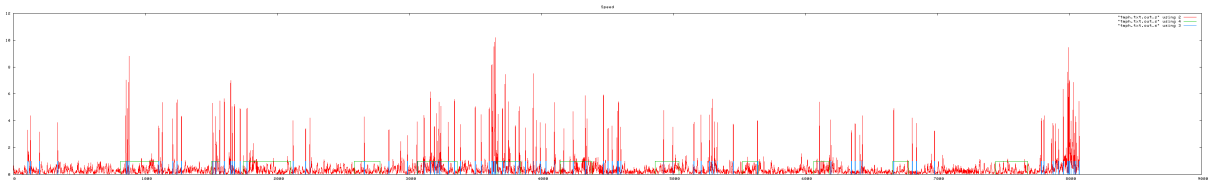


Figure B.326: Velocity profile. Velocity magnitude of each frame, Trial 18.

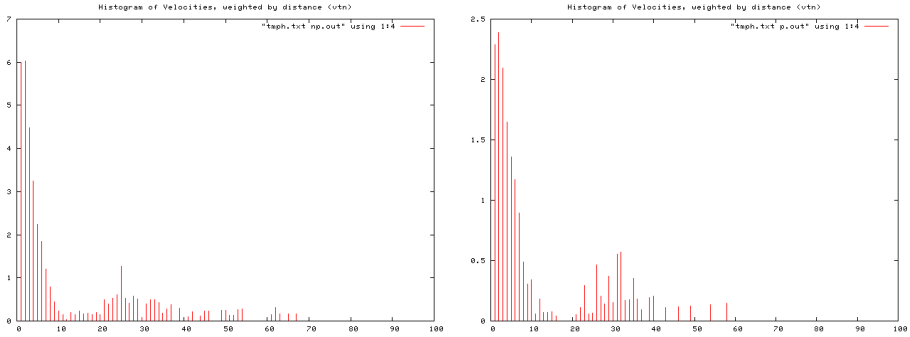



Figure B.327: Histogram of velocities during periods of no perturbation (left), and perturbation (right), Trial 18.

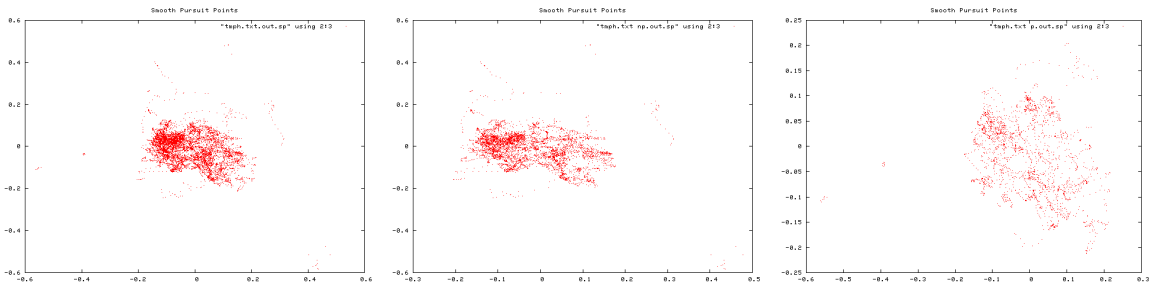



Figure B.328: Smooth pursuit gaze locations, Trial 18. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

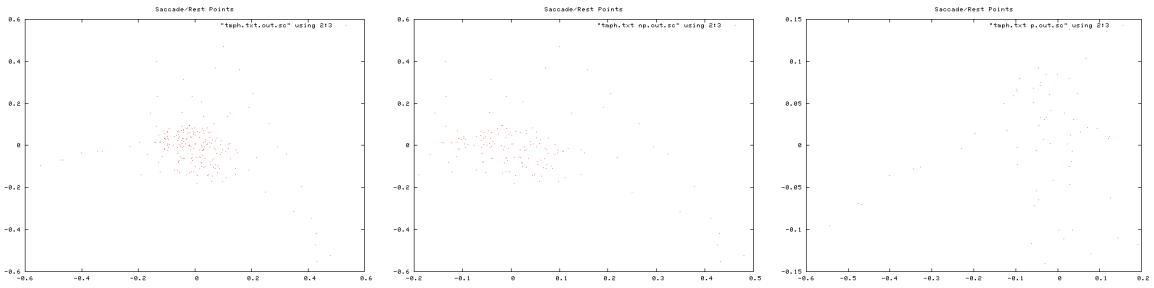



Figure B.329: Saccade gaze locations, Trial 18. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

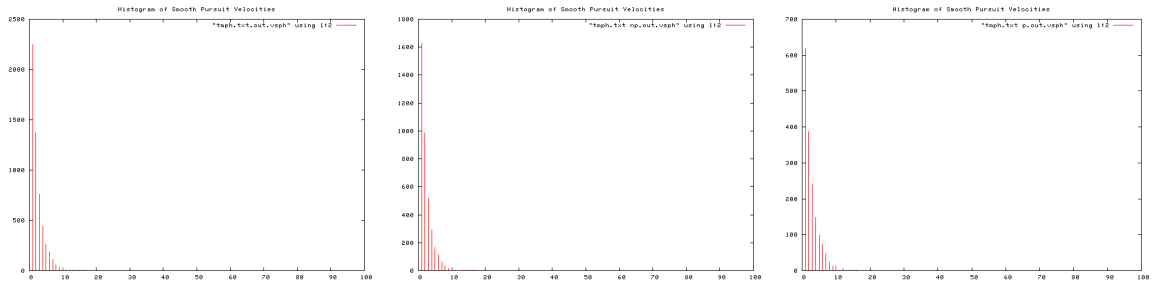



Figure B.330: Histogram of smooth pursuit velocities, Trial 18. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

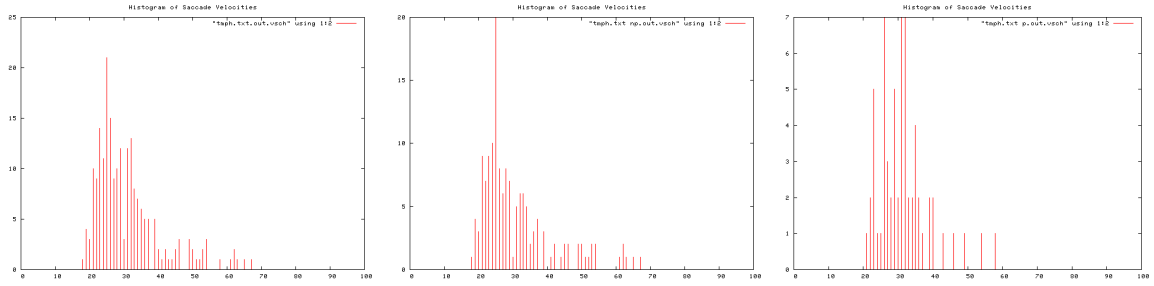



Figure B.331: Histogram of Saccade velocities, Trial 18. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

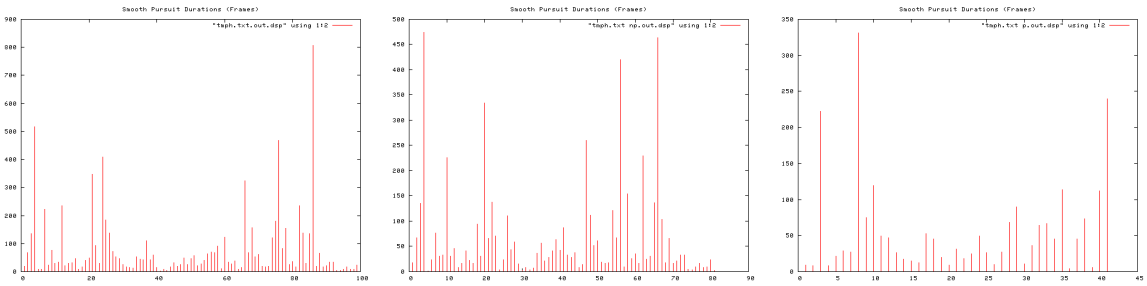



Figure B.332: Smooth pursuit durations, Trial 18. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

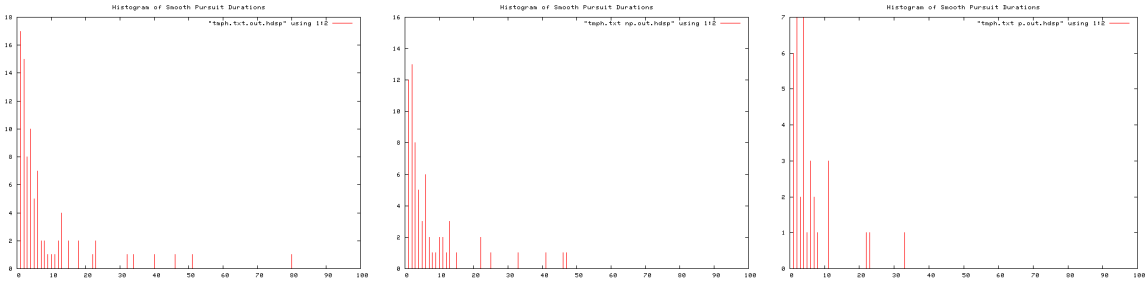



Figure B.333: Histogram of Smooth pursuit durations, Trial 18. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

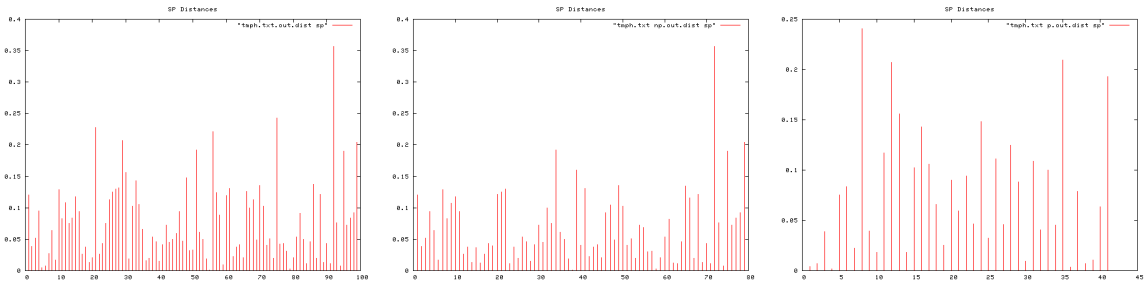



Figure B.334: Smooth pursuit distances, Trial 18. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

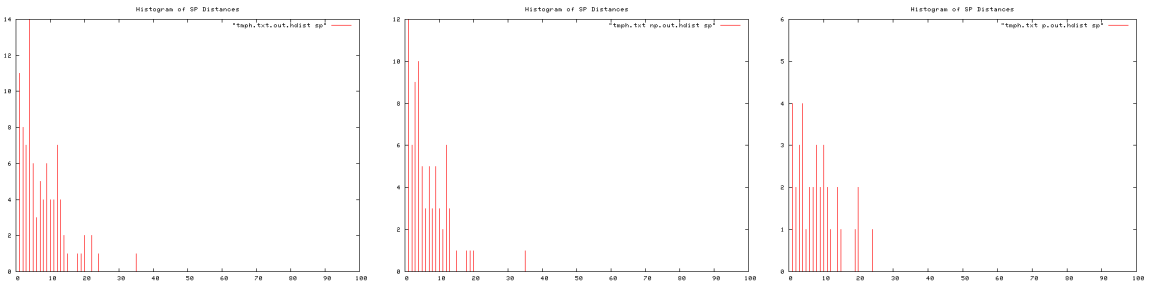



Figure B.335: Histogram of smooth pursuit distances, Trial 18. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

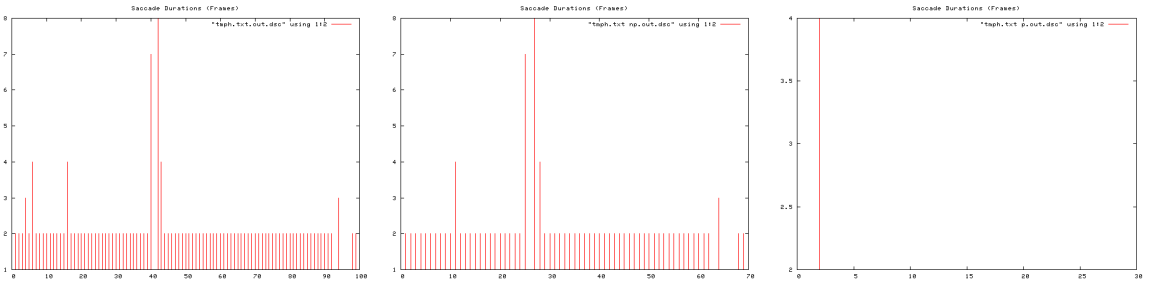



Figure B.336: Saccade durations, Trial 18. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

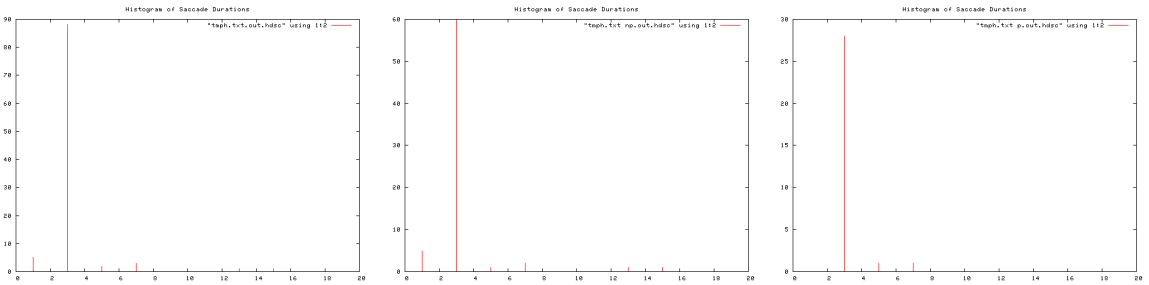



Figure B.337: Histogram of saccade durations, Trial 18. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

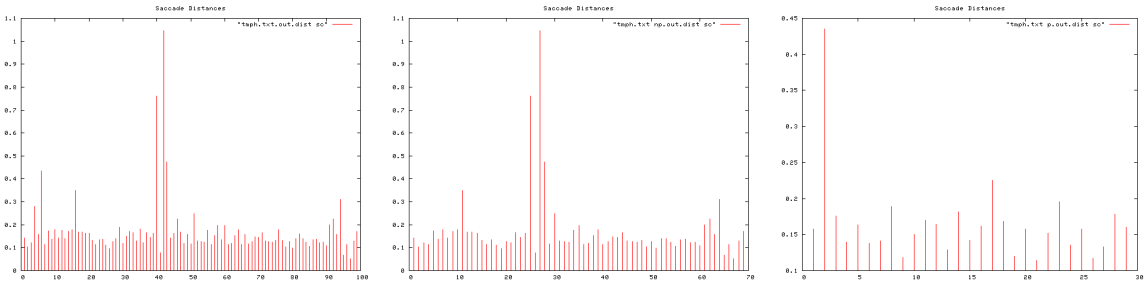



Figure B.338: Saccade distances, Trial 18. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

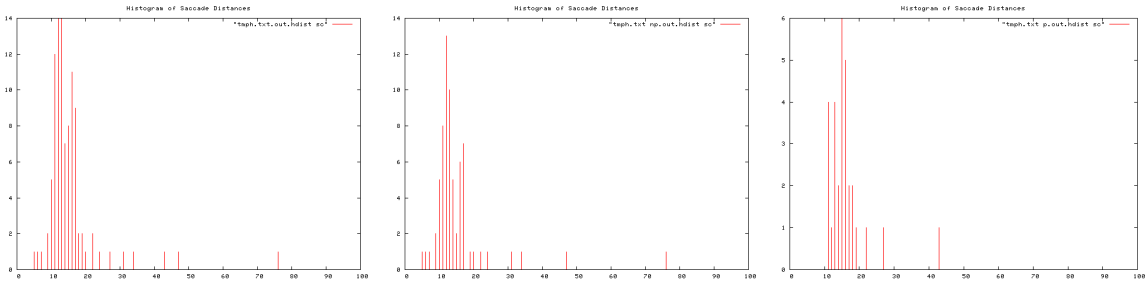



Figure B.339: Histogram of saccade distances, Trial 18. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

| Ti | To | T | I/O Event | Pear | Apple | Peach | Orange | |
|---|---|---|---|---|---|---|---|---|
| 58 | 1:03 | 5 | Orange In | | | | 2 | |
| 1:04 | 1:09 | 5 | Pear In | 2 | | | 2 | |
| 1:11 | 1:23 | 12 | | 3 | | | 3 | |
| 1:24 | 1:30 | 6 | Peach In | 2 | | 2 | 1 | |
| 1:31 | 1:42 | 11 | | 2 | | 2 | 1 | |
| 1:43 | 1:50 | 7 | | 1 | | 2 | 2 | |
| 1:51 | 1:59 | 8 | | 2 | | 2 | 2 | |
| 2:00 | 2:10 | 10 | Apple In, Peach Out | 1 | 2 | 2 | 2 | |
| 2:12 | 2:19 | 7 | Orange Out | 1 | 2 | | 2 | |
| 2:20 | 2:29 | 9 | Pear Out | 2 | 3 | | | |
| 2:31 | 2:43 | 12 | Apple Out | | 2 | | | |
| | | | TOTAL Rets | 16 | 9 | 10 | 17 | |
| | | | TOTAL T | 75 | 48 | 42 | 71 | **SD** |
| | | | **Av. Re-attention Period** | **4.7** | **5.3** | **4.2** | **4.2** | 0.5228129 |

Figure B.340: Re-attention period statistics, Trial 18.

### B.1.1.21 Trial 19

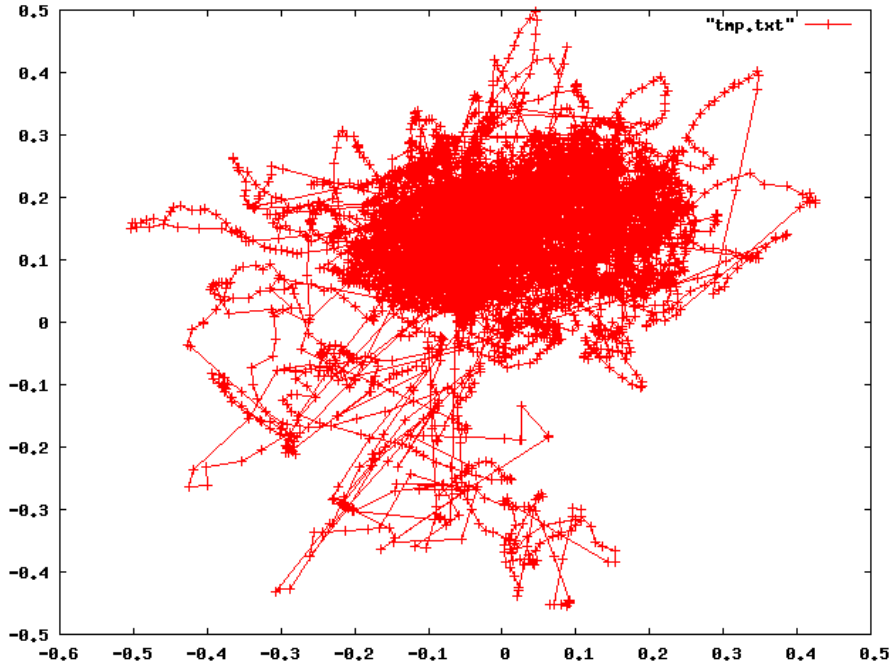



Figure B.341: Complete scan path, Trial 19.

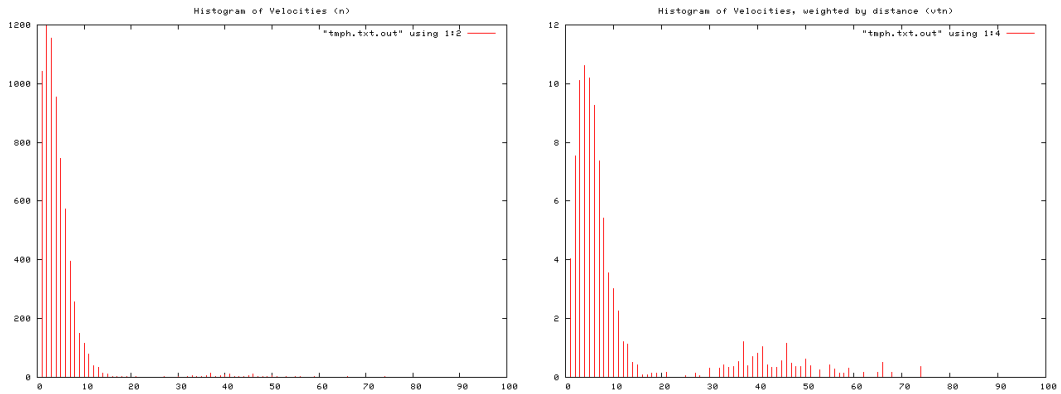



Figure B.342: Histogram of velocity magnitudes, Trial 19 (left). Histogram of distance weighted velocities, Trial 19 (right).

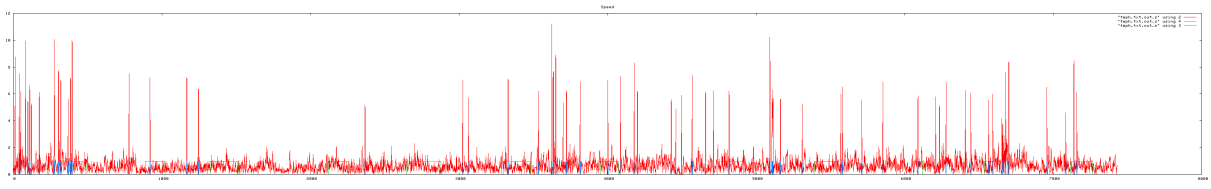


Figure B.343: Velocity profile. Velocity magnitude of each frame, Trial 19.

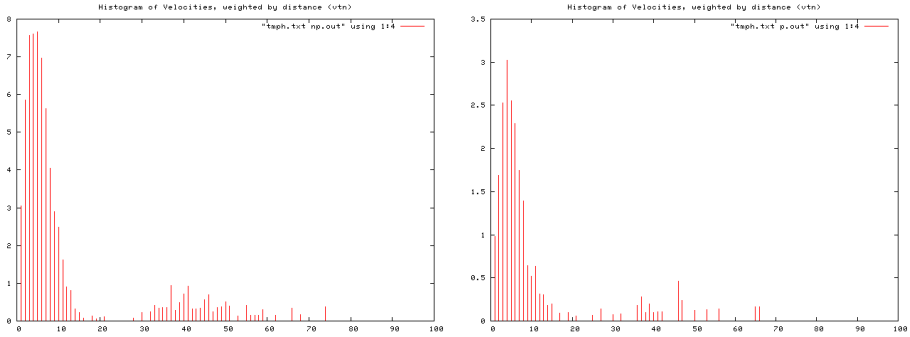



Figure B.344: Histogram of velocities during periods of no perturbation (left), and perturbation (right), Trial 19.


Figure B.345: Smooth pursuit gaze locations, Trial 19. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

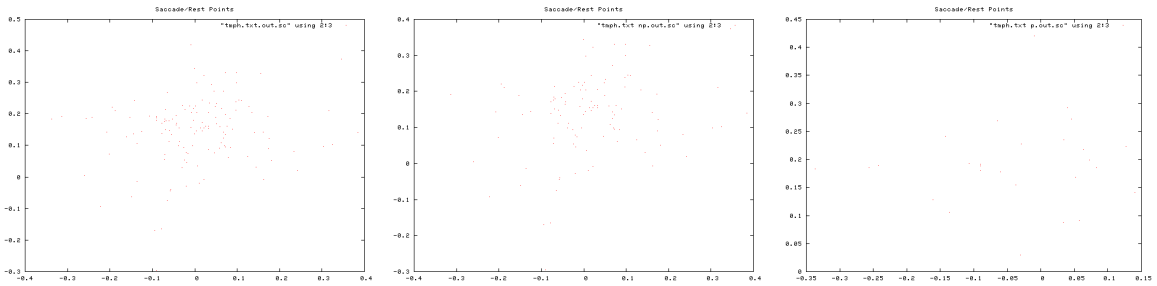



Figure B.346: Saccade gaze locations, Trial 19. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

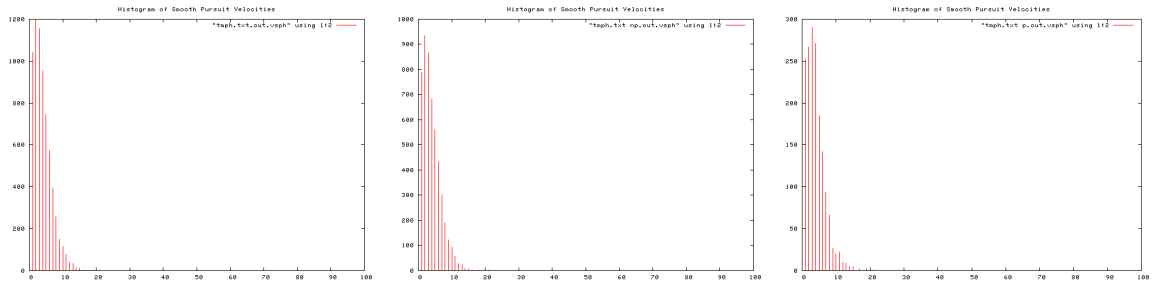



Figure B.347: Histogram of smooth pursuit velocities, Trial 19. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

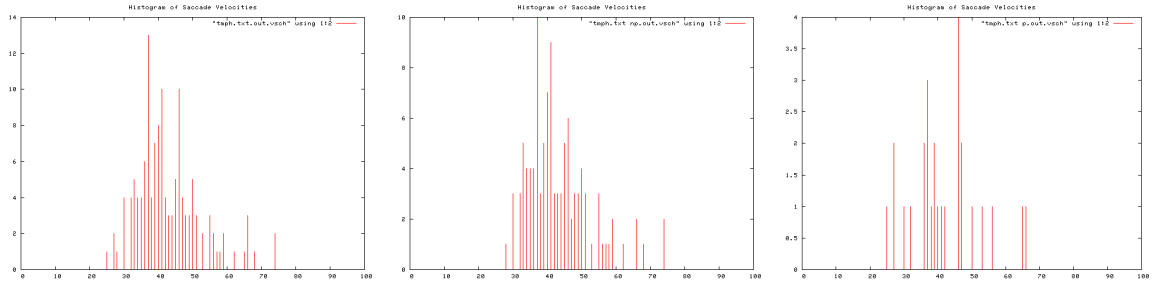



Figure B.348: Histogram of Saccade velocities, Trial 19. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

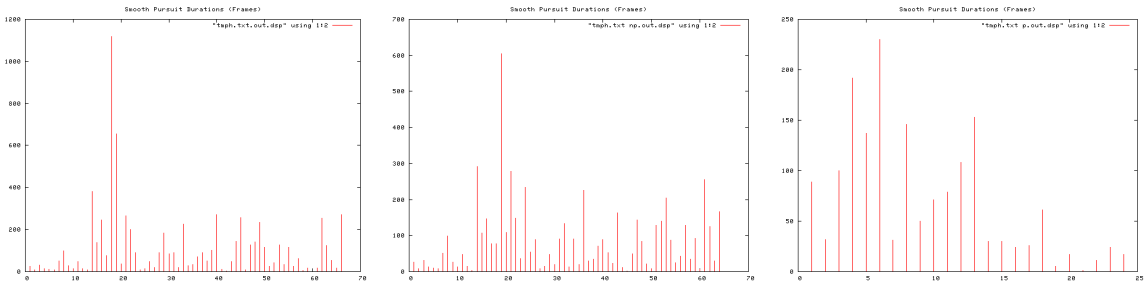



Figure B.349: Smooth pursuit durations, Trial 19. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

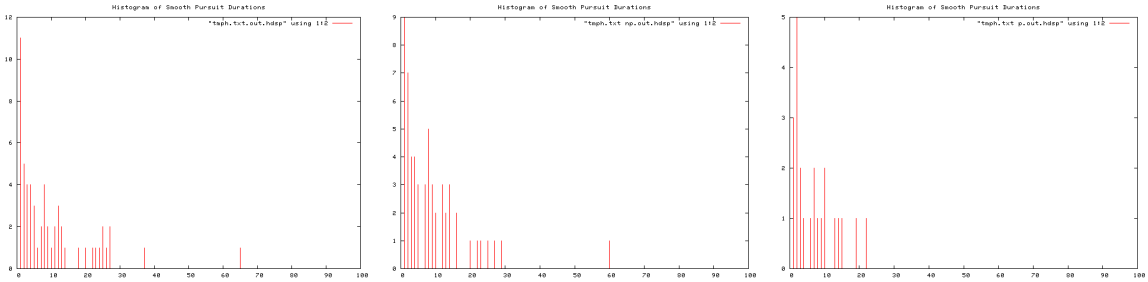



Figure B.350: Histogram of Smooth pursuit durations, Trial 19. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

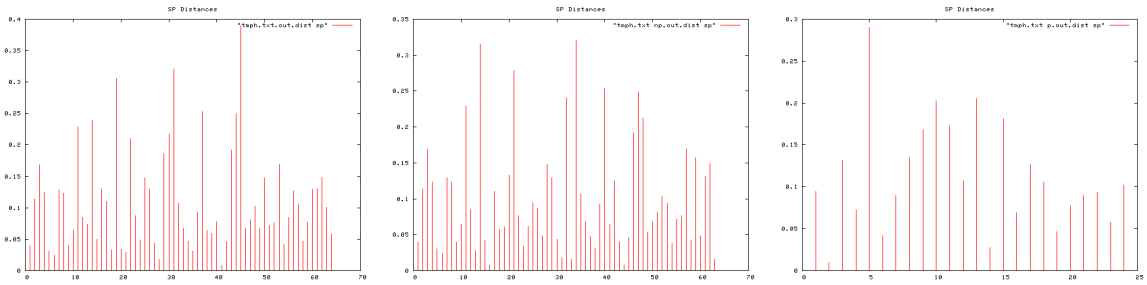



Figure B.351: Smooth pursuit distances, Trial 19. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

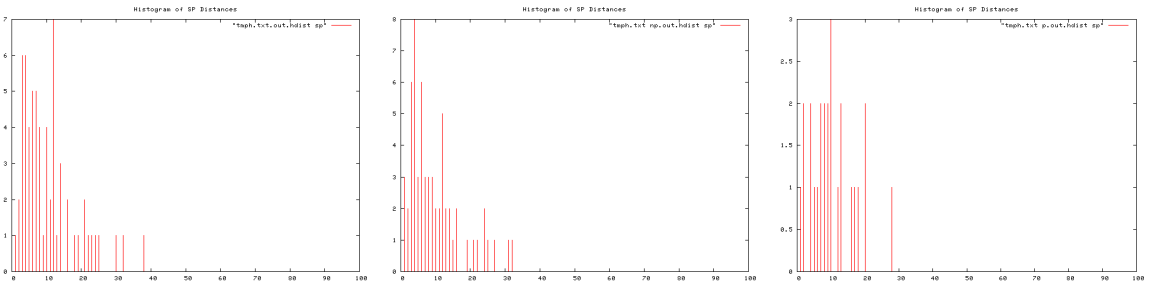



Figure B.352: Histogram of smooth pursuit distances, Trial 19. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

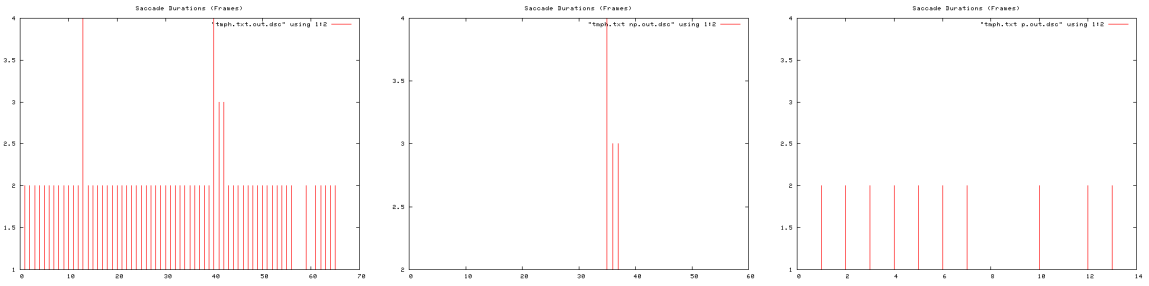



Figure B.353: Saccade durations, Trial 19. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

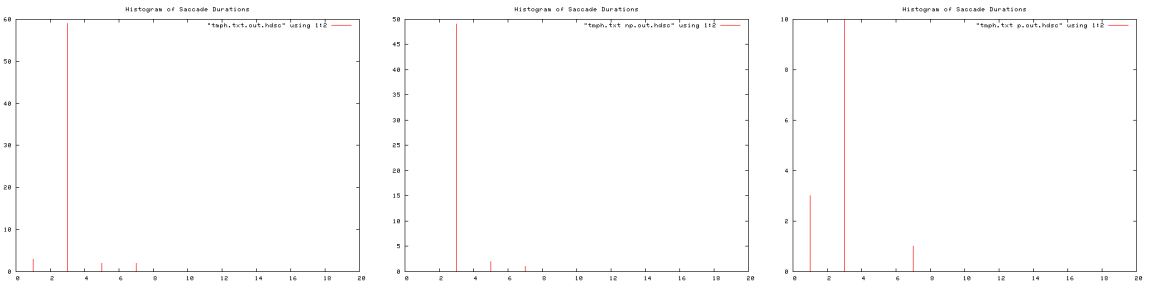



Figure B.354: Histogram of saccade durations, Trial 19. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

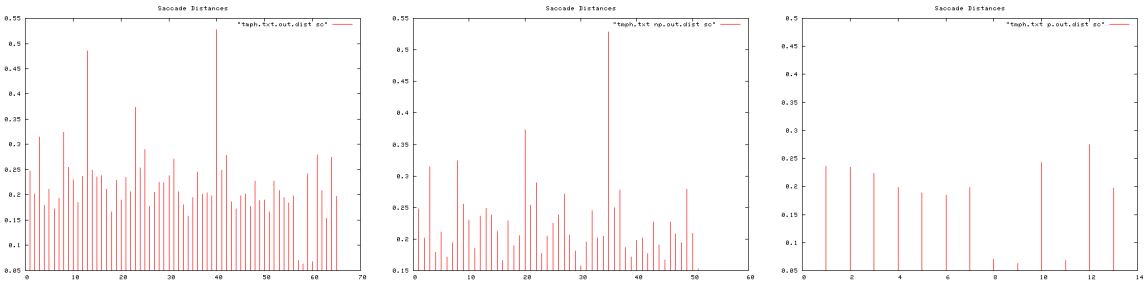



Figure B.355: Saccade distances, Trial 19. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

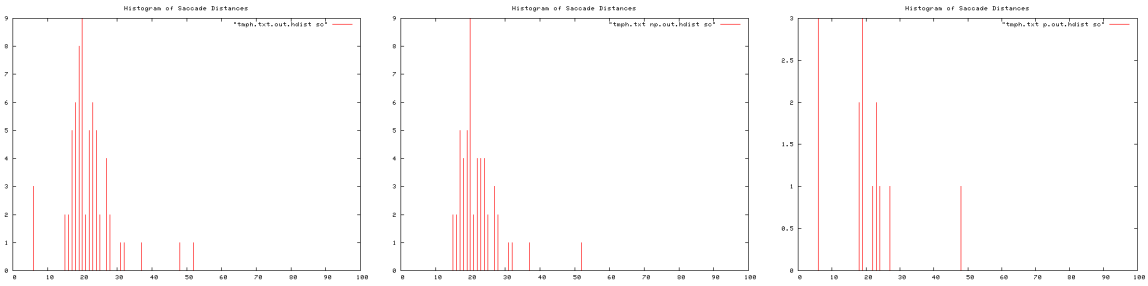



Figure B.356: Histogram of saccade distances, Trial 19. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

| Ti | To | T | I/O Event | Pear | Apple | Peach | Orange | |
|---|---|---|---|---|---|---|---|---|
| 7 | 14 | 7 | Orange In | | | | 2 | |
| 16 | 23 | 6 | Pear In | 2 | | | 1 | |
| 25 | 37 | 12 | | 3 | | | 2 | |
| 37 | 46 | 9 | Peach In | 1 | | 2 | 2 | |
| 47 | 0:56 | 9 | | 2 | | 2 | 2 | |
| 0:58 | 1:06 | 8 | | 3 | | 2 | 2 | |
| 1:08 | 1:19 | 11 | | 3 | | 3 | 2 | |
| 1:22 | 1:30 | 8 | Apple In, Peach Out | 1 | 3 | 2 | 2 | |
| 1:32 | 1:41 | 9 | Orange Out | 1 | 2 | | 2 | |
| 1:42 | 1:50 | 8 | Pear Out | 2 | 2 | | | |
| 1:52 | 1:59 | 7 | Apple Out | | 1 | | | |
| | | | TOTAL Rets | 17 | 8 | 11 | 17 | |
| | | | TOTAL T | 80 | 32 | 45 | 79 | **SD** |
| | | | **Av. Re-attention Period** | **4.7** | **4** | **4.1** | **4.6** | 0.35118846 |

Figure B.357: Re-attention period statistics, Trial 19.

### B.1.1.22 Trial 20

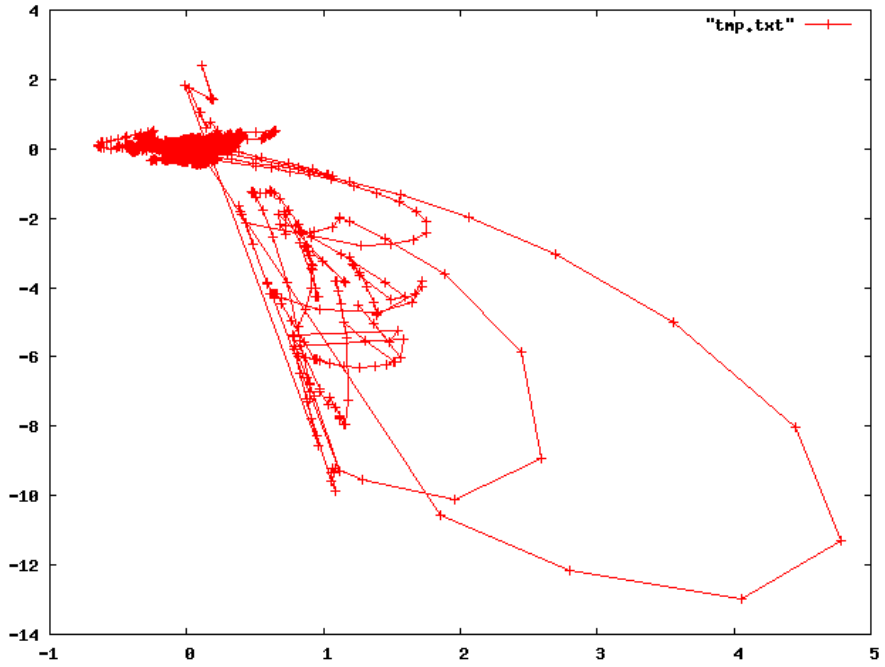



Figure B.358: Complete scan path, Trial 20.

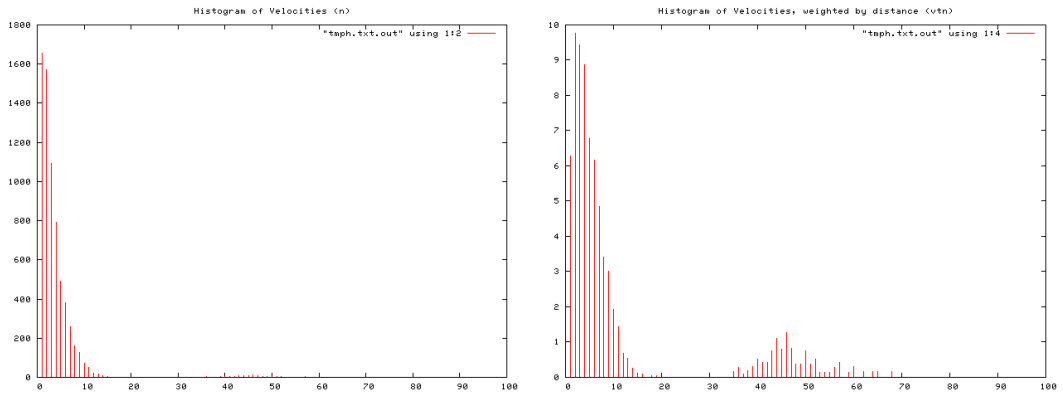



Figure B.359: Histogram of velocity magnitudes, Trial 20 (left). Histogram of distance weighted velocities, Trial 20 (right).

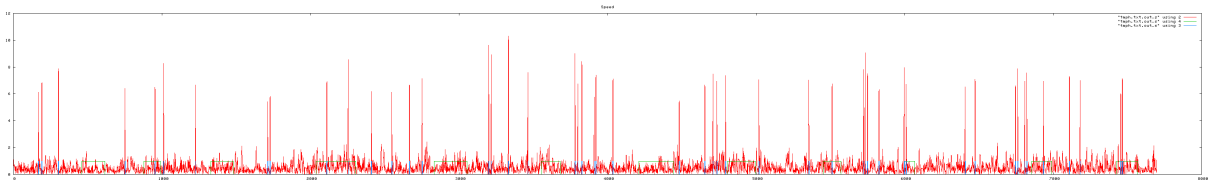


Figure B.360: Velocity profile. Velocity magnitude of each frame, Trial 20.

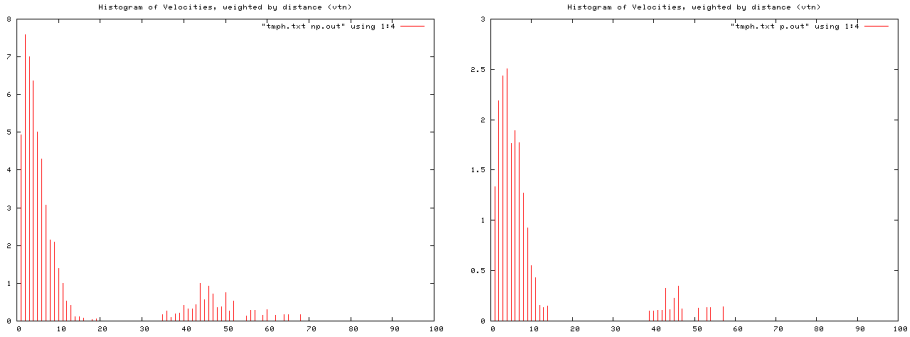



Figure B.361: Histogram of velocities during periods of no perturbation (left), and perturbation (right), Trial 20.

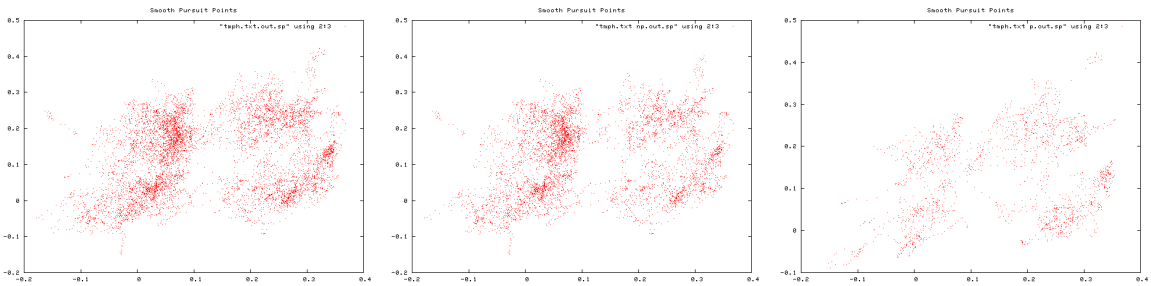



Figure B.362: Smooth pursuit gaze locations, Trial 20. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

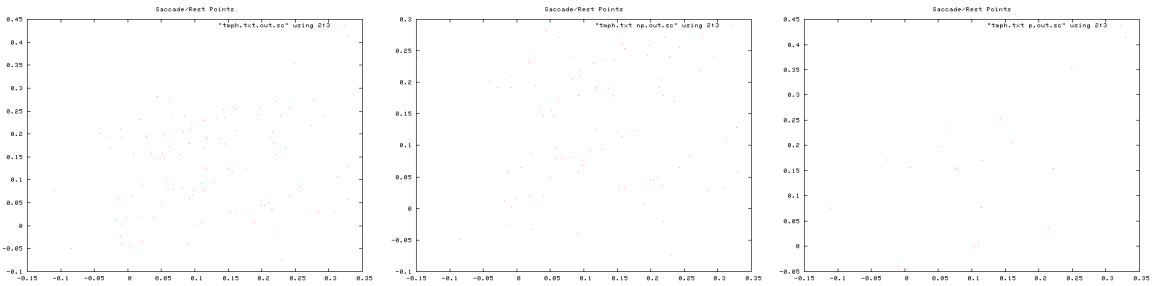



Figure B.363: Saccade gaze locations, Trial 20. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

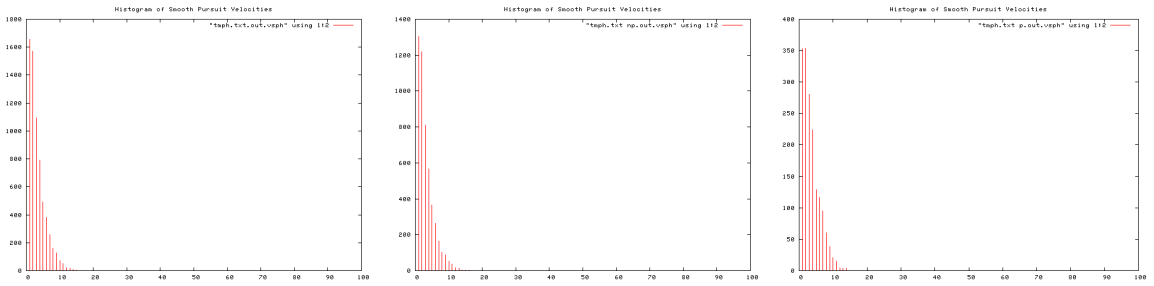



Figure B.364: Histogram of smooth pursuit velocities, Trial 20. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

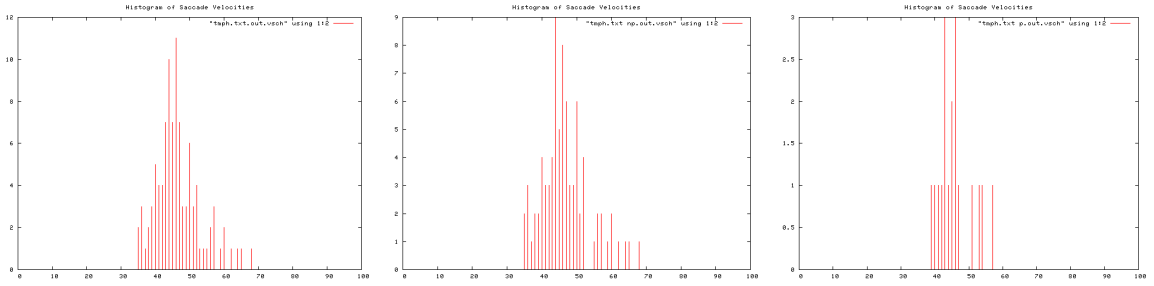



Figure B.365: Histogram of Saccade velocities, Trial 20. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

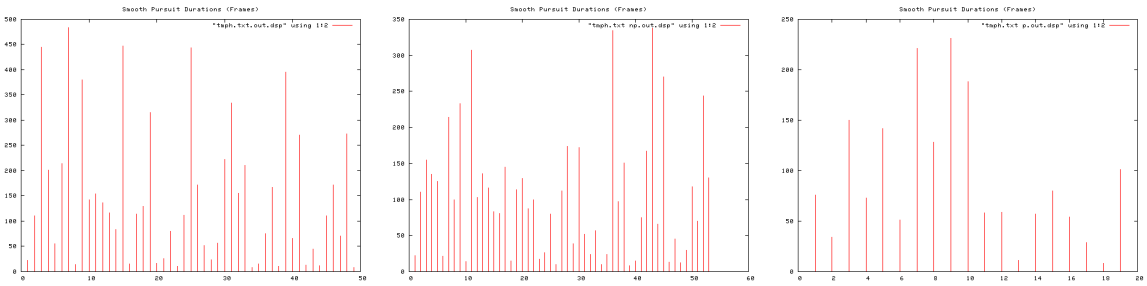



Figure B.366: Smooth pursuit durations, Trial 20. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

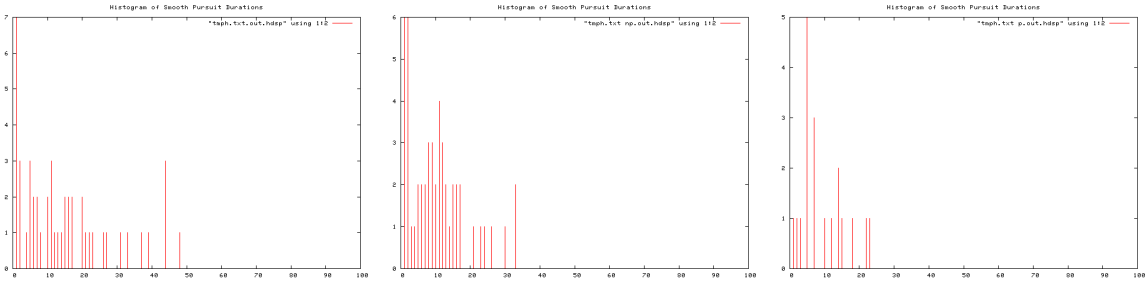



Figure B.367: Histogram of Smooth pursuit durations, Trial 20. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).


Figure B.368: Smooth pursuit distances, Trial 20. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

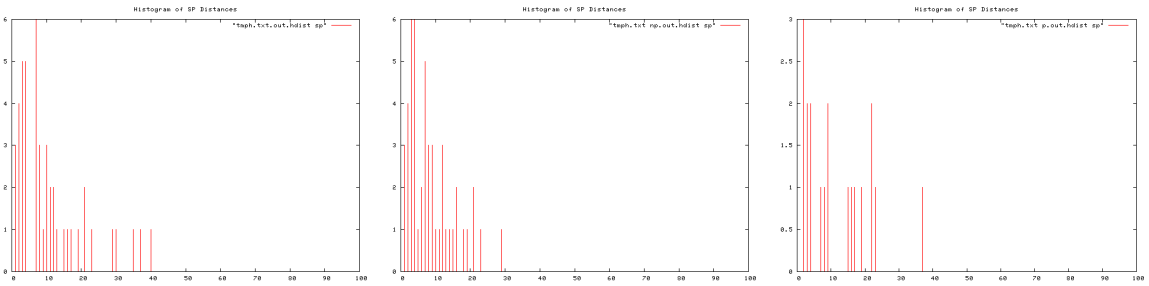



Figure B.369: Histogram of smooth pursuit distances, Trial 20. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

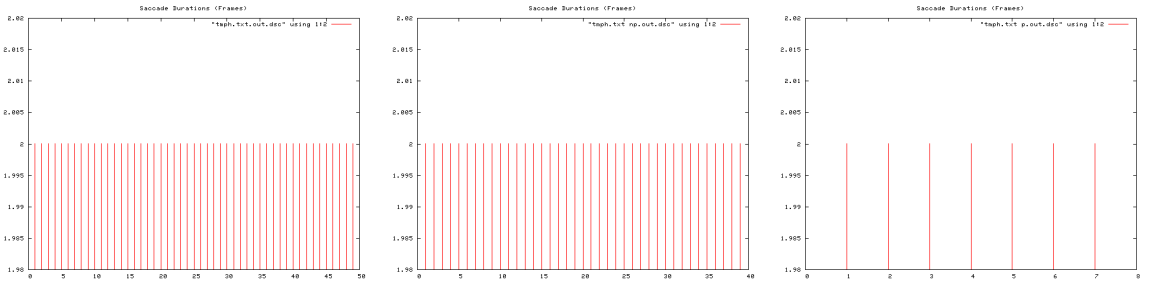



Figure B.370: Saccade durations, Trial 20. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

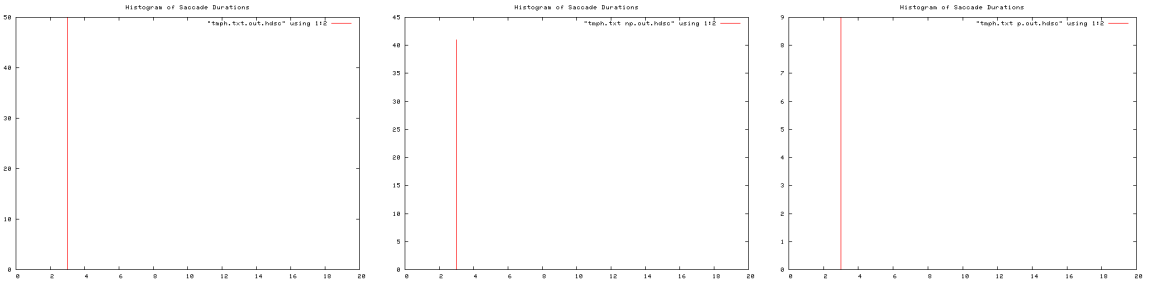



Figure B.371: Histogram of saccade durations, Trial 20. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

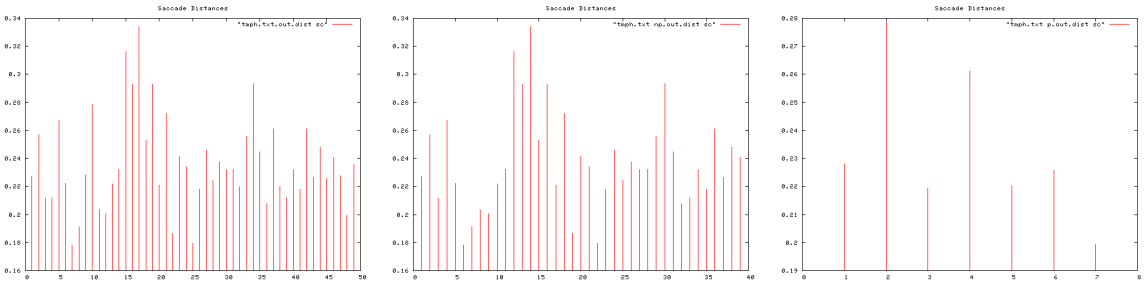



Figure B.372: Saccade distances, Trial 20. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

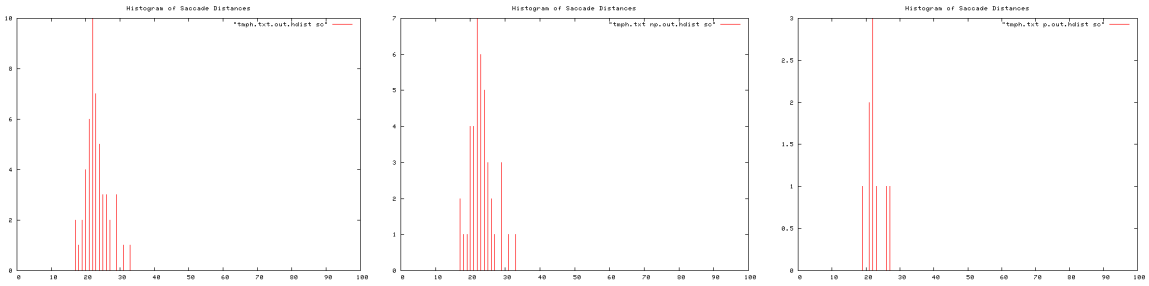



Figure B.373: Histogram of saccade distances, Trial 20. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

| Ti | To | T | I/O Event | Pear | Apple | Peach | Orange | |
|---|---|---|---|---|---|---|---|---|
| 6 | 13 | 7 | Orange In | | | | 1 | |
| 14 | 21 | 7 | Pear In | 2 | | | 1 | |
| 23 | 35 | 12 | | 3 | | | 3 | |
| 36 | 48 | 12 | Peach In | 2 | | 2 | 3 | |
| 49 | 0:55 | 6 | | 1 | | 2 | 1 | |
| 0:58 | 1:08 | 10 | | 2 | | 3 | 2 | |
| 1:10 | 1:23 | 13 | | 2 | | 3 | 3 | |
| 1:26 | 1:36 | 10 | Apple In, Peach Out | 1 | 3 | 1 | 1 | |
| 1:39 | 1:50 | 11 | Orange Out | 2 | 3 | | 2 | |
| 1:52 | 2:01 | 9 | Pear Out | 4 | 2 | | | |
| 2:04 | 4:15 | 10 | Apple Out | | 1 | | | |
| | | | TOTAL Rets | 19 | 9 | 11 | 17 | |
| | | | TOTAL T | 90 | 40 | 51 | 88 | **SD** |
| | | | **Av. Re-attention Period** | **4.7** | **4.4** | **4.6** | **5.2** | 0.34034296 |

Figure B.374: Re-attention period statistics, Trial 20.

## B.1.2 Group Statistics

### B.1.2.1 Processing Script Output

Table B.1: Extracted parameters, human trials (1 of 3).

| Param | P1 | P2 | T1 | T2 | T3 | T4 | T5 |
|---|---|---|---|---|---|---|---|
| $Spv_p$ | 0.476394 | 0.537571 | 0.483943 | 0.512705 | 0.635670 | 0.448225 | 0.476246 |
| $Spv_np$ | 0.484677 | 0.543956 | 0.470061 | 0.475582 | 0.606783 | 0.434325 | 0.490644 |
| $Scv_p$ | 6.797215 | 7.028750 | 5.380098 | 5.936850 | 6.614626 | 5.086491 | 6.778406 |
| $Scv_np$ | 7.168081 | 7.343113 | 6.212796 | 6.365770 | 6.740767 | 5.015597 | 7.360287 |
| $Scl$ | 0.265968 | 0.263172 | 0.196104 | 0.170443 | 0.228847 | 0.142702 | 0.278075 |
| $Spl$ | 0.102800 | 0.130395 | 0.054773 | 0.067970 | 0.106074 | 0.073668 | 0.090951 |
| $Scf$ | 289 | 182 | 282 | 358 | 194 | 109 | 163 |
| $Spf$ | 8068 | 9166 | 7411 | 7421 | 6461 | 7552 | 6216 |
| $Sc\%$ | 3.582053 | 1.985599 | 3.805155 | 4.824148 | 3.002631 | 1.443326 | 2.622265 |
| $Spt_p$ | 86.038462 | 88.407407 | 45.702703 | 42.742857 | 48.541667 | 64.318182 | 57.958333 |
| $Sct_p$ | 2.187500 | 2.000000 | 1.640000 | 1.833333 | 1.857143 | 1.636364 | 2.166667 |
| $Scl_p$ | 0.247815 | 0.234292 | 0.147056 | 0.181404 | 0.204738 | 0.138722 | 0.244776 |
| $Spl_p$ | 0.115915 | 0.161802 | 0.077886 | 0.112763 | 0.102770 | 0.072543 | 0.099557 |
| $Scf_p$ | 35 | 32 | 41 | 44 | 26 | 18 | 26 |
| $Spf_p$ | 2237 | 2387 | 1691 | 1496 | 1165 | 1415 | 1391 |
| $Sc_p\%$ | 1.564595 | 1.340595 | 2.424601 | 2.941176 | 2.231760 | 1.272085 | 1.869159 |
| $Spt_np$ | 46.269841 | 83.679012 | 43.325758 | 28.209524 | 56.935484 | 94.400000 | 67.000000 |
| $Sct_np$ | 2.228070 | 2.173913 | 1.991736 | 1.593909 | 2.024096 | 1.716981 | 2.322034 |
| $Scl_np$ | 0.266183 | 0.266055 | 0.206237 | 0.169108 | 0.227399 | 0.143528 | 0.284847 |
| $Spl_np$ | 0.092618 | 0.118117 | 0.051856 | 0.056486 | 0.098408 | 0.058626 | 0.074130 |
| $Scf_np$ | 254 | 150 | 241 | 314 | 168 | 91 | 137 |
| $Spf_np$ | 5831 | 6779 | 5720 | 5925 | 5296 | 6137 | 4825 |
| $Sc_np\%$ | 4.356028 | 2.212716 | 4.213287 | 5.299578 | 3.172205 | 1.482809 | 2.839378 |
| $Scp_r$ | 2.784124 | 1.650548 | 1.737724 | 1.801857 | 1.421392 | 1.165653 | 1.519067 |
| $Scl_r$ | 1.074120 | 1.135571 | 1.402441 | 0.932217 | 1.110682 | 1.034642 | 1.163707 |
| $Spl_r$ | 0.799019 | 0.730010 | 0.665801 | 0.500928 | 0.957561 | 0.808156 | 0.744593 |
| $Sct_r$ | 1.018546 | 1.086957 | 1.214473 | 0.869405 | 1.089898 | 1.049266 | 1.071708 |
| $Spt_r$ | 0.537781 | 0.946516 | 0.947991 | 0.659982 | 1.172920 | 1.467703 | 1.156003 |
| $Scv_r$ | 1.054561 | 1.044725 | 1.154774 | 1.072247 | 1.019070 | 0.986062 | 1.085843 |
| $Spv_r$ | 1.017385 | 1.011877 | 0.971315 | 0.927594 | 0.954557 | 0.968989 | 1.030232 |
| $Pr_{av}$ | 4.725 | 4.35 | 4.6 | 4.825 | 3.775 | 6.0 | 5.85 |
| $Pr_{sd}$ | 0.34 | 0.35 | 0.52 | 0.40 | 0.48 | 0.18 | 0.38 |
| $T$ | 132 | 119 | 118 | 122 | 103 | 112 | 96 |
| $Age$ | 28 | 29 | 32 | 26 | 25 | 26 | 30 |

Table B.2: Extracted parameters, human trials (2 of 3).

| Param | T6 | T7 | T8 | T9 | T10 | T11 | T12 |
|---|---|---|---|---|---|---|---|
| $Spv_p$ | 0.295954 | 0.377642 | 0.513933 | 0.714394 | 0.626854 | 0.351033 | 0.343403 |
| $Spv_np$ | 0.308856 | 0.360486 | 0.532035 | 0.654225 | 0.633180 | 0.339045 | 0.357335 |
| $Scv_p$ | 6.282551 | 6.171099 | 6.108873 | 7.661773 | 8.826044 | 13.398437 | 6.283213 |
| $Scv_np$ | 6.570229 | 9.148609 | 6.752817 | 10.165976 | 15.732483 | 32.566405 | 7.224471 |
| $Scl$ | 0.221303 | 0.323260 | 0.230835 | 0.313068 | 0.580876 | 1.681185 | 0.282256 |
| $Spl$ | 0.060289 | 0.085546 | 0.103825 | 0.058297 | 0.121427 | 0.129064 | 0.092076 |
| $Scf$ | 126 | 136 | 203 | 288 | 281 | 151 | 112 |
| $Spf$ | 7079 | 8006 | 6565 | 8786 | 7384 | 7371 | 7053 |
| $Sc\%$ | 1.779912 | 1.698726 | 3.092155 | 3.277942 | 3.805525 | 2.048569 | 1.587977 |
| $Spt_p$ | 61.150000 | 59.850000 | 52.906977 | 41.593750 | 57.902439 | 72.809524 | 90.684211 |
| $Sct_p$ | 2.000000 | 1.750000 | 2.064516 | 1.450000 | 2.931034 | 2.500000 | 2.000000 |
| $Scl_p$ | 0.209418 | 0.179990 | 0.210198 | 0.185160 | 0.431157 | 0.558268 | 0.209440 |
| $Spl_p$ | 0.072526 | 0.085483 | 0.098626 | 0.083659 | 0.116613 | 0.093650 | 0.086653 |
| $Scf_p$ | 16 | 14 | 64 | 29 | 85 | 25 | 14 |
| $Spf_p$ | 1223 | 1197 | 2275 | 1331 | 2374 | 1529 | 1723 |
| $Sc_p\%$ | 1.308258 | 1.169591 | 2.813187 | 2.178813 | 3.580455 | 1.635056 | 0.812536 |
| $Spt_np$ | 87.388060 | 101.611940 | 54.987179 | 51.406897 | 54.445652 | 126.978261 | 100.547170 |
| $Sct_np$ | 2.037037 | 2.259259 | 2.138462 | 1.962121 | 2.419753 | 3.600000 | 2.450000 |
| $Scl_np$ | 0.223063 | 0.344485 | 0.240677 | 0.332448 | 0.634479 | 1.953984 | 0.294999 |
| $Spl_np$ | 0.052493 | 0.093731 | 0.092262 | 0.057455 | 0.112225 | 0.105982 | 0.081818 |
| $Scf_np$ | 110 | 122 | 139 | 259 | 196 | 126 | 98 |
| $Spf_np$ | 5856 | 6809 | 4290 | 7455 | 5010 | 5842 | 5330 |
| $Sc_np\%$ | 1.878415 | 1.791746 | 3.240093 | 3.474178 | 3.912176 | 2.156796 | 1.838649 |
| $Scp_r$ | 1.435814 | 1.531943 | 1.151752 | 1.594528 | 1.092648 | 1.319096 | 2.262852 |
| $Scl_r$ | 1.065156 | 1.913906 | 1.145004 | 1.795468 | 1.471571 | 3.500082 | 1.408511 |
| $Spl_r$ | 0.723783 | 1.096498 | 0.935471 | 0.686773 | 0.962371 | 1.131682 | 0.944198 |
| $Sct_r$ | 1.018519 | 1.291005 | 1.035817 | 1.353187 | 0.825563 | 1.440000 | 1.225000 |
| $Spt_r$ | 1.429077 | 1.697777 | 1.039318 | 1.235928 | 0.940300 | 1.743979 | 1.108762 |
| $Scv_r$ | 1.045790 | 1.482493 | 1.105411 | 1.326844 | 1.782507 | 2.430612 | 1.149805 |
| $Spv_r$ | 1.043595 | 0.954570 | 1.035222 | 0.915776 | 1.010091 | 0.965848 | 1.040573 |
| $Pr_{av}$ | 5.875 | 5.925 | 9.775 | 3.85 | 10.375 | 4.1 | 4.125 |
| $Pr_{sd}$ | 0.26 | 0.56 | 1.13 | 0.17 | 0.76 | 0.39 | 0.17 |
| $T$ | 112 | 118 | 100 | 143 | 106 | 110 | 110 |
| $Age$ | 53 | 31 | 26 | 47 | 26 | 33 | 31 |

Table B.3: Extracted parameters, human trials (3 of 3).

| Param | T13 | T14 | T15 | T16 | T17 | T18 | T19 | T20 |
|---|---|---|---|---|---|---|---|---|
| $Spv_p$ | 0.369853 | 0.340640 | 0.956269 | 0.571606 | 0.291247 | 0.368432 | 0.656070 | 0.560402 |
| $Spv_np$ | 0.397492 | 0.316563 | 0.923015 | 0.653905 | 0.293808 | 0.327663 | 0.636751 | 0.497662 |
| $Scv_p$ | 6.401014 | 4.797172 | 8.963861 | 5.256359 | 6.477056 | 4.810778 | 6.360418 | 6.935337 |
| $Scv_np$ | 7.000278 | 6.789544 | 8.722185 | 6.382642 | 6.921927 | 4.716970 | 6.557405 | 7.095826 |
| $Scl$ | 0.237515 | 0.239663 | 0.299018 | 0.144104 | 0.227507 | 0.169224 | 0.222205 | 0.235565 |
| $Spl$ | 0.095461 | 0.073853 | 0.152817 | 0.051045 | 0.108080 | 0.076366 | 0.110198 | 0.107614 |
| $Scf$ | 207 | 74 | 437 | 625 | 46 | 214 | 135 | 100 |
| $Spf$ | 7285 | 7181 | 7467 | 6608 | 7203 | 7862 | 7298 | 7600 |
| $Sc\%$ | 2.841455 | 1.030497 | 5.852417 | 9.458232 | 0.638623 | 2.721954 | 1.849822 | 1.315789 |
| $Spt_p$ | 49.677419 | 109.222222 | 31.724638 | 16.623529 | 133.882353 | 55.833333 | 68.153846 | 95.300000 |
| $Sct_p$ | 2.105263 | 2.000000 | 2.017544 | 1.266667 | 2.000000 | 2.100000 | 1.928571 | 2.000000 |
| $Scl_p$ | 0.224597 | 0.159906 | 0.301416 | 0.110968 | 0.215902 | 0.168377 | 0.204442 | 0.231178 |
| $Spl_p$ | 0.076836 | 0.092113 | 0.168399 | 0.051147 | 0.109561 | 0.076056 | 0.108809 | 0.116596 |
| $Scf_p$ | 40 | 14 | 115 | 95 | 10 | 63 | 27 | 18 |
| $Spf_p$ | 1540 | 1966 | 2189 | 1413 | 2276 | 2345 | 1772 | 1906 |
| $Sc_p\%$ | 2.597403 | 0.712106 | 5.253540 | 6.723284 | 0.439367 | 2.686567 | 1.523702 | 0.944386 |
| $Spt_np$ | 61.106383 | 137.210526 | 31.041176 | 13.455959 | 164.200000 | 67.268293 | 85.000000 | 105.425926 |
| $Sct_np$ | 2.061728 | 2.307692 | 2.050955 | 1.417112 | 2.000000 | 2.157143 | 2.076923 | 2.000000 |
| $Scl_np$ | 0.240545 | 0.261136 | 0.298147 | 0.150749 | 0.230731 | 0.169586 | 0.226987 | 0.236528 |
| $Spl_np$ | 0.084115 | 0.053749 | 0.140224 | 0.050604 | 0.087533 | 0.067844 | 0.100205 | 0.083747 |
| $Scf_np$ | 167 | 60 | 322 | 530 | 36 | 151 | 108 | 82 |
| $Spf_np$ | 5745 | 5215 | 5278 | 5195 | 4927 | 5517 | 5526 | 5694 |
| $Sc_np\%$ | 2.906876 | 1.150527 | 6.100796 | 10.202117 | 0.730668 | 2.736995 | 1.954397 | 1.440112 |
| $Scp_r$ | 1.119147 | 1.615669 | 1.161273 | 1.517431 | 1.663000 | 1.018770 | 1.282664 | 1.524919 |
| $Scl_r$ | 1.071005 | 1.633064 | 0.989153 | 1.358493 | 1.068684 | 1.007181 | 1.110276 | 1.023141 |
| $Spl_r$ | 1.094733 | 0.583515 | 0.832685 | 0.989380 | 0.798938 | 0.892031 | 0.920927 | 0.718261 |
| $Sct_r$ | 0.979321 | 1.153846 | 1.016561 | 1.118773 | 1.000000 | 1.027211 | 1.076923 | 1.000000 |
| $Spt_r$ | 1.230064 | 1.256251 | 0.978456 | 0.809453 | 1.226450 | 1.204805 | 1.247178 | 1.106253 |
| $Scv_r$ | 1.093620 | 1.415322 | 0.973039 | 1.214271 | 1.068684 | 0.980501 | 1.030971 | 1.023141 |
| $Spv_r$ | 1.074729 | 0.929321 | 0.965226 | 1.143978 | 1.008793 | 0.889345 | 0.970554 | 0.888043 |
| $Pr_{av}$ | 3.1 | 6.35 | 3.075 | 7.4 | 4.9625 | 3.425 | 5.65 | 4.475 |
| $Pr_{sd}$ | 0.48 | 0.45 | 0.57 | 0.4 | 0.39 | 0.29 | 0.54 | 0.17 |
| $T$ | 108 | 110 | 120 | 108 | 110 | 109 | 114 | 132 |
| $Age$ | 28 | 36 | 26 | 41 | 23 | 31 | 25 | 28 |

### B.1.2.2 Normality Checks

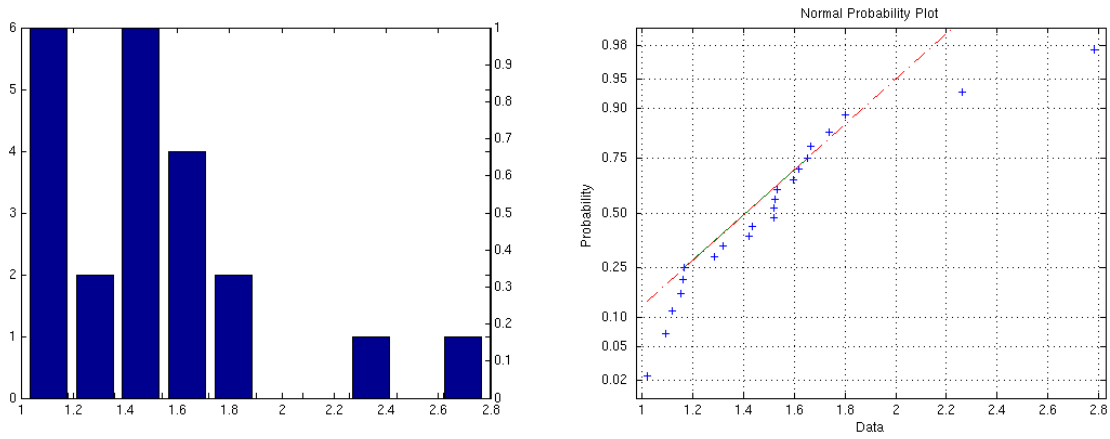



Figure B.375: $Sc_r$ parameter. Histogram (left), and normality plot (right).

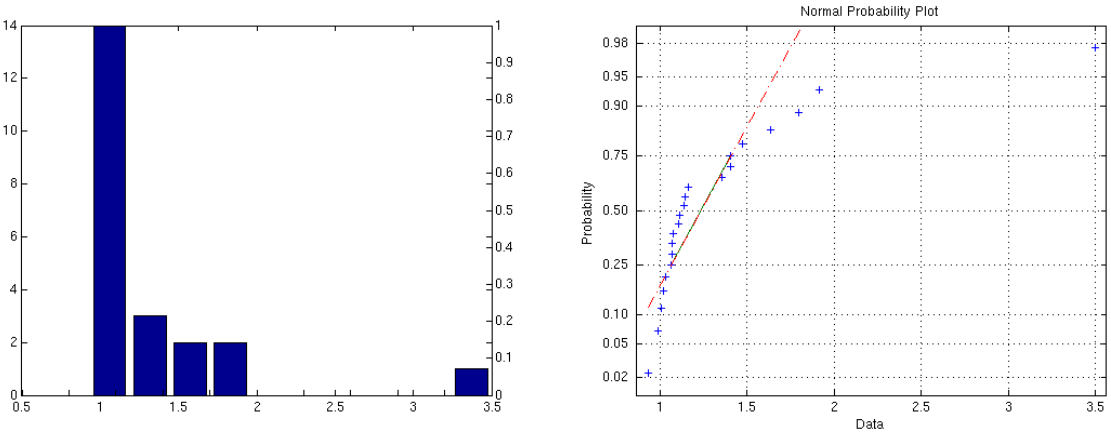



Figure B.376: $Scl_r$ parameter. Histogram (left), and normality plot (right).

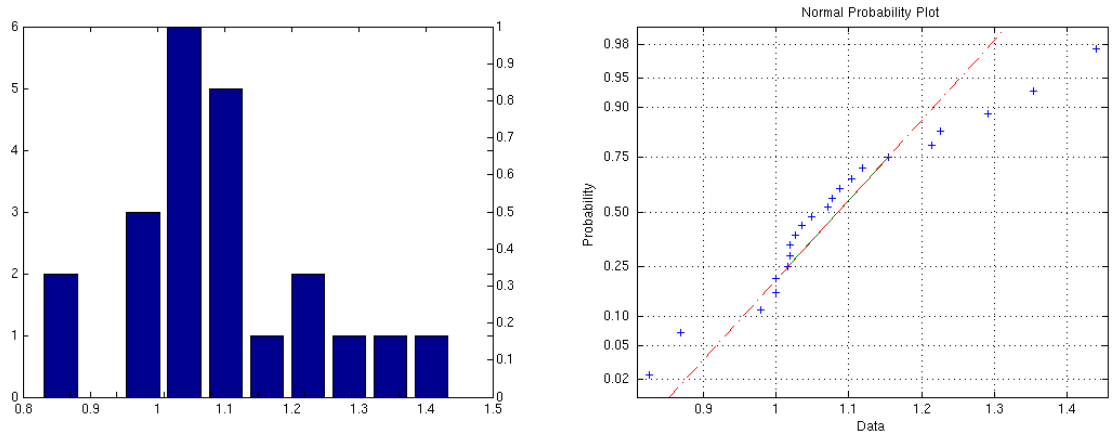



Figure B.377: $Sct_r$ parameter. Histogram (left), and normality plot (right).

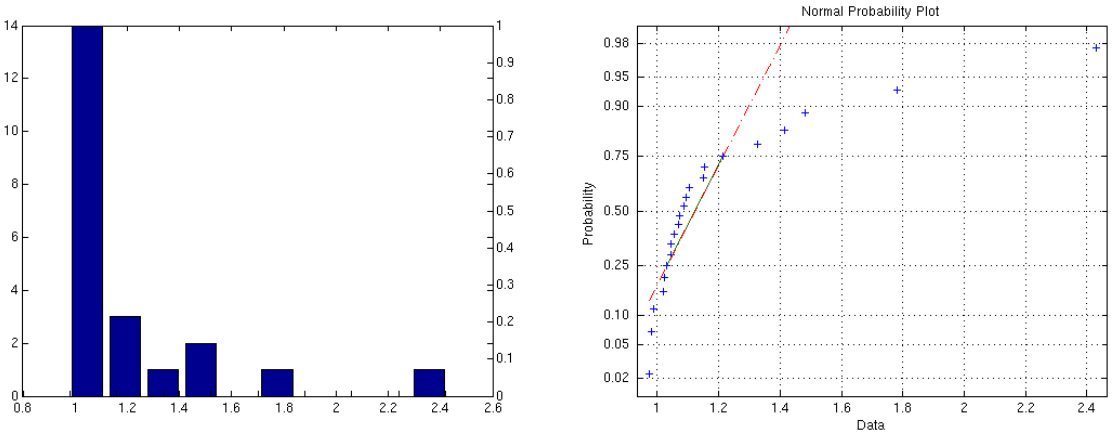



Figure B.378: $Scv_r$ parameter. Histogram (left), and normality plot (right).

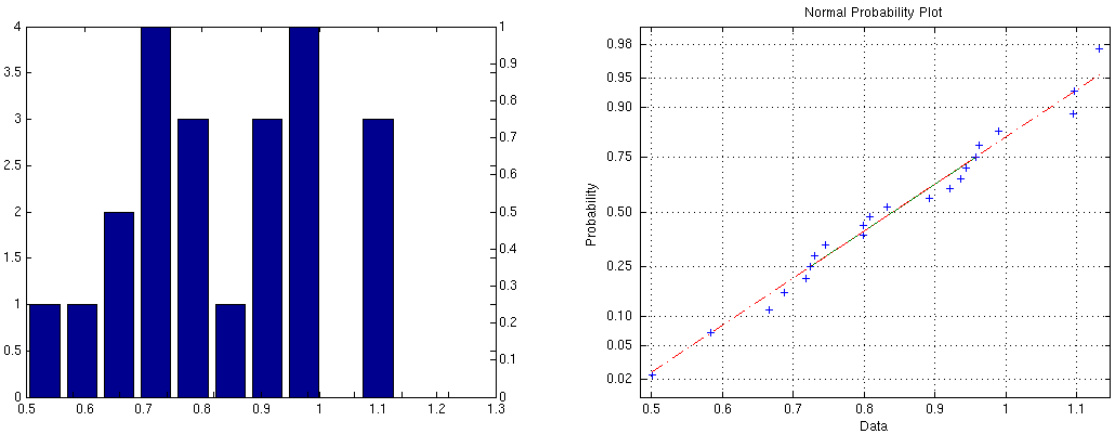



Figure B.379: $Spl_r$ parameter. Histogram (left), and normality plot (right).

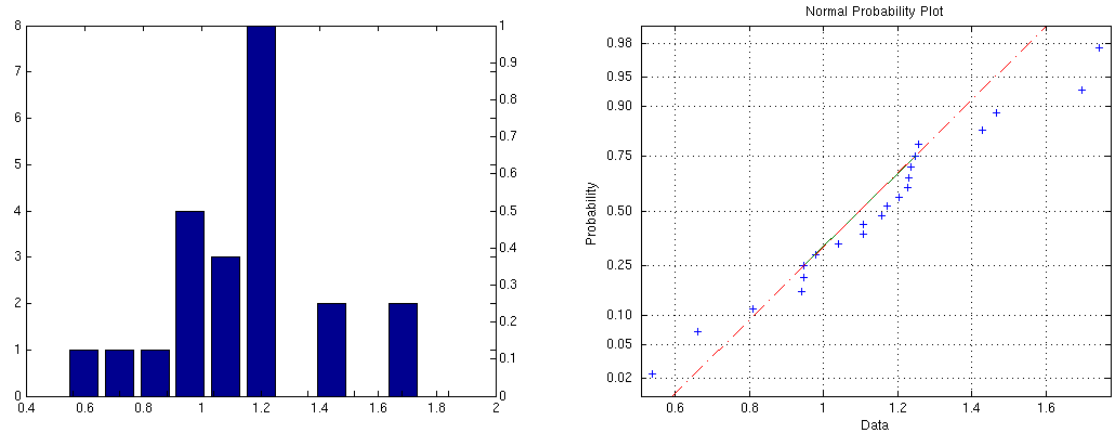



Figure B.380: $Spt_r$ parameter. Histogram (left), and normality plot (right).

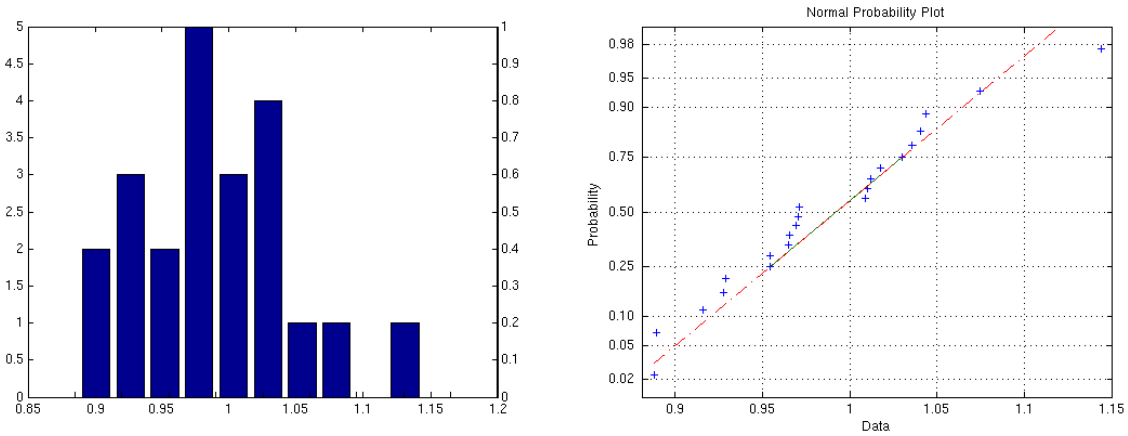



Figure B.381: $Spv_r$ parameter. Histogram (left), and normality plot (right).

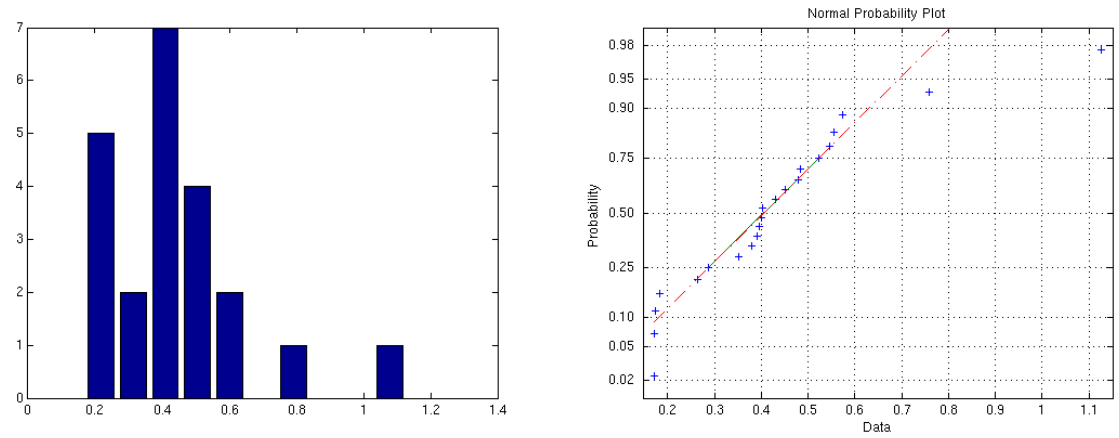



Figure B.382: $Pr$ parameter. Histogram (left), and normality plot (right).

#### B.1.2.3 Bootstrapping

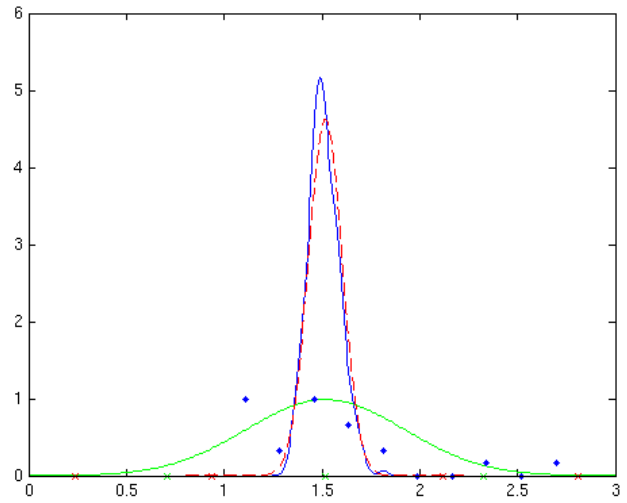



Figure B.383: $Sc_r$ parameter.

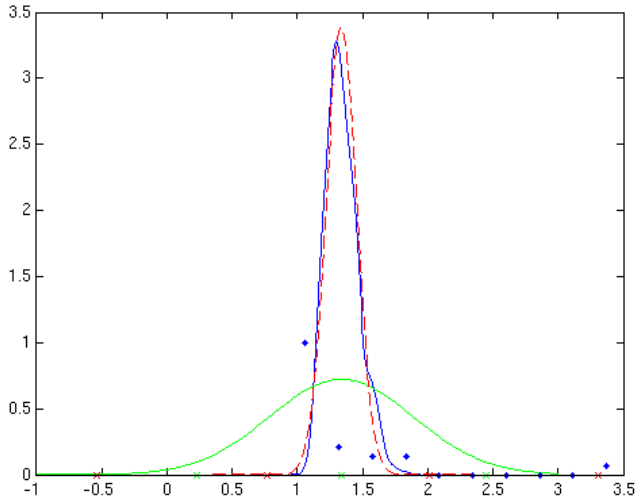



Figure B.384: $Scl_r$ parameter.

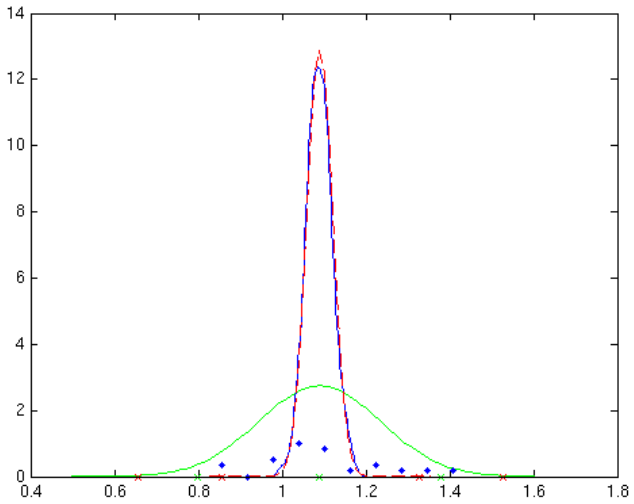



Figure B.385: $Sct_r$ parameter.

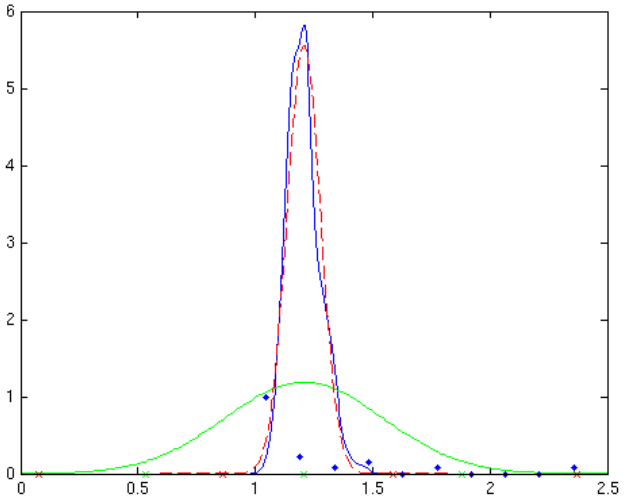



Figure B.386: $Scv_r$ parameter.

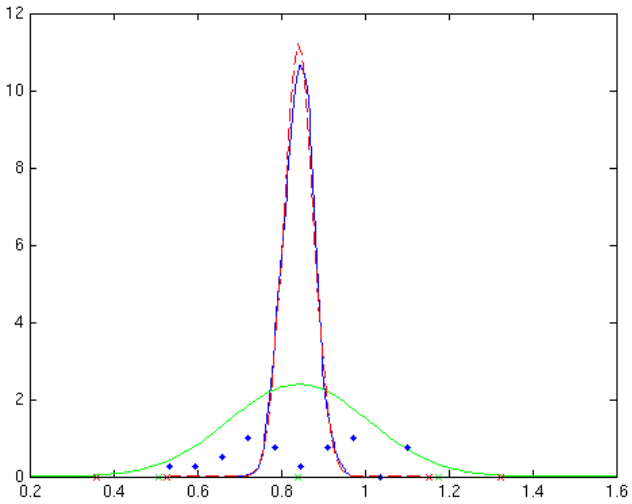



Figure B.387: $Spl_r$ parameter.

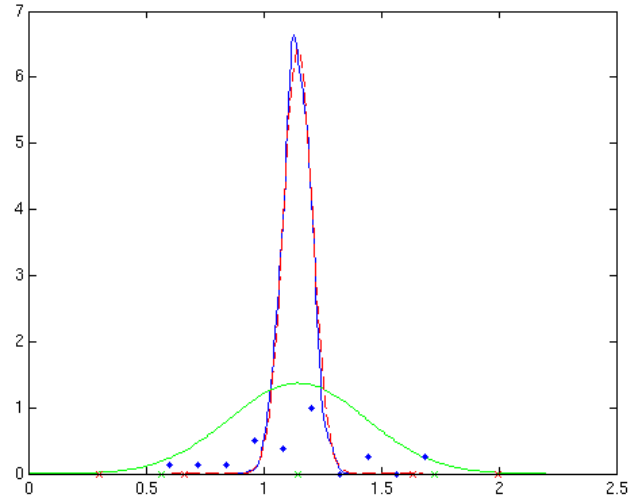



Figure B.388: $Spt_r$ parameter.

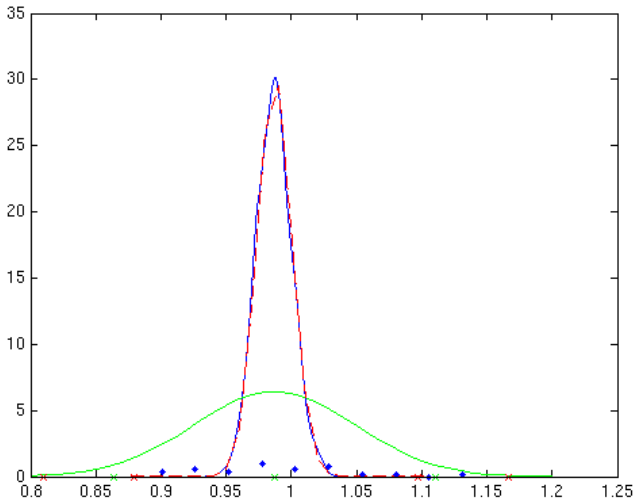



Figure B.389: $Spv_r$ parameter.

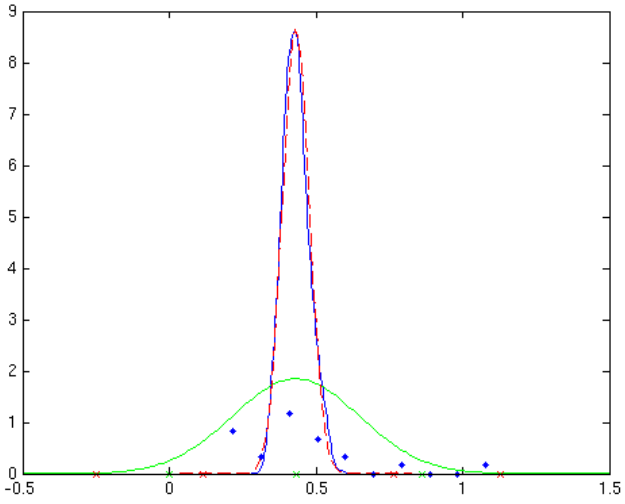



Figure B.390: $Pr$ parameter.

## B.2 Synthetic Trials

### B.2.1 Individual Trials

### B.2.1.1 Trial 1

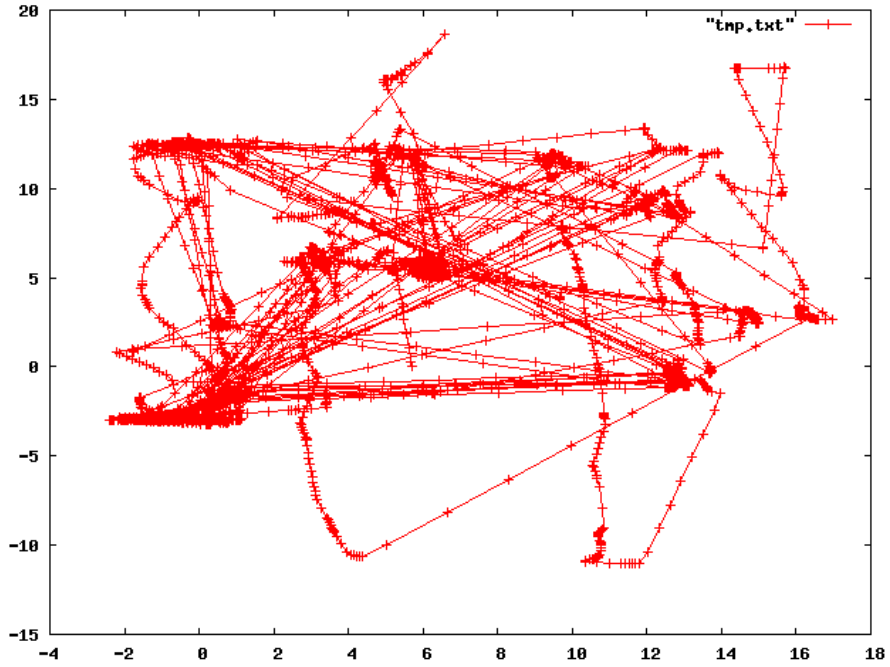



Figure B.391: Complete scan path, Trial 1.

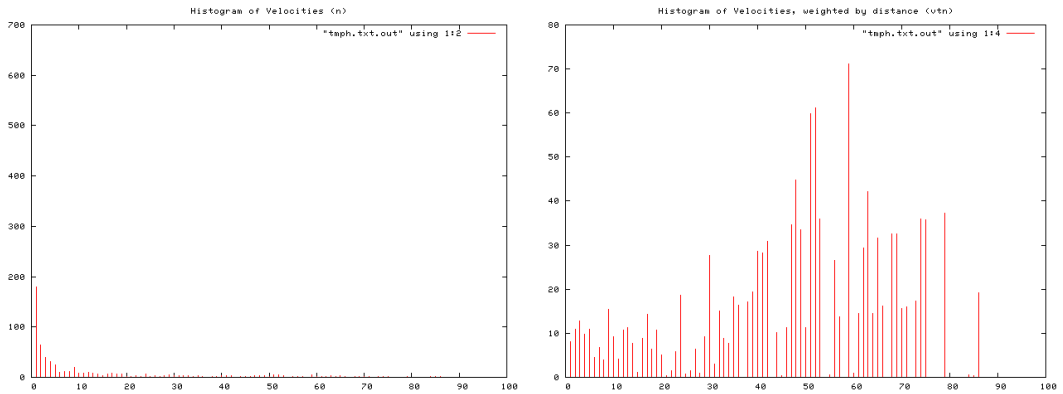



Figure B.392: Histogram of velocity magnitudes, Trial 1 (left). Histogram of distance weighted velocities, Trial 1 (right).

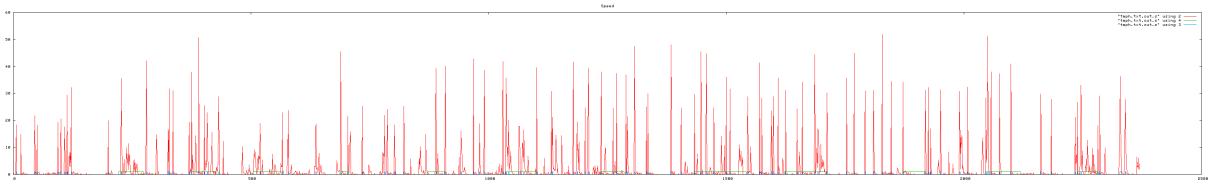


Figure B.393: Velocity profile. Velocity magnitude of each frame, Trial 1.

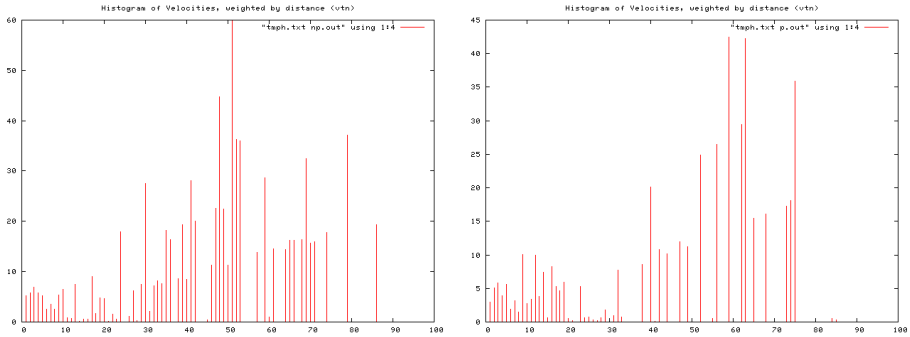



Figure B.394: Histogram of velocities during periods of no perturbation (left), and perturbation (right), Trial 1.

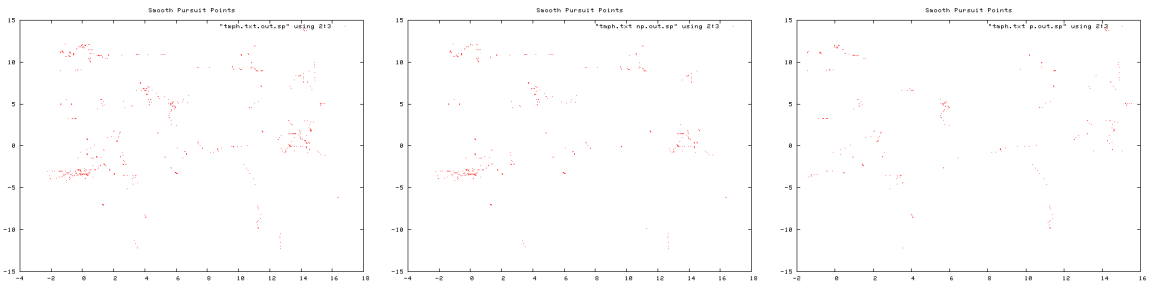



Figure B.395: Smooth pursuit gaze locations, Trial 1. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

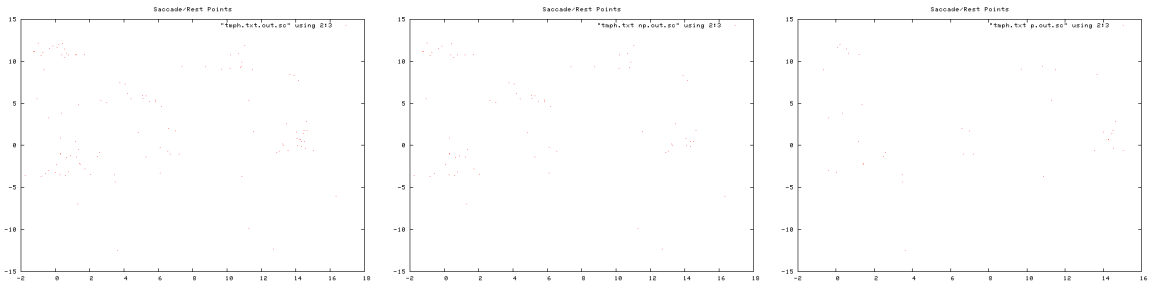



Figure B.396: Saccade gaze locations, Trial 1. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

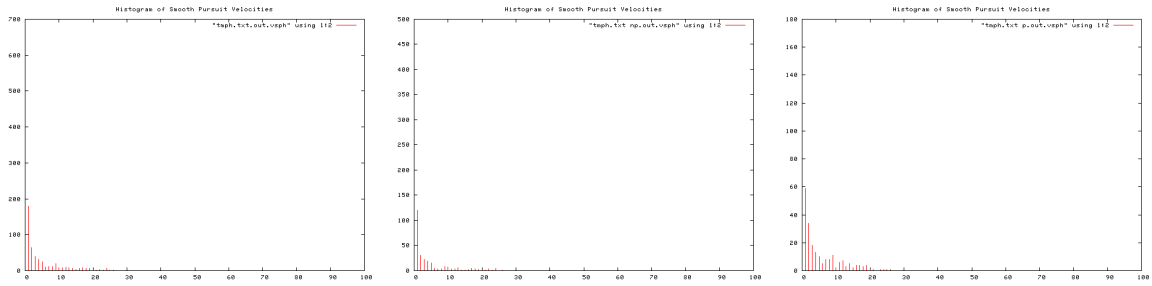



Figure B.397: Histogram of smooth pursuit velocities, Trial 1. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

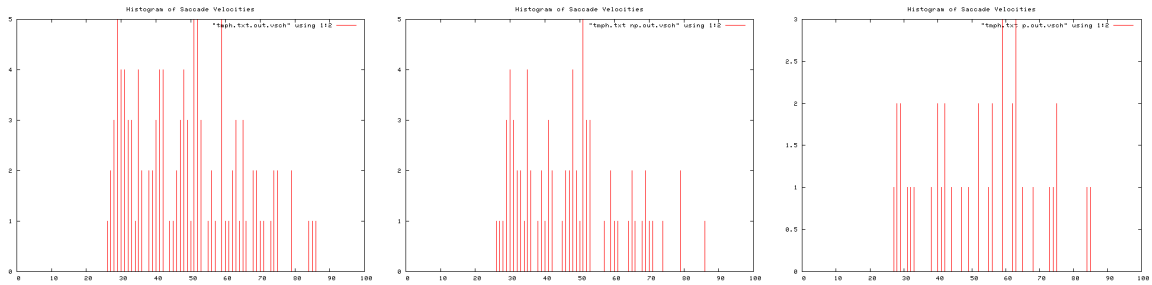



Figure B.398: Histogram of Saccade velocities, Trial 1. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

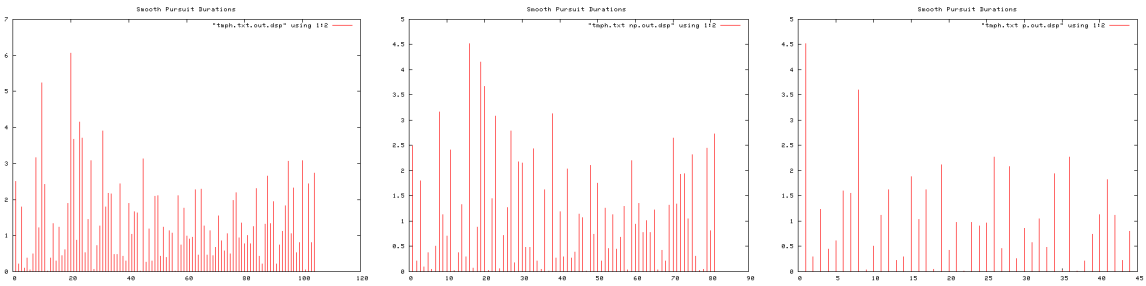



Figure B.399: Smooth pursuit durations, Trial 1. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

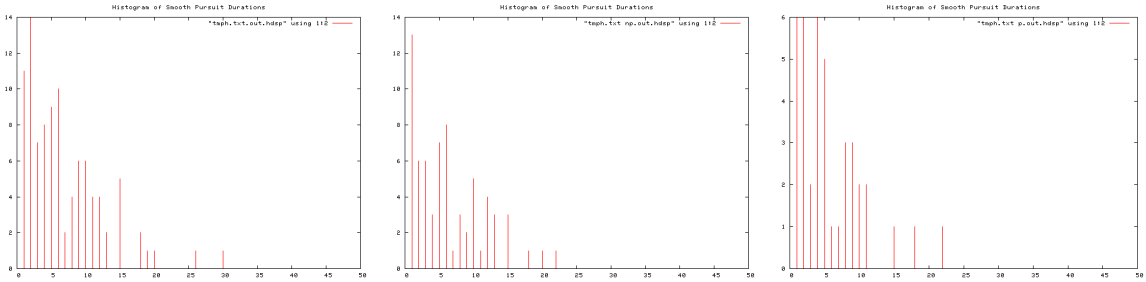



Figure B.400: Histogram of Smooth pursuit durations, Trial 1. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

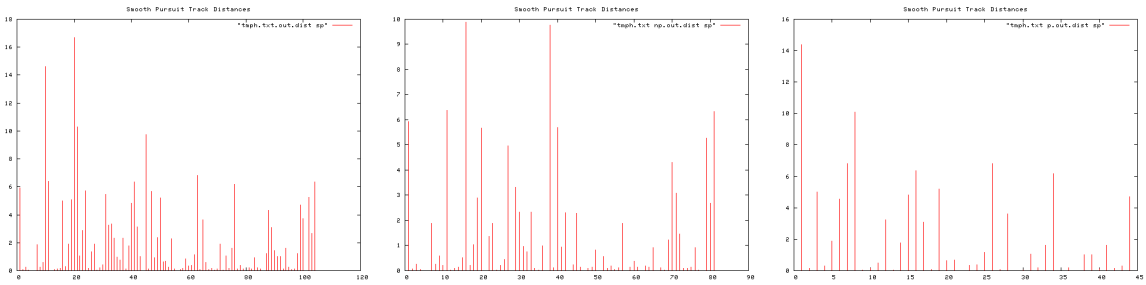



Figure B.401: Smooth pursuit distances, Trial 1. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

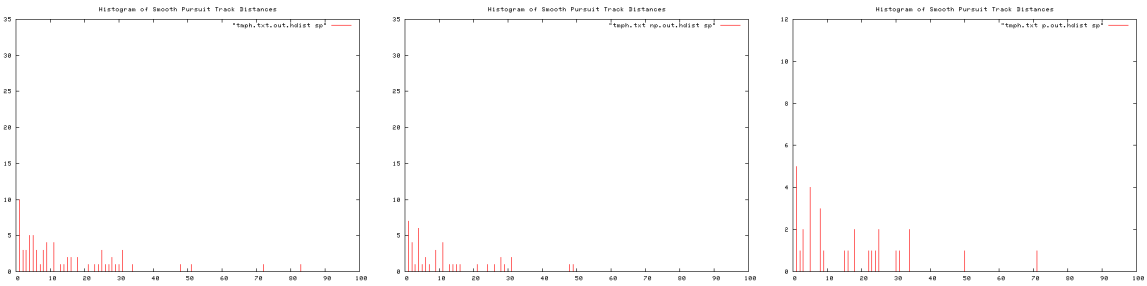



Figure B.402: Histogram of smooth pursuit distances, Trial 1. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

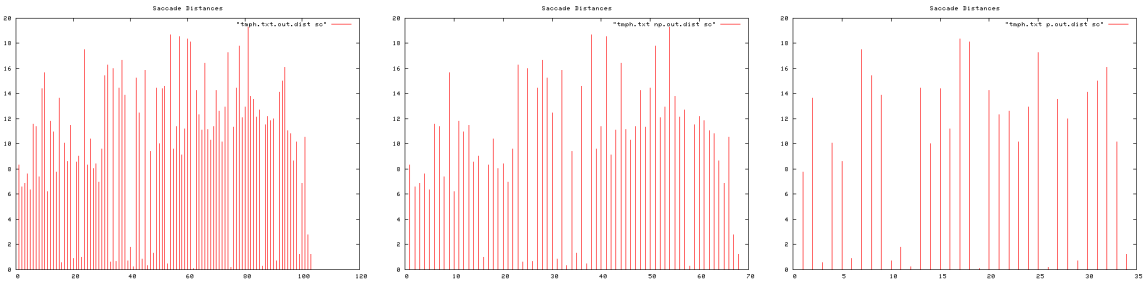



Figure B.403: Saccade distances, Trial 1. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

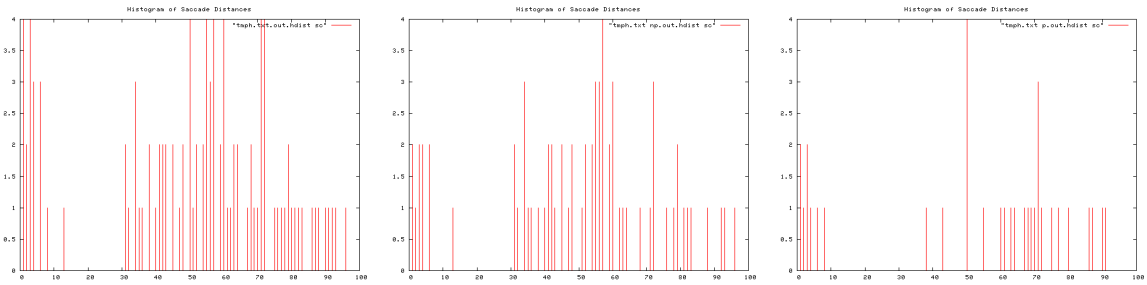



Figure B.404: Histogram of saccade distances, Trial 1. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

| Ti | To | T | I/O Event | Pear | Apple | Peach | Orange | |
|---|---|---|---|---|---|---|---|---|
| 20 | 27 | 7 | Orange In | | | | 1 | |
| 30 | 36 | 6 | Pear In | 1 | | | 2 | |
| 42 | 53 | 11 | | 1 | | | 1 | |
| 54 | 1:04 | 10 | Peach In | 3 | | 2 | 2 | |
| 1:10 | 1:19 | 9 | | 1 | | 2 | 3 | |
| 1:26 | 1:35 | 9 | | 2 | | 1 | 3 | |
| 1:41 | 1:49 | 8 | | 2 | | 2 | 2 | |
| 1:58 | 2:05 | 7 | Apple In, Peach Out | 2 | 1 | 1 | 2 | |
| 2:12 | 2:24 | 12 | Orange Out | 2 | 3 | | 2 | |
| 2:26 | 2:36 | 10 | Pear Out | 2 | 3 | | | |
| 2:44 | 2:50 | 6 | Apple Out | | 1 | | | |
| | | | TOTAL Rets | 16 | 8 | 8 | 18 | |
| | | | TOTAL T | 82 | 35 | 43 | 79 | **SD** |
| | | | **Av. Re-attention Period** | **5.1** | **4.4** | **5.4** | **4.4** | 0.5057997 |

Figure B.405: Re-attention period statistics, Trial 1.

### B.2.1.2 Trial 2

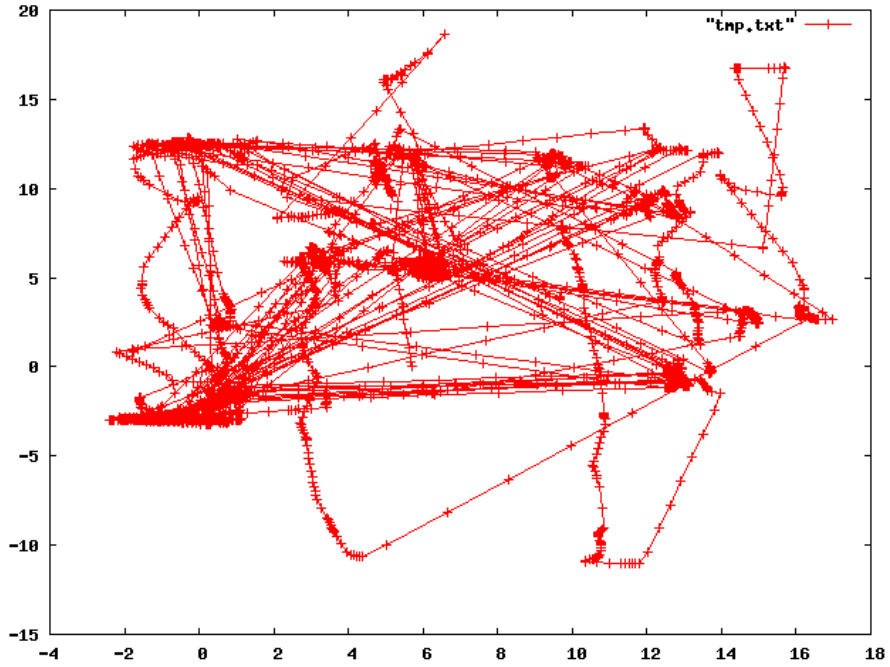



Figure B.406: Complete scan path, Trial 2.

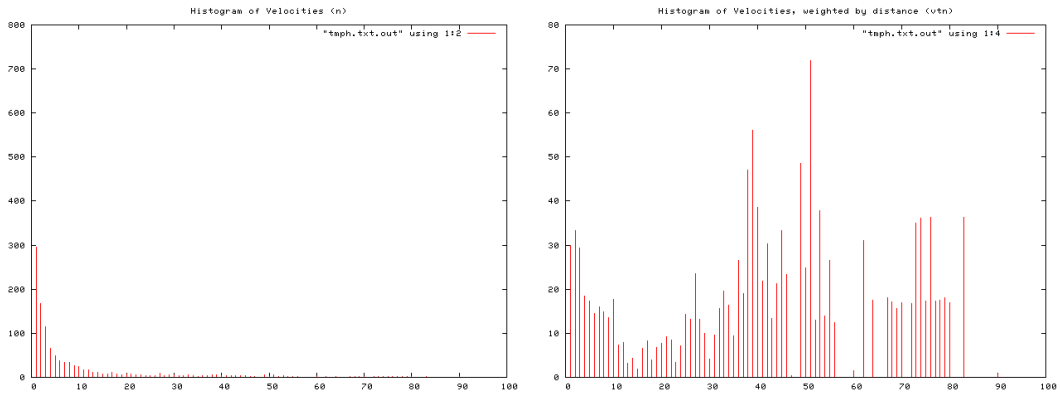



Figure B.407: Histogram of velocity magnitudes, Trial 2 (left). Histogram of distance weighted velocities, Trial 1 (right).

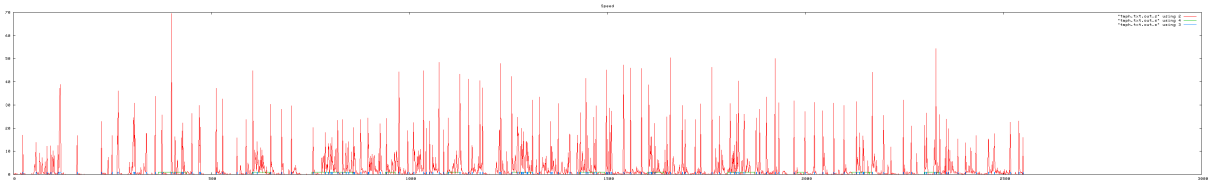


Figure B.408: Velocity profile. Velocity magnitude of each frame, Trial 2.

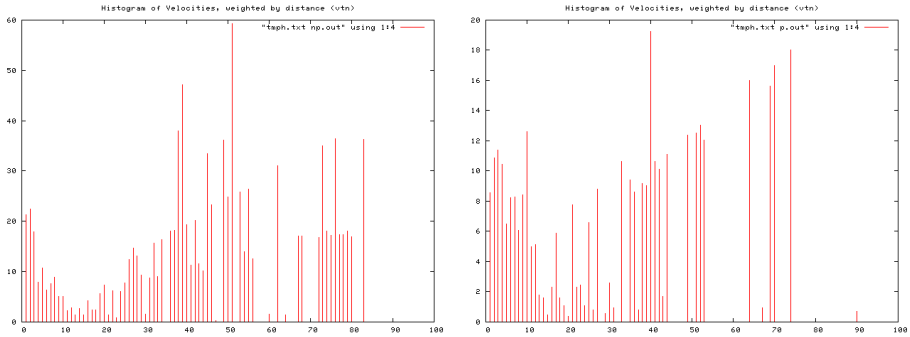



Figure B.409: Histogram of velocities during periods of no perturbation (left), and perturbation (right), Trial 2.

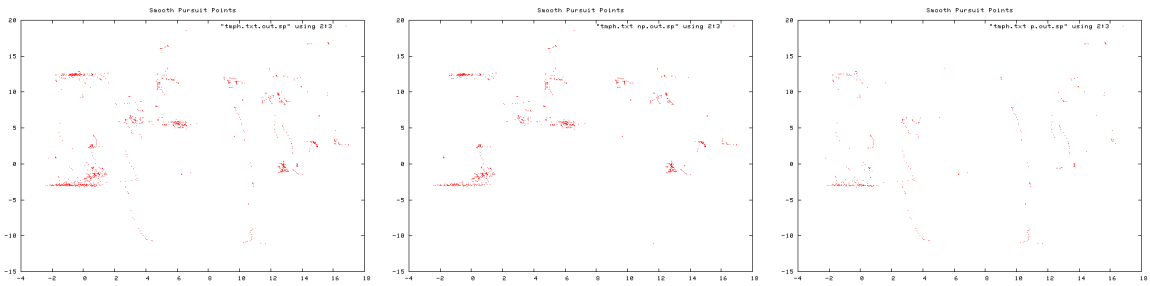



Figure B.410: Smooth pursuit gaze locations, Trial 2. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

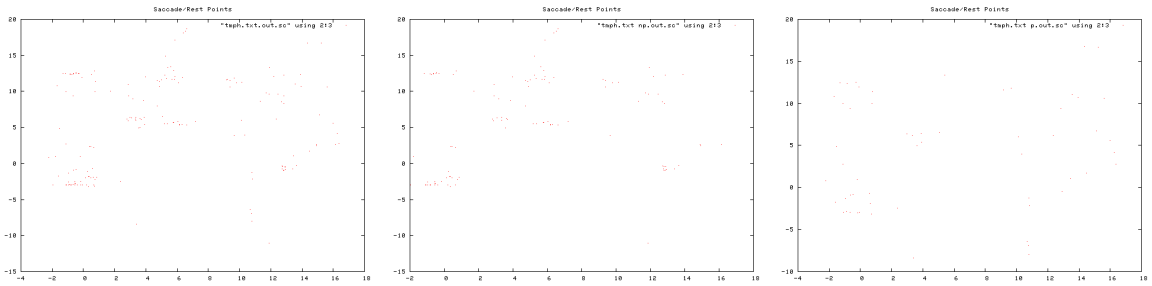



Figure B.411: Saccade gaze locations, Trial 2. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

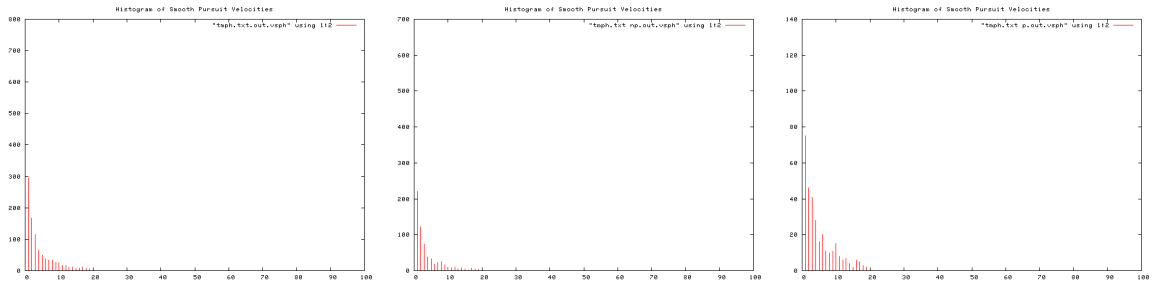



Figure B.412: Histogram of smooth pursuit velocities, Trial 2. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

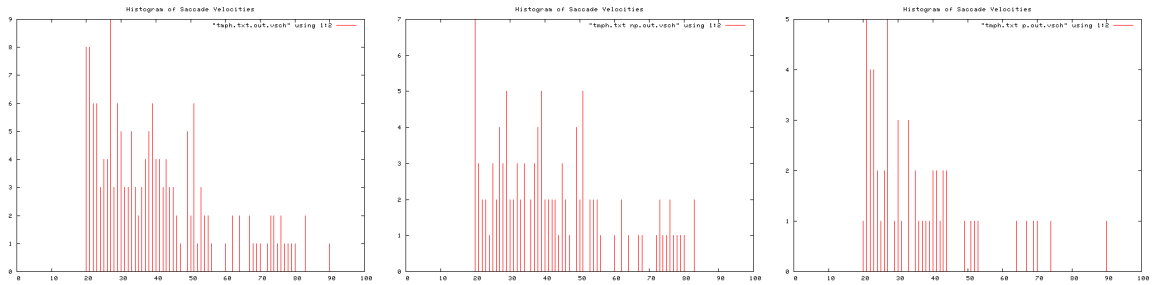



Figure B.413: Histogram of Saccade velocities, Trial 2. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

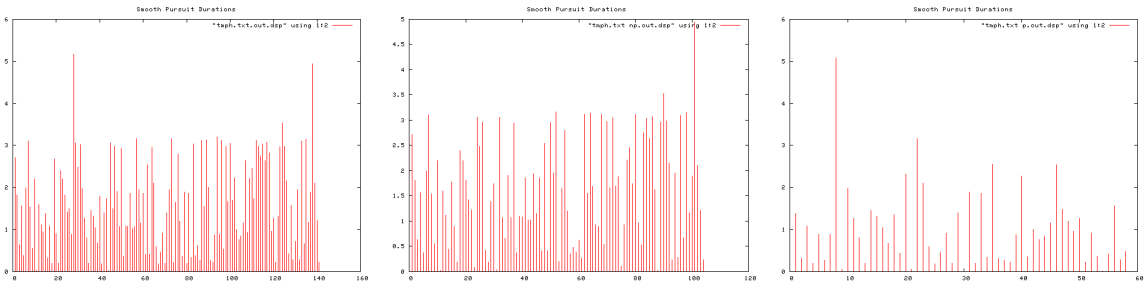



Figure B.414: Smooth pursuit durations, Trial 2. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

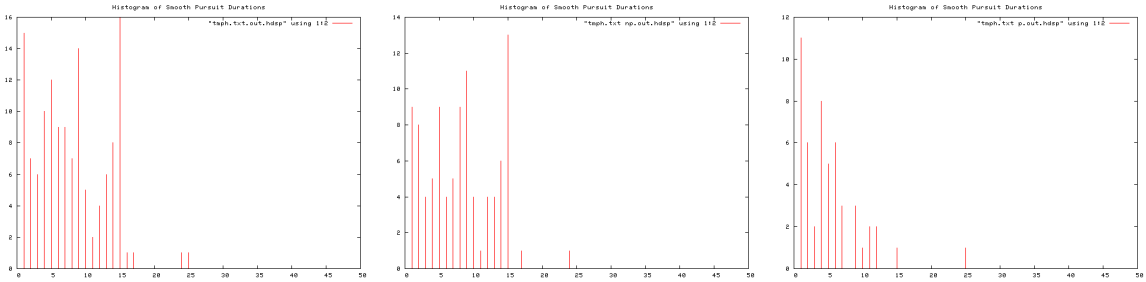



Figure B.415: Histogram of Smooth pursuit durations, Trial 2. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

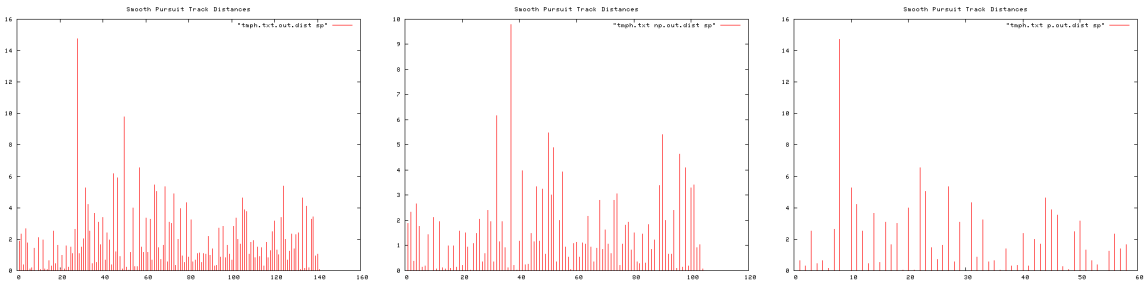



Figure B.416: Smooth pursuit distances, Trial 2. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

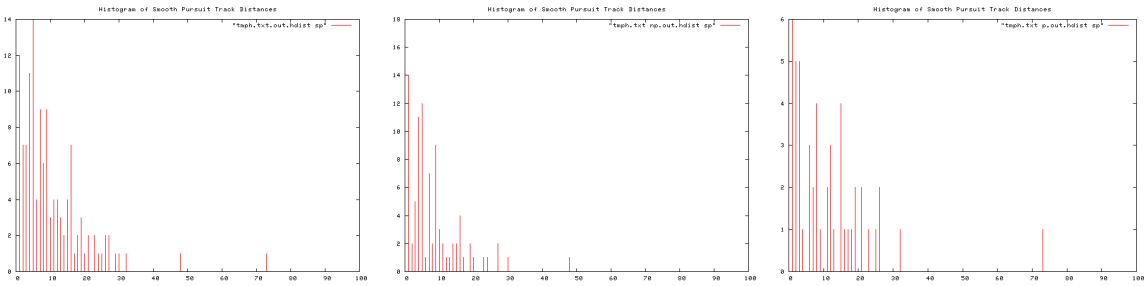



Figure B.417: Histogram of smooth pursuit distances, Trial 2. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

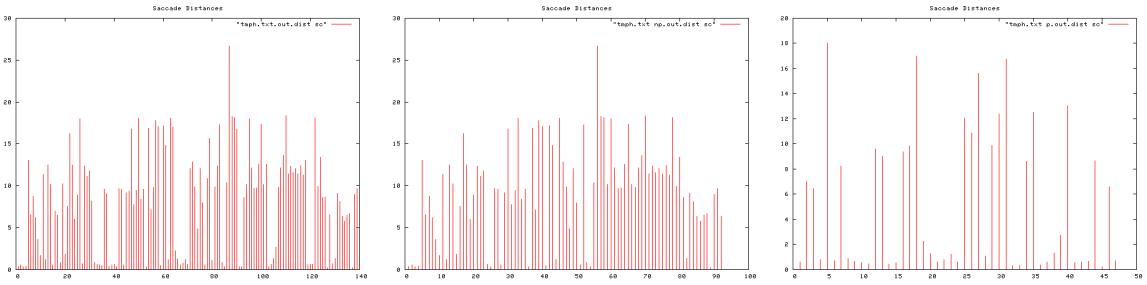



Figure B.418: Saccade distances, Trial 2. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

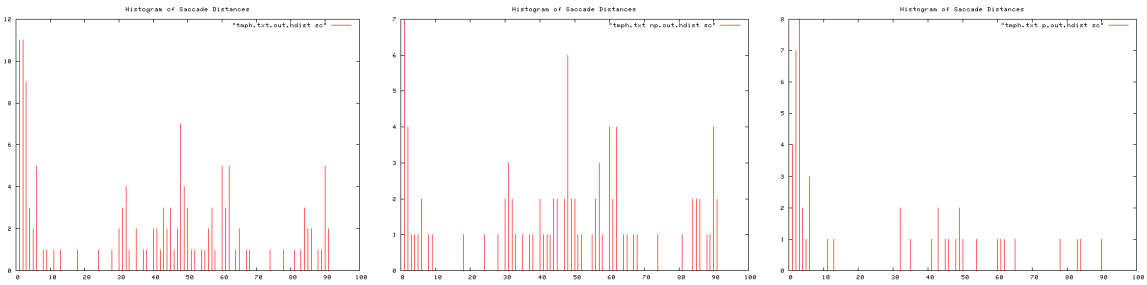



Figure B.419: Histogram of saccade distances, Trial 2. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

| Ti | To | T | I/O Event | Pear | Apple | Peach | Orange | |
|---|---|---|---|---|---|---|---|---|
| 33 | 46 | 13 | Orange In | | | | 2 | |
| 53 | 1:02 | 9 | Pear In | 1 | | | 1 | |
| 1:07 | 1:18 | 11 | | 2 | | | 2 | |
| 1:23 | 1:42 | 9 | Peach In | 4 | | 3 | 3 | |
| 1:47 | 2:01 | 14 | | 3 | | 1 | 3 | |
| 2:05 | 2:20 | 15 | | 2 | | 3 | 2 | |
| 2:24 | 2:39 | 15 | | 2 | | 2 | 3 | |
| 2:47 | 3:06 | 19 | Apple In, Peach Out | 2 | 3 | 1 | 3 | |
| 3:13 | 3:26 | 13 | Orange Out | 2 | 2 | | 1 | |
| 3:31 | 3:45 | 14 | Pear Out | 2 | 2 | | | |
| 3:50 | 4:03 | 13 | Apple Out | | 2 | | | |
| | | | TOTAL Rets | 20 | 9 | 10 | 20 | |
| | | | TOTAL T | 119 | 59 | 72 | 118 | **SD** |
| | | | **Av. Re-attention Period** | **6** | **6.6** | **7.2** | **5.9** | 0.60207973 |

Figure B.420: Re-attention period statistics, Trial 2.

### B.2.1.3 Trial 3

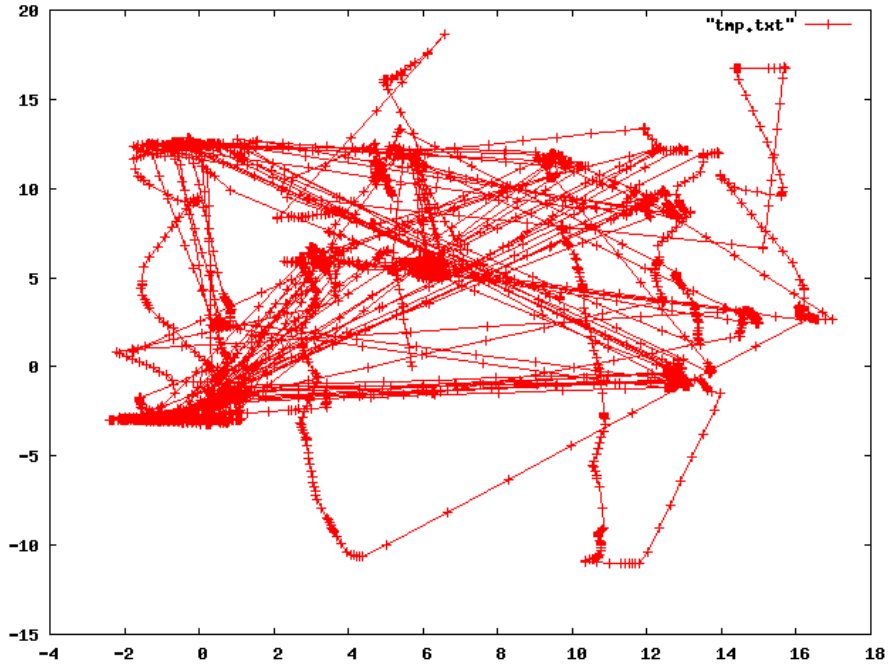



Figure B.421: Complete scan path, Trial 3.

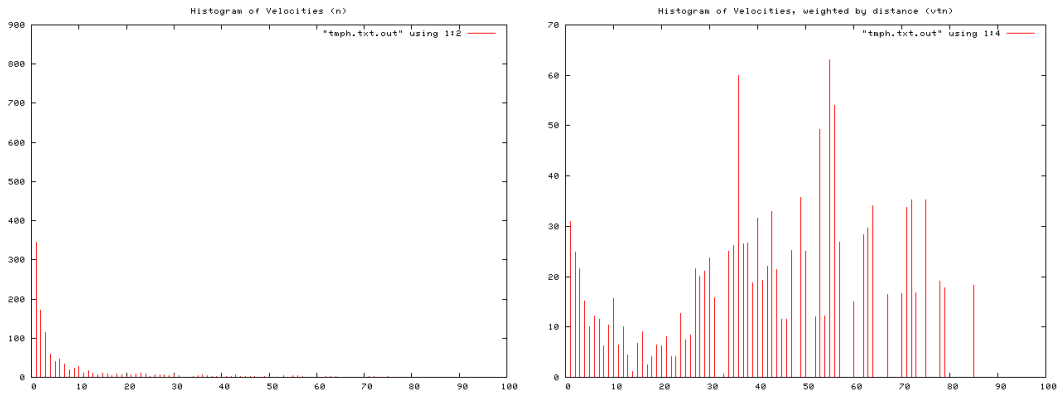



Figure B.422: Histogram of velocity magnitudes, Trial 3 (left). Histogram of distance weighted velocities, Trial 3 (right).

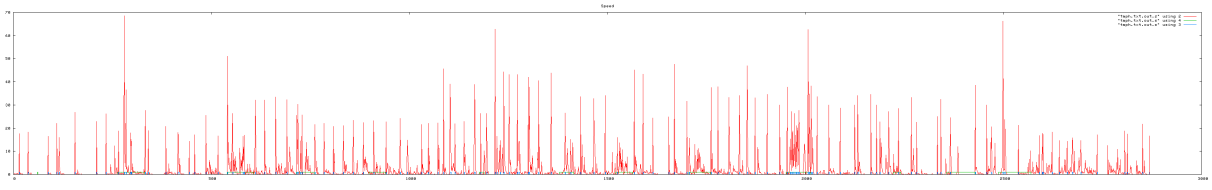


Figure B.423: Velocity profile. Velocity magnitude of each frame, Trial 3.

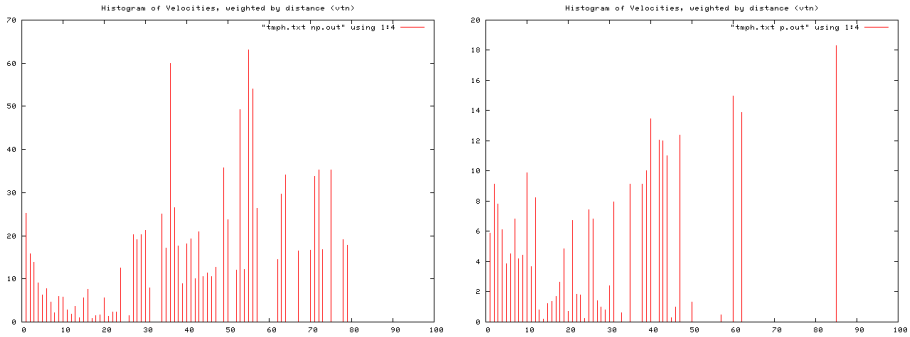



Figure B.424: Histogram of velocities during periods of no perturbation (left), and perturbation (right), Trial 3.

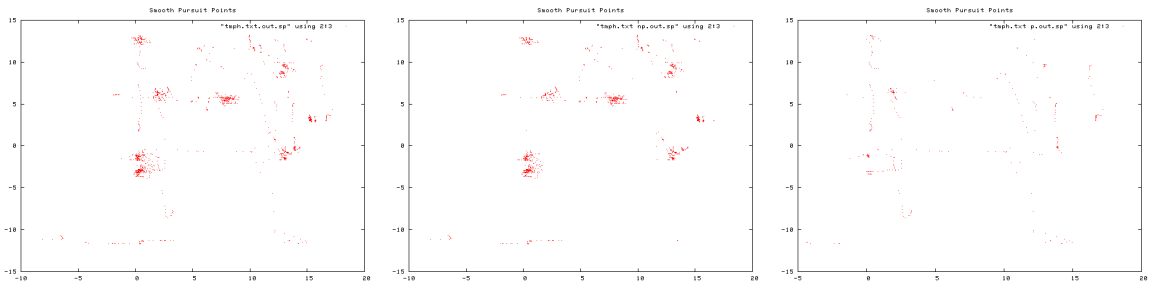



Figure B.425: Smooth pursuit gaze locations, Trial 3. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

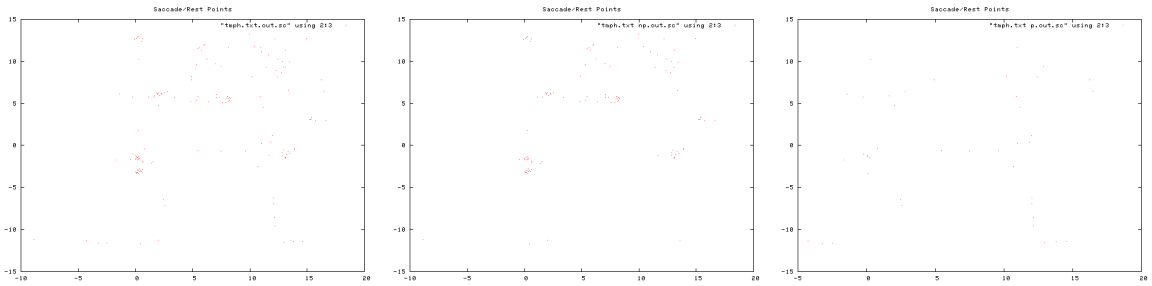



Figure B.426: Saccade gaze locations, Trial 3. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

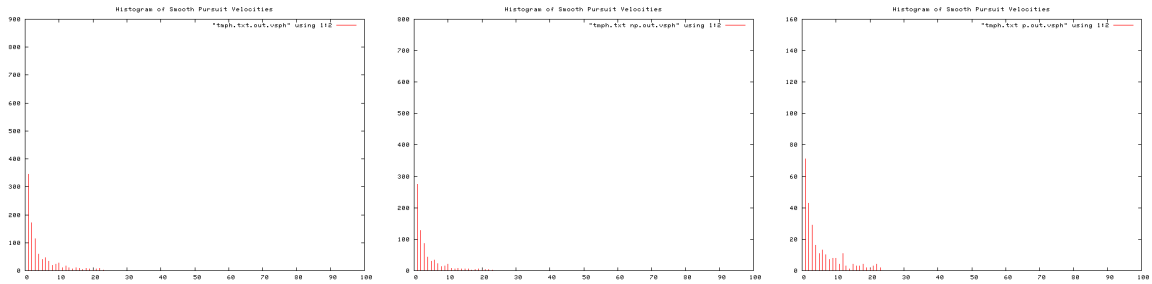



Figure B.427: Histogram of smooth pursuit velocities, Trial 3. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

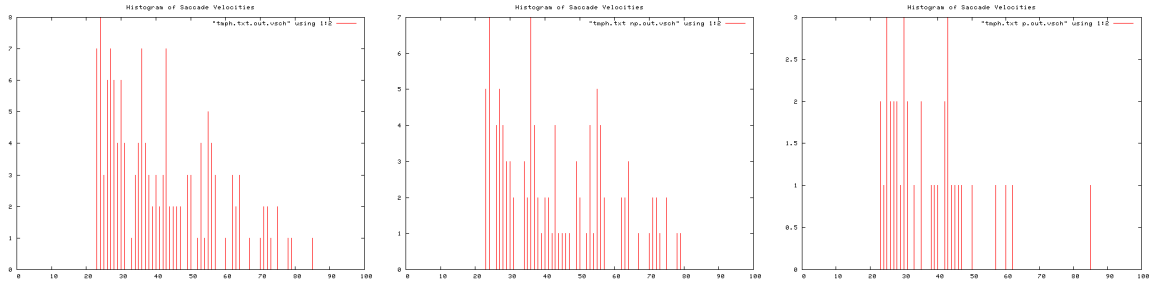



Figure B.428: Histogram of Saccade velocities, Trial 3. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

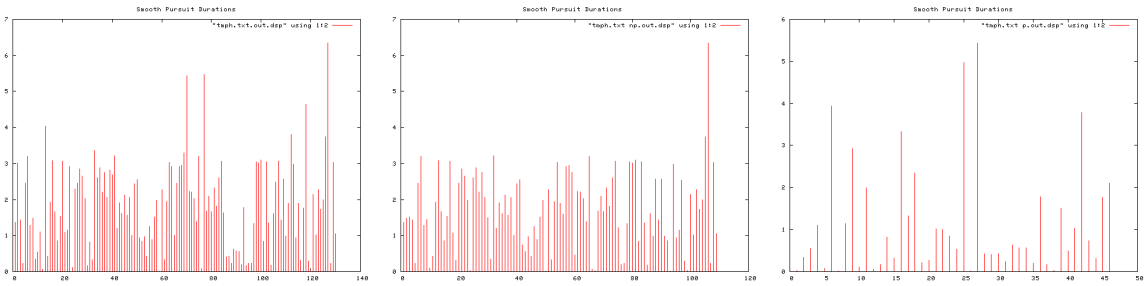



Figure B.429: Smooth pursuit durations, Trial 3. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

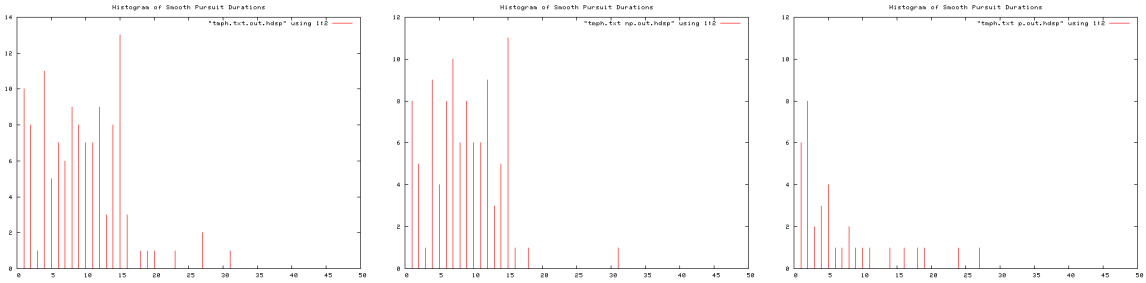



Figure B.430: Histogram of Smooth pursuit durations, Trial 3. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

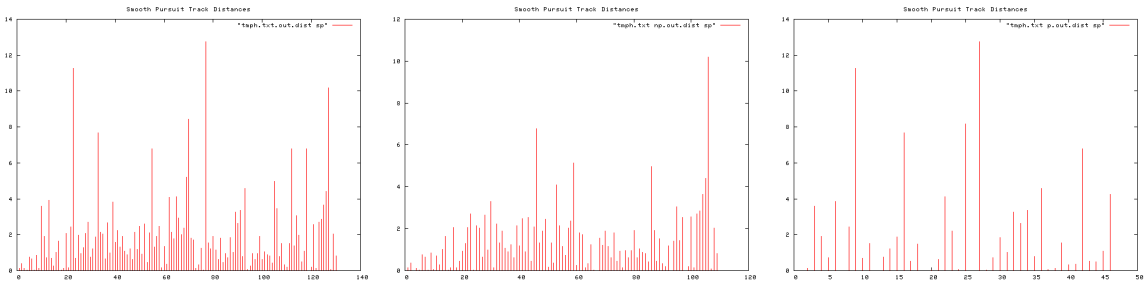



Figure B.431: Smooth pursuit distances, Trial 3. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

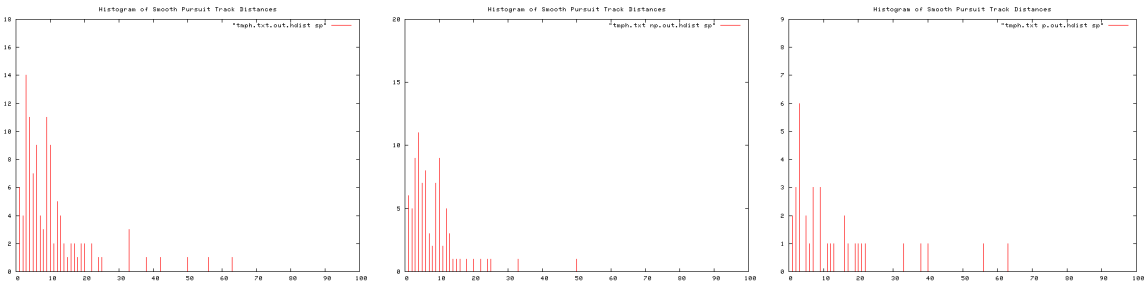



Figure B.432: Histogram of smooth pursuit distances, Trial 3. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

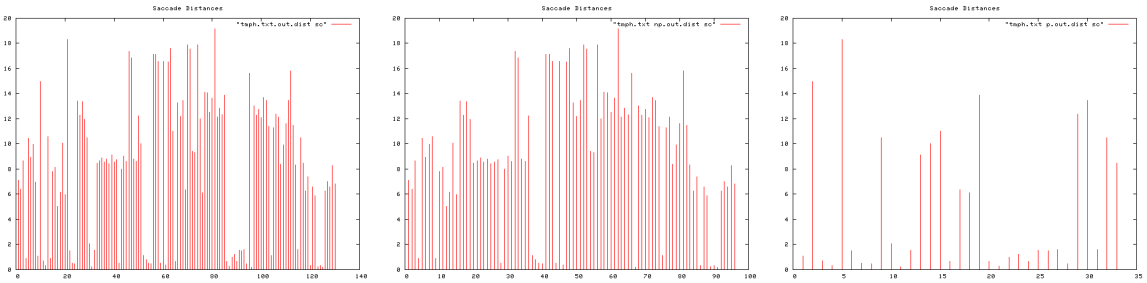



Figure B.433: Saccade distances, Trial 3. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

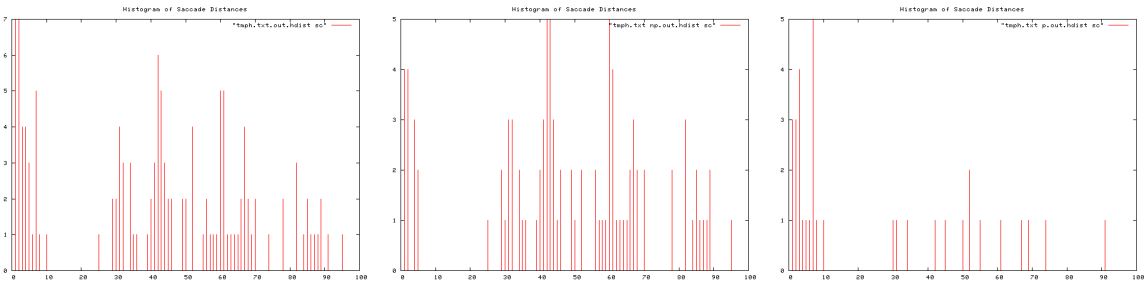



Figure B.434: Histogram of saccade distances, Trial 3. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

| Ti | To | T | I/O Event | Pear | Apple | Peach | Orange | |
|---|---|---|---|---|---|---|---|---|
| 24 | 43 | 19 | Orange In | | | | 2 | |
| 49 | 1:01 | 12 | Pear In | 3 | | | 3 | |
| 1:07 | 1:20 | 13 | | 3 | | | 2 | |
| 1:27 | 1:50 | 23 | Peach In | 3 | | 4 | 5 | |
| 1:55 | 2:11 | 16 | | 5 | | 4 | 3 | |
| 2:16 | 2:27 | 11 | | 2 | | 2 | 2 | |
| 2:32 | 2:46 | 14 | | 3 | | 1 | 2 | |
| 2:52 | 3:10 | 18 | Apple In, Peach Out | 3 | 2 | 2 | 1 | |
| 3:17 | 3:40 | 23 | Orange Out | 1 | 5 | | 3 | |
| 3:42 | 3:52 | 10 | Pear Out | 2 | 2 | | | |
| 3:58 | 4:03 | 5 | Apple Out | | 1 | | | |
| | | | TOTAL Rets | 25 | 10 | 13 | 23 | |
| | | | TOTAL T | 140 | 56 | 82 | 149 | **SD** |
| | | | **Av. Re-attention Period** | **5.6** | **5.6** | **6.3** | **6.5** | 0.46904158 |

Figure B.435: Re-attention period statistics, Trial 3.

### B.2.1.4 Trial 4

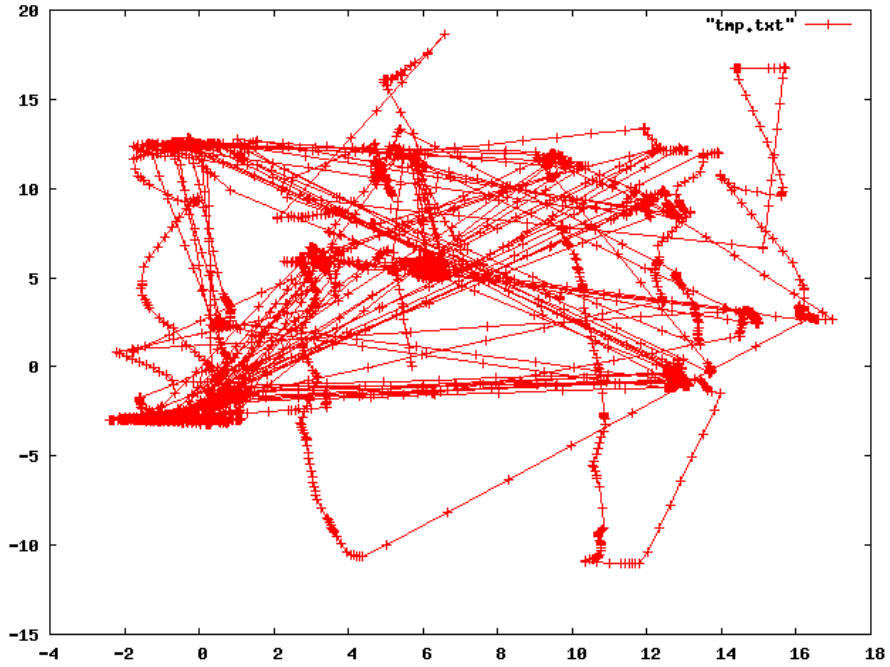



Figure B.436: Complete scan path, Trial 4.

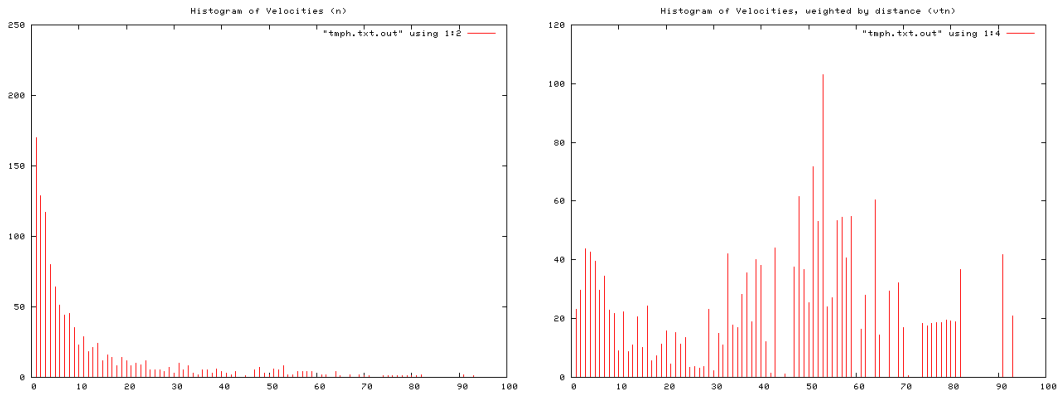



Figure B.437: Histogram of velocity magnitudes, Trial 4 (left). Histogram of distance weighted velocities, Trial 1 (right).

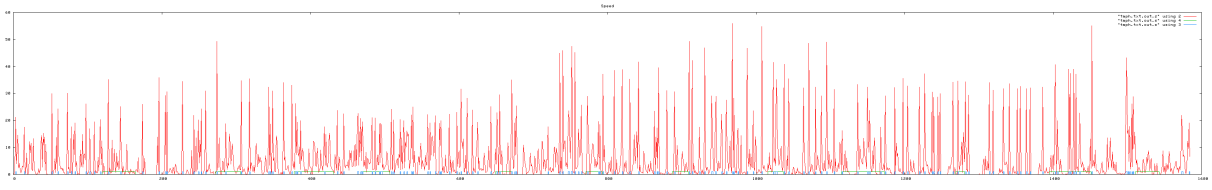


Figure B.438: Velocity profile. Velocity magnitude of each frame, Trial 4.

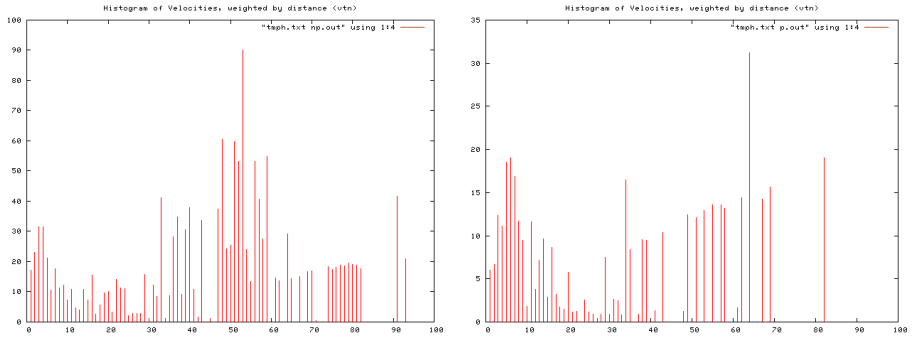



Figure B.439: Histogram of velocities during periods of no perturbation (left), and perturbation (right), Trial 4.

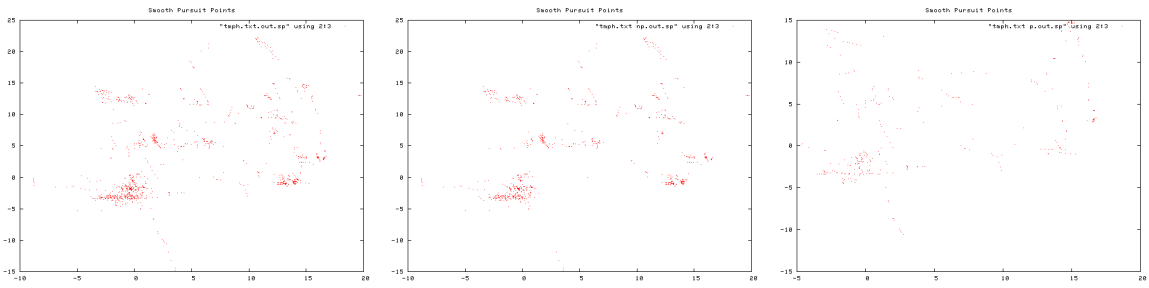



Figure B.440: Smooth pursuit gaze locations, Trial 4. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

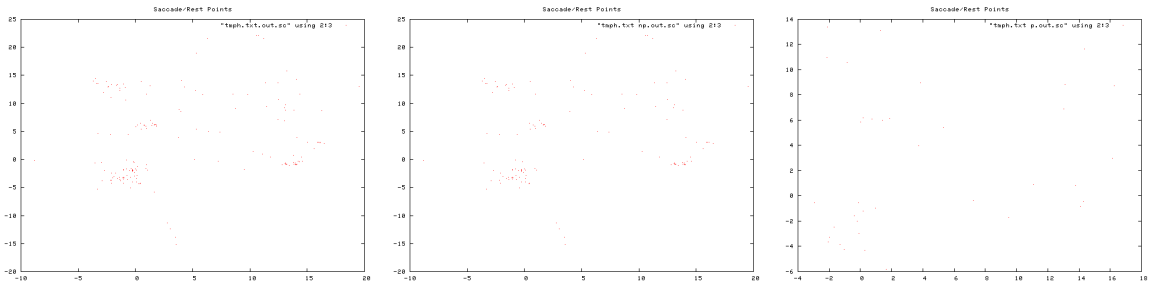



Figure B.441: Saccade gaze locations, Trial 4. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

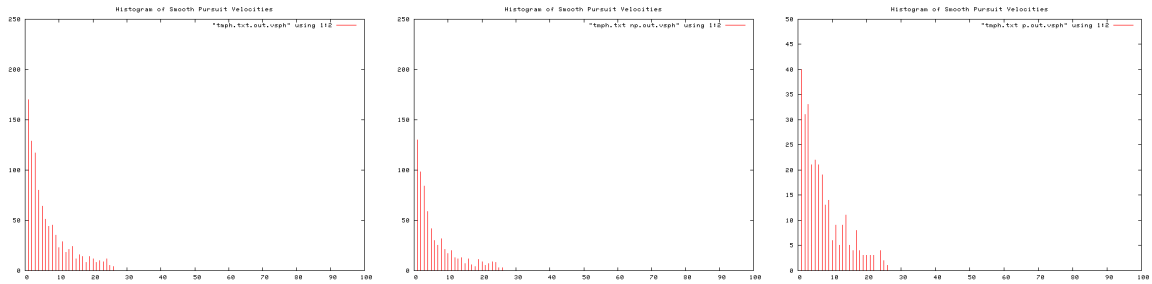



Figure B.442: Histogram of smooth pursuit velocities, Trial 4. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

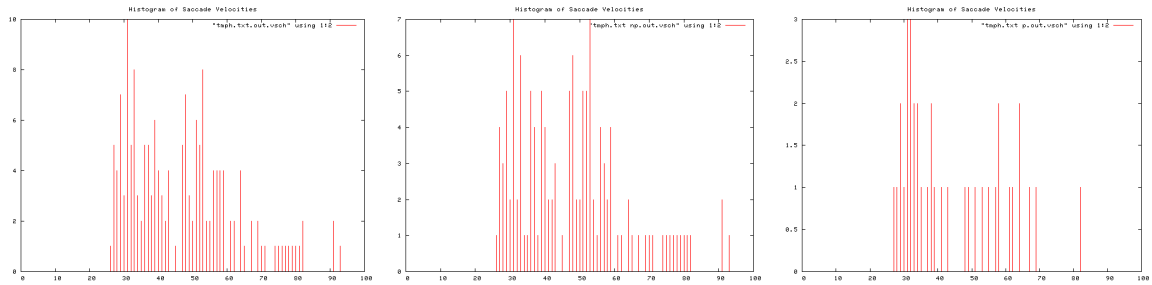



Figure B.443: Histogram of Saccade velocities, Trial 4. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

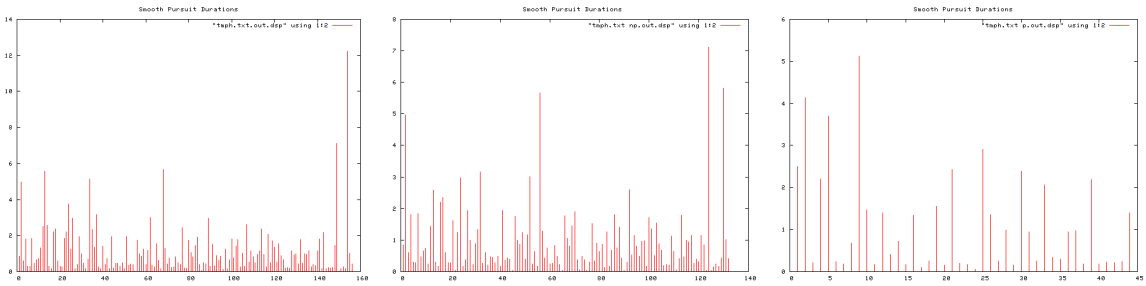



Figure B.444: Smooth pursuit durations, Trial 4. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

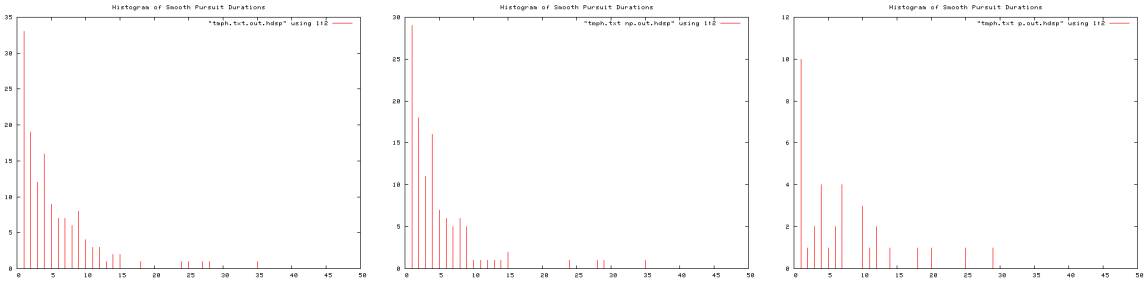



Figure B.445: Histogram of Smooth pursuit durations, Trial 4. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

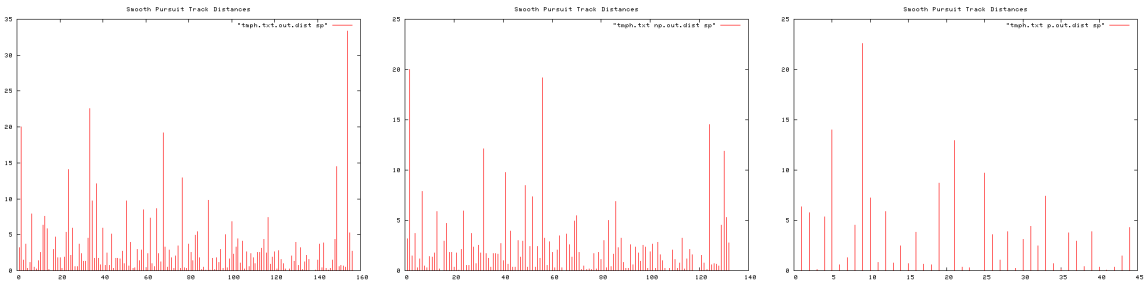



Figure B.446: Smooth pursuit distances, Trial 4. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

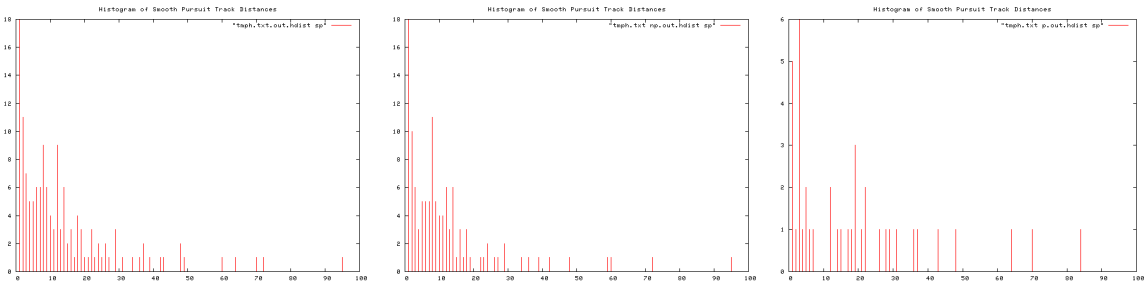



Figure B.447: Histogram of smooth pursuit distances, Trial 4. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

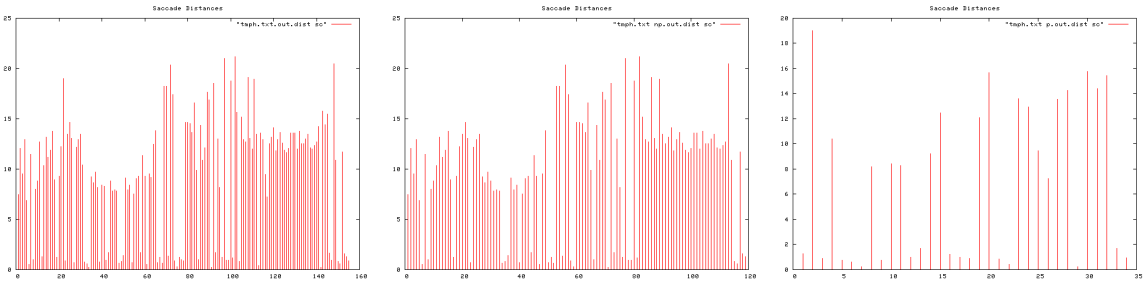



Figure B.448: Saccade distances, Trial 4. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

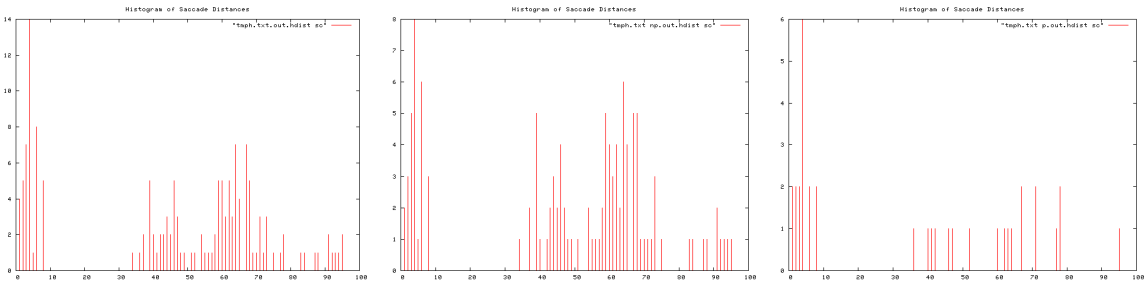



Figure B.449: Histogram of saccade distances, Trial 4. Over entire trial (left), during non-perturbation periods (middle), and during perturbation periods (right).

| Ti | To | T | I/O Event | Pear | Apple | Peach | Orange | |
|---|---|---|---|---|---|---|---|---|
| 21 | 38 | 17 | Orange In | | | | 3 | |
| 45 | 0:55 | 10 | Pear In | 4 | | | 3 | |
| 1:01 | 1:12 | 11 | | 2 | | | 3 | |
| 1:18 | 1:37 | 9 | Peach In | 5 | | 1 | 6 | |
| 1:42 | 1:53 | 11 | | 2 | | 2 | 3 | |
| 1:57 | 2:08 | 11 | | 2 | | 4 | 5 | |
| 2:11 | 2:26 | 15 | | 4 | | 3 | 4 | |
| 2:30 | 2:40 | 10 | Apple In, Peach Out | 2 | 3 | 2 | 2 | |
| 2:48 | 3:01 | 13 | Orange Out | 6 | 5 | | 0 | |
| 3:02 | 3:17 | 15 | Pear Out | 5 | 4 | | | |
| 3:25 | 3:33 | 8 | Apple Out | | 2 | | | |
| | | | TOTAL Rets | 32 | 14 | 12 | 29 | |
| | | | TOTAL T | 105 | 46 | 56 | 107 | **SD** |
| | | | **Av. Re-attention Period** | **3.3** | **3.3** | **4.7** | **3.7** | 0.66080759 |

Figure B.450: Re-attention period statistics, Trial 4.

## B.2.2 Group Statistics

### B.2.2.1 Processing Script Output

Table B.4: Extracted parameters, synthetic trials.

| Param | S1 | S2 | S3 | S4 |
|---|---|---|---|---|
| $Spv_p$ | 1.298856 | 2.013923 | 1.741268 | 3.665578 |
| $Spv_np$ | 0.615771 | 0.990874 | 0.973078 | 2.773834 |
| $Scv_p$ | 31.553287 | 22.700747 | 25.823510 | 26.807705 |
| $Scv_np$ | 28.509815 | 25.715175 | 26.616577 | 28.999305 |
| $Scl$ | 10.056636 | 8.028029 | 8.329611 | 9.263824 |
| $Spl$ | 2.030133 | 1.961543 | 1.947704 | 3.150480 |
| $Scf$ | 33.582401 | 40.772499 | 39.065098 | 46.388102 |
| $Spf$ | 145.568399 | 230.463301 | 239.910502 | 180.944398 |
| $Sc\%$ | 18.745325 | 15.032123 | 14.003052 | 20.405398 |
| $Spt_p$ | 1.089046 | 1.003941 | 1.107530 | 1.197154 |
| $Sct_p$ | 0.300003 | 0.204587 | 0.194662 | 0.232634 |
| $Scl_p$ | 10.132018 | 5.284343 | 5.048391 | 7.072878 |
| $Spl_p$ | 2.263494 | 2.032493 | 2.157840 | 3.876390 |
| $Scf_p$ | 10.500100 | 9.820199 | 6.618500 | 8.142201 |
| $Spf_p$ | 50.096099 | 59.232501 | 52.053900 | 55.069099 |
| $Sc_p\%$ | 17.327985 | 14.221311 | 11.280432 | 12.880926 |
| $Spt_np$ | 1.164296 | 1.630770 | 1.707787 | 0.946431 |
| $Sct_np$ | 0.334526 | 0.332820 | 0.334501 | 0.316082 |
| $Scl_np$ | 10.018398 | 9.444125 | 1.9.479729 | 9.897569 |
| $Spl_np$ | 1.329795 | 1.510686 | 1.397552 | 2.378282 |
| $Scf_np$ | 23.082300 | 30.952300 | 32.446598 | 38.245901 |
| $Spf_np$ | 95.472300 | 171.230800 | 187.856602 | 125.875299 |
| $Sc_np\%$ | 19.469763 | 15.309044 | 14.728155 | 23.303450 |
| $Spt_r$ | 1.069097 | 1.624368 | 1.541978 | 0.790567 |
| $Spl_r$ | 0.587497 | 0.743268 | 0.647662 | 0.613530 |
| $Scl_r$ | 0.988786 | 1.787190 | 1.877772 | 1.399369 |
| $Spv_r$ | 0.474087 | 0.492012 | 0.558833 | 0.756725 |
| $Scv_r$ | 0.903545 | 1.132790 | 1.030711 | 1.081753 |
| $Scp_r$ | 1.123602 | 1.076486 | 1.305637 | 1.809144 |
| $Pr_{av}$ | 4.825 | 6.425 | 6 | 3.75 |
| $Pr_{sd}$ | 0.5057997 | 0.60207973 | 0.46904158 | 0.66080759 |
| $T$ | 155 | 218 | 227 | 203 |

#### B.2.2.2 Bootstrapping

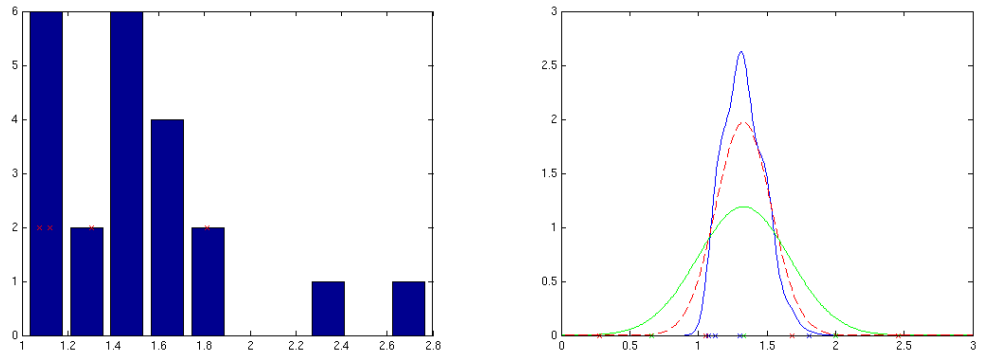



Figure B.451: $Sc_r$ parameter.

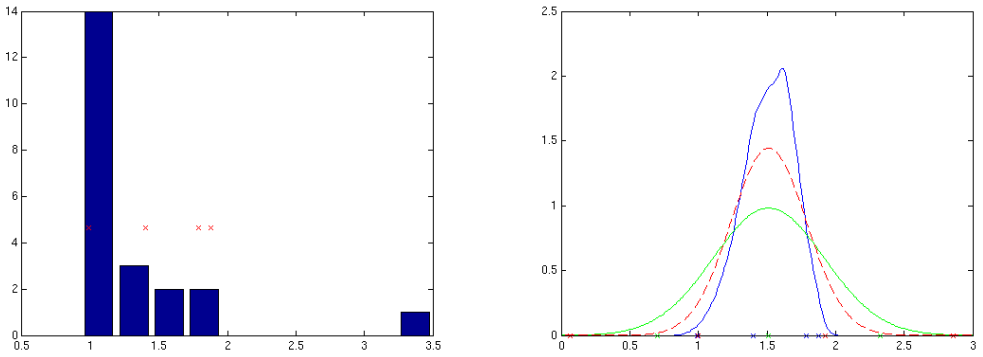



Figure B.452: $Scl_r$ parameter.

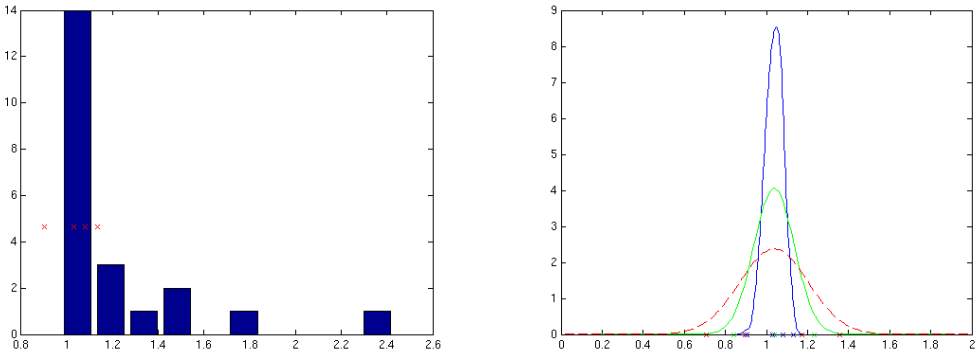


Figure B.453: $Scv_r$ parameter.

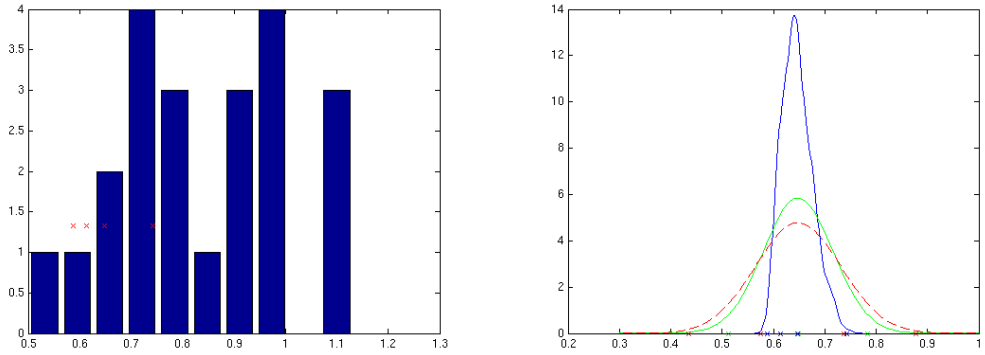


Figure B.454: $Spl_r$ parameter.

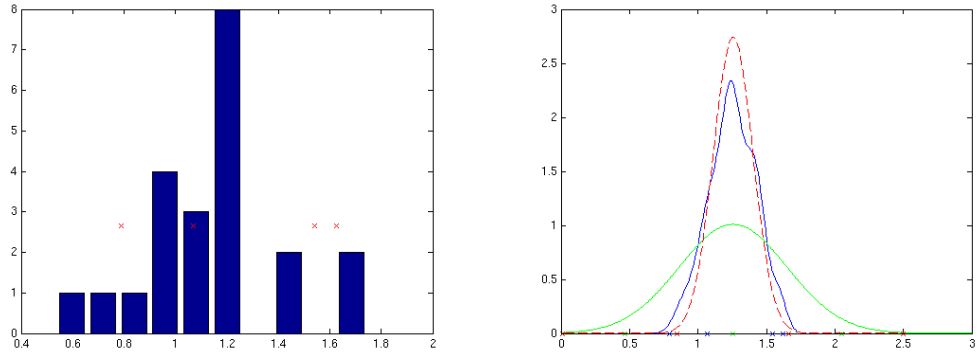


Figure B.455: $Spt_r$ parameter.

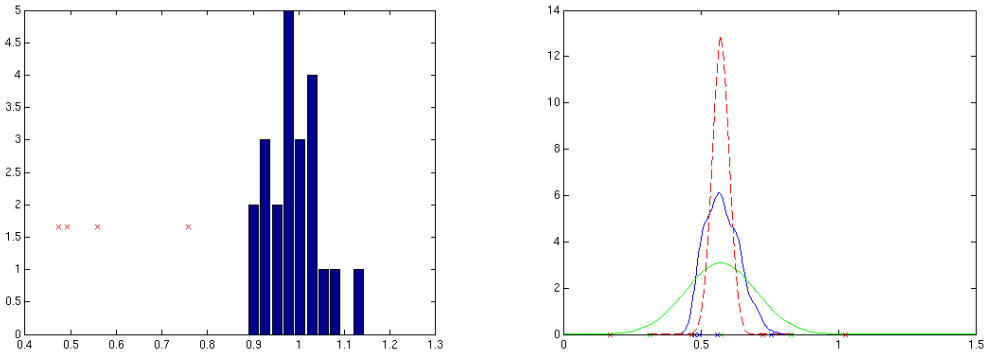



Figure B.456: $Spv_r$ parameter.

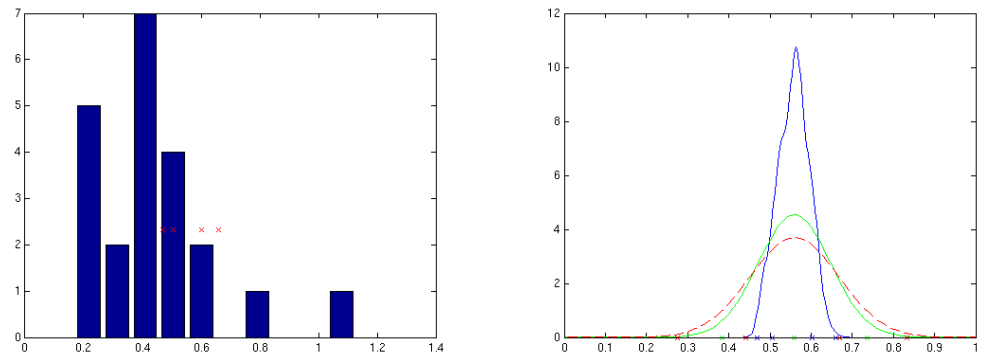



Figure B.457: $Spv_r$ parameter.

# Appendix C

# Demonstration Footage

## C.1 DVD Index

The following directory structure exists in Appendix C on the accompanying DVD:

Chapter 1 - Introduction (no movies)
Chapter 2 - Primate Vision System (no movies)
Chapter 3 - Synthetic Primate Vision System (no movies)
Chapter 4 - Active Vision Platform
Chapter 5 - Active Rectification
Chapter 6 - Spatial Perception
Chapter 7 - Coordinated Fixation
Chapter 8 - Active Attention
Chapter 9 - Human Trials
Chapter 10 - Synthetic Trials

The above directories are populated as follows:

### C.1.1 Chapter 4 - Active Vision Platform

CeDAR - 3dof head.avi
Image-based stabilisation.mpg
ATR 7dof head.mp4
Gyro-based stabilisation.mp4

### C.1.2 Chapter 5 - Active Rectification

Mosaic construction - manual movement.mpg
Mosaic construction - sinusoidal movement.avi
Mosaicing of saliency map.avi

### C.1.3 Chapter 6 - Spatial Perception

Occupancy grid.mpg
Flow.mpg
Object segmentation.mpg
Ground plane.mpg

### C.1.4 Chapter 7 - Coordinated Fixation

Early ZDF.avi
MRF ZDF theory.mp4
MRF ZDF demo.avi
Bimodal.mp4

### C.1.5 Chapter 8 - Active Attention

Static IOR.avi
Dynamic IOR.avi
Active saliency.avi
Fixation map.avi
Active attention fixation.avi
Active attention scene.avi

### C.1.6 Chapter 9 - Human Trials

P1.mp4
P2.mp4
T1.mp4
T2.mp4
T3.mp4
T4.mp4
T5.mp4
T6.mp4
T7.mp4
T8.mp4
T9.mp4
T10.mp4
T11.mp4
T12.mp4
T13.mp4
T14.mp4
T15.mp4
T16.mp4
T17.mp4
T18.mp4
T19.mp4
T20.mp4

### C.1.7 Chapter 10 - Synthetic Trials

S1.mp4
S2.mp4
S3.mp4
S4.mp4
S1proc.mp4
S2proc.mp4
S3proc.mp4
S4proc.mp4

# References


Allen, J.G., Xu, R.Y.D. & Jin, J.S. (2003). Object tracking using camshift algorithm and multiple quantized feature spaces. In *Conf. in Research and Practice in Inf. Tech.* 189

Allman, J., Miezin, F. & McGuinness, E. (1985). Stimulus specific responses from beyond the classical receptive field: neurophysiological mechanisms for local-global comparisions in visual neurons. In *Annu. Rev. Neurosci.*, 8:405–430. 48

Aloimonos, J., Weiss, I. & Bandyopadhyay, A. (1988). Active vision. In *IEEE International Journal on Computer Vision*, 333–356. 19, 21

Aloimonos, J., Weiss, I. & Bandyopadhyay, A. (1993). Active perception. In *Lawrence Erlbaum Associates*. 19

Azoz, Y., Devi, L. & Sharma, R. (1998). Tracking hand dynamics in unconstrained environments. In *IEEE International Conference on Face and Gesture Recognition Nara, Japan*. 165

Bajcsy, R. (1985). From active perception to active cooperation - fundamental processes of intelligent behavior. In *The Third IEEE Workshop on Computer Vision, pp. 55-59*. 19

Bajczy, R. (1988). Active perception. In *IEEE International Journal on Computer Vision*, 8:996–1005. 21

Ballard, D. (1991). Animate vision. In *Artificial Intelligence*, 48(1):57–86. 10, 19, 21

Baloh, R.W., Yee, R.D. & Honrubia, V. (1981). Eye movements in patients with wallenberg's syndrome. In *Ann NY Acad Sci*. 95

Banks, J. & Corke, P. (1991). Quantitative evaluation of matching methods and validity measures for stereo vision. *IEEE International Journal of Robotics Research*, **20**. 126, 151, 200

Bellingham, J., Wilkie, S.E., Morris, A.G., Bowmaker, J.K. & Hunt, D.M. (1997). Characterisation of the ultraviolet-sensitive opsin gene in the honey bee, apis mellifera. In *European Journal of Biochemistry, Vol 243, 775-781*. 16

Blake, A. & Isard, M. (1998). Active contours. In *Springer-Verlag*. 165

Blake, A. & Yuille, A. (1992). Active vision. In *MIT Press*. 19

Blaser, E., Sperling, G. & Lu, Z.L. (1999). Measuring the amplification of attention. In *National Acad Sciences 96, 20:11681-11686*. 57

Born, R. & Bradley, D. (2005). Structure and function of visual area mt. In *Annu Rev Neurosci 28: 157-8*. 47

Boykov, Y., Veksler, O. & Zabih, R. (1997). Markov random fields with efficient approximations. Tech. Rep. TR97-1658, Computer Science Department, Cornell University Ithaca, NY 14853. 174, 175

Braddick, O.J. & O'Brian, J.M.D. (2001). Brain areas sensitive to visual motion. In *Perception 30: 61-72*. 46

Bradski, G.R. (1998). Computer vision face tracking for use in a perceptual user interface. In *Intel. Tech Journ.*. 166

Braun, J. (1993). Shape-from-shading is independent of visual attention and may be a texon. In *Spat. Vis.*, 7:311–322. 48

Braun, J. & Julesz, B. (1998). Withdrawing attention at little or no cost: detection and discrimination of tasks. In *Percept. Psychophys.*, 60:1–23. 56

Brooks, A., Dickins, G., Zelinsky, A., Kieffer, L. & Abdallah, S. (1997). A high-performance camera platform for real-time active vision. In *Int. Conf. Field and Service Robotics*. 83, 84

Brooks, A., Zelinsky, A. & Kieffer, L. (1998). A multimodal approach to real-time active vision. In *Int. Conf. Intelligent Robotics*. 84

Burt, P. (1988). Smart sensing within a pyramid vision machine. In *Proceedings of the IEEE vol. 76, pp. 1007-1015*. 19

Buswell, G.T. (1922). Fundamental reading habits: A study of their development. In *Chicago, IL: University of Chicago Press*. 229

Buswell, G.T. (1935). How people look at pictures. In *Chicago, IL: University of Chicago Press*. 229

Buswell, G.T. (1937). How adults read. In *Chicago, IL: University of Chicago Press*. 229

Campbell, F. & Green, D.G. (1965). Optical and retinal factors affecting visual resolution. In *J. Physiol. 181:576-593*. 39

Carandini, M. & Heeger, D. (1994). Summation and division by neurons in primate visual cortex. In *Science*, 264:1333–1336. 56

Carrasco, M., Penpeci-Talgar, C. & Eckstein, M. (2000). Spatial convert attention increases contrast sensitivity across the csf: support for signal enhancement. In *Vision Res.*, 40:1203–1215. 56, 202

Cave, K.R. (1999). The featuregate model of visual selection. In *Psychological Research*. 67, 68

Cham, T. & Rehg, J. (1999). Dynamic feature ordering for efficient registration. In *IEEE International Conference on Computer Vision, volume 2, Corfu, Greece*. 165, 167

Chandon, P., Young, S.H. & Hutchinson, J.W. (2001). Measuring the value of point-of-purchase marketing with commercial eye-tracking data. In *J. Marketing Research*. 229

Cheng, Y. (1995). Mean shift, mode seeking, and clustering. In *IEEE Trans. Pattern and Machine Intelligence*, 17:790–799. 166

Clark, R.N. (2005). Notes on the resolution and other details of the human eye. In *http://clarkvision.com/imagedetail/eye-resolution.html*. 39

Comaniciu, D., Ramesh, V. & Meer, P. (2003). Kernel-based object tracking. In *IEEE Trans. Patt Anal. and Mach. Int.*, 25:5:564–575. 189

Coombs, D. & Brown, C. (1992). Real-time smooth pursuit tracking for a moving binocular robot. In *Proc., International Conference on Computer Vision and Pattern Recognition 23-28*. 168

Crick, F. (1998). Consciousness and neuroscience. In *Cerebral Cortex, 8:97-107*. 54

Cronin, T.W. & King, C.A. (1989). Spectral sensitivity of vision in the mantis shrimp, gonodactylus oerstedii, determined using noninvasive optical techniques. In *Biol. Bull. 176: 308-316*. 16

Curcio, C.A., Sloan, K.R. & Kalina, R.E. (1990). Human photoreceptor topography. In *J Comp Neurol. 292: 497-523*. 39

D Kahneman, A.T. (1984). Changing views of attention and automaticity. In *In R. Parasuraman and D.R. Davies (Eds.), Varieties of Attention, (pp. 29-61). New York: Academic Press*. 57

Dacey, M. (1996). Circuitry for color coding in the primate retina. In *Proc. Nat. Acad. Sci.*, 93:582–588. 48, 217

Dankers, A. (2002). Multiple cue horopter tracking with cedar. In *Honours Thesis, Aust Nat. Univ*. 169

DeSchepper, B. & Treisman, A. (1996). Visual memory for novel shapes: implicit coding without attention. In *Exp Psychol Learn Mem Cogn*. 54

Desimone, R. & Duncan, J. (1995). Neural mechanisms of selective visual attention. In *Annu. Rev. Neurosci. 18:193-222*. 54, 55, 57

Djeraba, C. (2006). Eye gaze tracking in web, image and video documents. In *ACM Multimedia*. 229

Dorris, M.C., Pare, M. & Munoz, D.P. (1997). Neuronal activity in monkey superior colliculus related to the initiation of saccadic eye movements. In *J. Neuroscience*. 216

Efron, B. & Tibshirani, R.J. (1993). An introduction to the bootstrap. In *Chapman and Hall*. 263

Elfes, A. (1989). Using occupancy grids for mobile robot perception and navigation. *IEEE Computer Magazine*, 46–57. 135, 137, 138, 208

Fischer, R.F. & Tadic, B. (2000). Optical system design. In *McGraw-Hill Professional*. 39

Fletcher, L., Loy, G., Barnes, N. & Zelinsky, A. (2005). Correlating driver gaze with the road scene for driver assistance systems. In *Robotics and Autonomous Systems*. 230

Ford, L. & Fulkerson, D. (1962). *Flows in Networks*. Princeton University Press. 176

Fukunaga, K. (1990). Introduction to statistical pattern recognition, 2nd edition. In *Academic Press*. 166

Fusiello, A., Trucco, E. & Verri, A. (2000). A compact algorithm for rectification of stereo pairs. *Machine Vision and Applications*, **12**, 16–22. 114, 115

Gavrila, D. & Davis, L. (1996). 3-d model-based tracking of human motion in action. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 165, 167

Gavrila, D.M. (1999). The visual analysis of human movement - a survey. In *Computer Vision and Image Understanding*. 165

Geman, S. & Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 721–741. 173

Gilbert, C., ito, M., Kapadia, M. & Westheimer, G. (2000). Interactions between attention, context and learning in primary visual cortex. In *Vision Res.*, 40:1217–1226. 49, 70

Gluck, M.A., Reifsnider, E.S. & Thompson, R.F. (1990). Adaptive signal processing and the cerebellum: Models of classical conditioning and vor adaptation. In *Neuroscience and connectionist theory*. 95

Gosselin, C., St-Pierre, E. & Gagne, M. (1996). On the development of the agile eye. In *IEEE Robotics and Automation Magazine 29–37*. 82

Gottlieb, J.P., Kusunoki, M. & Goldberg, M.E. (1998). The representation of visual salience in monkey parietal cortex. In *Nature 391(6666):481-4*. 55

Halder, G., Callaerts, P. & Gehring, W.J. (1995). New perspectives on eye evolution. In *Curr. Opin. Genet. Dev. 5 (pp. 602 609)*. 15

Harris, C. & Stephens, M. (1988). A combined corner and edge detector. In *4th Alvey Vision Conf.*, 147–151. 111

Hartley, R. & Zisserman, A. (2004). *Multiple View Geometry in Computer Vision, Second Edition*. Cambridge University Press. 106

Henderson, J.M. (1992). Visual attention and eye movement control during reading and picture viewing. In *Eye movements and visual cognition*. 69

Henkel, R.D. (1999). Stereovision by coherence detection. 133, 134

Holst, G. (1998). Ccd arrays, cameras, and displays. In *2nd ed. Winter Park, FL : JCD Publishing*. 9

Horowitz, T.S. & Wolfe, J.M. (1998). Visual search has no memory. In *Nature*. 56, 297

Imagawa, K., Lu, S., & Igi, S. (1998). Color-based hands tracking system for sign language recognition. In *IEEE International Conference on Face and Gesture Recognition, Nara, Japan*. 165, 167

Irwin, D.E. & Zelinsky, G.J. (2002). Eye movements and scene perception: Memory for things observed. In *Perception and Psychophysics*. 57, 297

Isard, M. & Blake, A. (1998). Condensation: conditional density propatation for visual tracking. In *Int. Journal of Comp. Vis.*, 29:1:5–28. 166

Itti, L. (2005). Models of bottom-up attention and saliency. In *Neurobiology of Attention*. 55, 67

Itti, L. & Koch, C. (1998). A model for saliency-based visual attention for rapid scene analysis. In *IEEE Trans. Patt. Anal. Mach. Int.*, 20:1254–1259. 58, 67, 196

Itti, L. & Koch, C. (2000). Feature combination strategies for saliency-based visual attention systems. In *J. Electronic Imaging*. 56, 197, 203, 205

Itti, L. & Koch, C. (2001). Computational modelling of visual attention. In *Nature Neuroscience 2:194-203*. 48, 54, 197

Jennings, C. (1999). Robust finger tracking with multiple cameras. In *Proceedings of the International Workshop on Recognition, Analysis and Tracking of Faces and Gestures in Real-Time Systems, Corfu, Greece*. 165

J.Joseph & LaViola, J. (2003). A comparision of unscented and extended kalman filtering for estimation quaternion motion. In *Proceedings of American Control Conference*. 167

Jojic, N., Turk, M. & Huang, T. (1999). Tracking self-occluding articulated objects in dense disparity maps. In *IEEE International Conference on Computer Vision, volume 1, Corfu, Greece*. 165, 167

Kagami, S., Okada, K., Inaba, M. & Inoue, H. (2000). Realtime 3d depth flow generation and its application to track to walking human being. In *IEEE International Conf. on Robotics and Automation*, 4:197–200. 65, 153, 202

Kandel, E.R., Schwartz, J.H. & Jessell, T.M. (2000). Principles of neural science, 4th edition. In *McGraw-Hill Medical*. 38, 39, 43, 50, 51, 95, 104, 131

Kenner, N. & Wolfe, J. (2003). An exact picture of your target guides visual search better than any other representation. In *J Vision*. 58

Klein, R. (1980). Does oculomotor readiness mediate cognitive control of visual attention? In *Attention and performance VIII (259276). R S Nickerson (Ed.), Hillsdale, NJ: Erlbaum*. 69

Klein, R.M. (2000). Inhibition of return. In *Trends in Cognitive Sciences*. 56

Knuth, D.E. (1997). Seminumerical algorithms: The art of computer programming. In *Addison-Wesley*. 289

Koch, C. & Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. In *Hum Neurobiol 4:219-27*. 66, 68

Kolmogorov, V. & Zabih, R. (2002a). Multi-camera scene reconstruction via graph cuts. In *Europuan Conference on Comupter Vision*, 82–96. 176

Kolmogorov, V. & Zabih, R. (2002b). What energy functions can be minimized via graph cuts? In *Europuan Conf. on Comupter Vision*, 65–81. 176

Kovesi, P. (2003). Phase congruency detects corners and edges. In *Aust. Patt. Rec. Soc.*, 309–318. 224

Kuniyoshi, Y., Kita, N., Rougeaux, S. & Suchiro, T. (1995). Active stereo vision system with foveated wide angle lenses. In *Asian Conf. Comp Vis*. 82

Kustov, A.A. & Robinson, D.L. (1996). Shared neural control of attentional shifts and eye movements. In *Nature 384:74-77*. 55

L Matthies, A.E. (1988). Integration of sonar and stereo range data using a grid-basedrepresentation. In *Robotics and Automation*. 136

Labayrade, R., Aubert, D. & Tarel, J. (2002). Real time obstacle detection on non flat road geometry through 'v-disparity' representation. In *IEEE Intelligent Vehicle Symposium*, vol. 2, 646–651. 156

Lamme, V.A.F., Supr, H., Landman, R., Roelfsema, P.R. & Spekreijse, H. (2000). The role of primary visual cortex (v1) in visual awareness. In *Vision Research, 40:1507-1521*. 45, 130

Land, M.F. & Nilsson, D. (2002). Animal eyes. In *Oxford Animal Biology series, Oxford University Press*. 15, 16

Lappe, M. (2006). Psychophysical investigations of gaze and coi during simulated driving in humans. In *EcoVision Rep*. 229

Lee, D., Itti, L., Koch, C. & Braun, J. (1999). Attention activates winner-take-all competition among visual filters. In *Nature Neurosci.*, 2: 375–381. 56

Loy, G., Fletcher, L., Apostoloff, N. & Zelinsky, A. (2002). An adaptive fusion architecture for target tracking. In *Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, 261. 165

MacCormick, J. & Blake, A. (1999). A probabilistic exclusion principle for tracking multiple objects. In *IEEE International Conference on Computer Vision, volume 1, Corfu, Greece*. 167

Martin, J., Devin, V. & Crowley, J. (1998). Active hand tracking. In *IEEE International Conference on Face and Gesture Recognition, Nara, Japan*. 165, 167

Merriam, E., Genovese, C. & Colby, C. (2003). Spatial updating in human parietal cortex. In *Neuron*, 39:351–373. 24, 56, 67, 104, 196, 207

Metaxas, D. (1999). Deformable model and hmm-based tracking, analysis and recognition of gestures and faces. In *Proceedings of the International Workshop on Recognition, Analysis and Tracking of Faces and Gestures in Real-Time Systems, pages 136-140, Corfu, Greece*. 165, 167

Milanese, R., Wechsler, H., Gill, S. & J M Bost, T.P. (1994). Integration of bottom-up and top-down cues for visual attentionusing non-linear relaxation. In *Computer Vision and Pattern Recognition*. 67

Miller, E.K. (2000). The prefrontal cortex and cognitive control. In *Nat Rev Neurosci 1:59-65*. 54

Mitchell, T.M. (1997). Machine learning. In *McGraw-Hill*. 263, 289

Moran & Desimone (1985). Selective attention gates visual processing in the extrastriate cortex. In *Science 229(4715)*. 47, 50

Moravec, H. (1989). Sensor fusion in certainty grids for mobile robots. In *Al Magazine, Carnegie Mellon University Pennsylvania*, 9:61–74. 136, 137, 138

Moravec, H. (1996). Robot spatial perception by stereoscopic vision and 3d evidence grids. Tech. Rep. CMU-RI-TR-96-34, Carnegie Melon Univ. 135

Murray, D. & Little, J.J. (2000). Using real-time stereo vision for mobile robot navigation. In *Autonomous Robots*. 136

Murray, W.W., Du, F., McLauchlan, P.F., Reid, I.D., Sharkey, P.M. & Brady, M. (1992). Active vision. In *Active Vision 155–172*. 84, 90

Murthy, A., Thompson, K.G. & Schall, J.D. (2001). Dynamic dissociation of visual selection from saccade programming in frontal eye field. In *Am Physiological Soc*. 55

Najemnik, J. & Geisler, W.S. (2004). Optimal eye movement strategies in visual search. In *Nature*. 58

Nakayama & Mackeben (1989). Sustained and transient components of focal visual attention. In *Vision Research 29:11, pp1631-1647*. 23

Navalpakkam, V. & Itti, L. (2004). Modeling the influence of task on attention. In *Vision Research*. 68

Navalpakkam, V., Arbib, M. & Itti, L. (2005). Attention and scene understanding. In *Neurobiology of Attention*. 55, 208

Neri, P., Bridge, H. & Heege, D.J. (2004). Stereoscopic processing of absolute and relative disparity in human visual cortex. In *Neurophysiol. 92: 18801891*. 51, 62, 104, 130

Nieuwenhuis, S. & Yeung, N. (2005). Neural mechanisms of attention and control: losing our inhibitions? In *Nature*, 8:1631–1633. 57, 68, 210, 214

Nishida, S., Ledgeway, T. & Edwards, M. (2001). Dual multiple-scale processing for motion in the human visual system. In *Vision Research 37: 2685-2698*. 50

Nolte, J. (2002). The human brain: An introduction to its functional anatomy. In *5th Ed. St. Louis: Mosby, 410-447*. 44

Nothdurf, H. (1990). Texture discrimination by cells in the cat lateral geniculate nucleus. In *Exp. Brain Res*, 82:48–66. 48, 197

Noton, D. & Stark, L. (1971). Scanpaths in saccadic eye movements while viewing and recognizing patterns. In *Vision Res*. 66, 69

Nummiaro, K., Koller-Meier, E. & Gool, L.V. (2002). A colour-based particle filter. In *Proc. Int. Workshop on Generative Model Based Vis. in conj. ECCV*, 53–60. 166

Ohzawa, I., DeAngelis, G.C. & Freeman, R.D. (1997). Encoding of binocular disparity by simple cells in the cat's visual cortex. In *Journal of Neurophysiology*. 130

Oliva, A. (2005). Gist of a scene. In *Neurobiology of Attention*. 58, 69

Ong, E. & Gong, S. (1999). A dynamic human model using hybrid 2d-3d representations in hierarchical pca space. In *British Machine Vision Conference, volume 1, pages 33-42, Nottingham, UK, BMVA*. 165, 167

Oshiro, N., Maru, N., Nishikawa, A. & Miyazaki, F. (1996). Binocular tracking using log polar mapping. In *IROS*, 791–798. 168, 190

Pahlavan, K. & Eklundh, J.O. (1992). A head-eye system - analysis and design. In *Image Understanding, CVGIP*. 82

Pahlavan, K., Uhlin, T. & Eklundh, J.O. (1993). Dynamic fixation. In *In 4th International Conference on Computer Vision (ICCV), pp. 412-419*. 19

Panerai, F., Metta, G. & Sandini, G. (2000). Visuo-inertial stabilization in space-variant binocular systems. In *Robotics and Autonomous Systems*. 94

Parker, A. (2003). In the blink of an eye; how vision sparked the big bang of evolution. In *Perseus, Basic Books*. 15, 16, 17

Pasupathy, A. & Connor, C. (1999). Responses to contour features in macaque area v4. In *Journal of Neurophysiology*, 82:2490–2502. 48

Poggio, G.F., Gonzalez, F. & Krause, F. (1988). Stereoscopic mechanisms in monkey visual cortex: Binocular correlation and disparity selectivity. In *J. Neuroscience 8(12): 4531 -I550*. 51, 130

Qiu, F.T. & von der Heydt, R. (2005). Figure and ground in the visual cortex: V2 combines stereoscopic cues with gestalt rules. In *Neuron 47: 155-166*. 46

Rae, R. & Ritter, H. (1998). 3d real-time tracking of points of interest based on zero-disparity filtering. In *Workshop Dynamische Perzeption, Proceedings in Artificial Intelligence*, 105–111. 168, 190

Rao, R.P.N., Zelinsky, G.J., Hayhoe, M.M. & Ballard, D.H. (1997). Eye movements in visual cognition: A computational study. In *Technical Report*. 68

Rasmussen, C. & Hager, G. (1998). Joint probabilistic techniques for tracking multi-part objects. In *IEEE Conference on Computer Vision and Pattern Recognition, pages 16-21, Santa Barbara, CA*. 167

Rensink, R.A. (2000). The dynamic representation of scenes. In *Visual Cognition*. 66

Reynolds, J.H., Pasternak, T. & Desimone, R. (2000). Attention increases sensitivity of v4 neurons. In *Neuron, 26 3:703-714*. 54, 70, 188, 225

Rizzolatti, G., Riggio, L., I, I.D. & Umilt, C. (1987). Reorienting attention across the horizontal and vertical meridians: evidence in favor of a premotor theory of attention. In *Neuropsychologia 25(1A):31-4*. 69

Rodieck, R.W. (1998). First steps in seeing. In *Sinuaer Associates*. 14, 17, 33, 38, 50, 88, 95, 128

Rougeaux, S. (1999). Real-time active vision for versatile interaction. In *PhD Thesis, University Evry, France*. 10

Rougeaux, S. & Kuniyoshi, Y. (1997a). Robust real-time tracking on an active vision head. In *IEEE International Conference on Intelligent Robots and Systems*, 873–879. 168

Rougeaux, S. & Kuniyoshi, Y. (1997b). Velocity and disparity cues for robust real-time binocular tracking. In *IEEE International CVPR*. 168, 190, 224

Rougeaux, S., Kita, N., Kuniyoshi, Y., Sakano, S. & Chavand, F. (1994). Binocular tracking based on virtual horopters. In *IROS*. 168, 190

Rullen, R.V. (2003). Visual saliency and spike timing in the ventral visual pathway. In *Journal of Physiology*. 58

Rybak, I.A., Gusakova, V.I., Golovan, A.V., Podladchikova, L.N. & Shevtsova, N.A. (1998). A model of attention-guided visual perception and recognition. In *Vision Research 38:23872400*. 54, 69

Schwartz, E. (1980). A quantitative model of the functional architecture of human striate cortex with application to visual illusion and cortical texture analysis. In *Biological Cybernetics*, 37(2):63–76. 21

Se, S., Lowe, D. & Little, J. (2001). Vision-based mobile robot localization and mapping using scale-invariant features. In *IEEE International Conf. on Intelligent Robots and Systems*, 2051–2058. 111

Sharkey, P.M., Murray, D.W. & Heuring, J.J. (1997). On the kinematics of robot heads. In *IEEE Trans. Robotics and Automation 437-442*. 82

Shen, C., Brooks, M.J. & Hengel, A. (2005). Fast global kernel density mode seeking with application to localisation and tracking. In *IEEE Int. Conf. on Comp. Vis*. 166, 189

Sherman, S.M. (2006). The role of the thalamus in cortical function: not just a simple relay. In *Journal of Vision*. 44

Sherwin, F. & Armitage, M. (2003). Trilobites the eyes have it! In *CRSQCreation Research Society Quarterly, Vol. 40, No. 3*. 15

Siagian, C. & Itti, L. (2007). Rapid biologically-inspired scene classification using features shared with visual attention. In *IEEE Trans. Patt Anal and Mach. Intel*. 69

Sillito, A., Grieve, K., Jones, H., Cudeiro, J. & Davis, J. (1995). Visual cortical mechanisms detecting focal orientation discontinuities. In *Nature*, 378:492–496. 49

Suder, K. & Worgotter, F. (2003). The control of low-level information flow in the visual system. In *Rev Neurosci. 11(2-3):127-46*. 55

Sugrue, L., Corrado, G.S. & Newsome, W.T. (2005). Choosing the greater of two goods: Neural currencies for valuation and decision making. In *Neuroscience*. 23, 216

Sun & Bonds (1994). Two-dimensional receptive field organization in striate cortical neurons of the cat. In *Vis Neurosci.*, 11: 703–720. 203

Sutherland, O., Rougeaux, S., Abdallah, S. & Zelinsky, A. (2000). Tracking with hybrid-drive active vision. In *ISER*. 81, 91

Tian, T. & Shah, M. (1997). Recovering 3d motion of multiple objects using adaptive hough transform. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **19**, 1178–1183. 156

Torralba, A. (2005). Contextual influences on saliency. In *Neurobiology of Attention*. 68

Toyama, K. & Horvitz, E. (2000). Bayesian modality fusion: Probabilistic integration of multiple vision algorithms for head tracking. In *Asian Conference on Computer Vision, Tapei, Taiwan*. 165

Treue, S. & Maunsell, J. (1996). Attentional modulation of visual motion processing in cortical areas mt and mst. In *Nature*, 382:539–541. 48

Trieisman, A. & Gelade, G. (1980). A feature-integration theory of attention. In *Cogn. Psychol.*, 12:97–136. 57, 58, 70, 188, 225

Triesch, J. & von der Malsburg, C. (2000). Self-organized integration of adaptive visual cues for face tracking. In *IEEE International Conference on Face and Gesture Recognition Grenoble, France*. 165

Troung, H. (1998). Active vision head. In *Thesis, Australian Nat Univ.*. 81, 85, 89

Truong, H., Abdallah, S., Rougeaux, S. & Zelinsky, A. (2000). A novel mechanism for stereo active vision. In *Australian Conf. on Robotics and Automation*. 83, 88

Tsotsos, J.K. (1989). The complexity of perceptual search tasks. In *International Joint Conference on Artificial Intelligence (IJCAI)*. 22

Tsotsos, J.K. (1992). Active verses passive perception, which is more efficient? In *International Journal of Computer Vision (IJCV)*. 21

Tsotsos, J.K., Culhane, S., Wai, W., Lai, Y., Davis, N. & Nuflo, F. (1995). Modeling visual attention via selective tuning. In *Artificial Intelligence 78(1-2),p 507 - 547,*. 22, 67

Ude, A., Wyart, V., Lin, M. & Cheng, G. (2005). Distributed visual attention on a humanoid robot. In *Report*. 25, 82, 197

Ullman, S. (1984). Visual routines. In *Cognition, 18:97-159*. 19

Vidyasagar, T.R. (1999). A neuronal model of attentional spotlight: parietal guiding the temporal. In *Brain Res Rev*. 53

Vinje, W.E. & Gallant, J.L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. In *Science 287:12731276*. 70

Wandell, B.A. (1995). *Foundations of vision*. Sunderland. 21, 164

Weichselgartner, E. & Sperling, G. (1987). Attention increases sensitivity of v4 neurons. In *Science 238 4828:778-780*. 54, 57, 70

Wilson, F.A., Scalaidhe, S.P. & Goldman-Rakic, P.S. (1993). Dissociation of object and spatial processing domains in primate prefrontal cortex. In *Science 260:1955-1958*. 57

Wilson, H.R. & Cowan, J.D. (1972). Excitatory and inhibitory interactions in localized populations of model neurons. In *Journal of Biophys 1-24*. 84

Wolfe, J. (1996). Attention. 58, 68

Wren, C., Clarkson, B. & Pentland, A. (2000). Understanding purposeful human motion. In *IEEE International Conference on Face and Gesture Recognition, pages 378-383, Grenoble, France*. 165, 167

Wylie, D., Bischof, W. & Frost, B. (1998). Common reference frame for neural coding of translational and rotational optic flow. In *Nature*, 392:278–282. 203

Yarbus, A.L. (1967). Eye movements and vision. In *Thesis, New York*. 229

Yu, H. & Baozong, Y. (1996). Zero disparity filter based on wavelet representation in the active vision system. In *Proc. Int. Conf. Signal Proc.*, 279–282. 168, 190

Zajac, J.L. (1960). Convergence, accommodation, and visual angle as factors in perception of size and distance. In *American Journal of Psychology, Vol. 73, No. 1*. 52, 64, 130

Zetzsche, C. (1998). Investigation of a sensorimotor system for saccadic scene analysis: an integrated approach. In *5th Intl. Conf. Sim. Adaptive Behav.*, 5:120–126. 54, 202

Zhaoping, L. (2005). The primary visual cortex creates a bottom-up saliency map. In *Neurobiology of Attention 93:570-575*. 55