Facial Expression Recognition In The Wild: From Individual To Group

Abhinav Dhall

A thesis submitted for the degree of
Doctor of Philosophy
of the Australian National University

19 June, 2014

THE AUSTRALIAN NATIONAL UNIVERSITY

Research School of Computer Science
College of Engineering and Computer Science
The Australian National University
Canberra, ACT, Australia

This thesis is dedicated to my family.

Declaration

This thesis is an account of the doctoral research undertaken at the Research School of Computer Science, College of Engineering and Computer Science, The Australian National University, Canberra, ACT, Australia. This research was mainly supported by the Australian Leadership Award Scholarship (2010-2013) given by the the Australian Government's overseas aid program (AusAID). Further, the research was also supported by the Dean's travel award.

The results and analysis presented in this thesis are my own original work, accomplished under the supervision of Associate Professor Dr. Roland Goecke, Professor Dr. Tom Gedeon and Associate Professor Dr. Simon Lucey, except where otherwise acknowledged. This thesis has not been submitted for any other degree.

Abhinav Dhall
Research School of Computer Science
College of Engineering and Computer Science
The Australian National University
Canberra, Australia
19 June, 2014

Acknowledgements

I am grateful to my supervisory panel: Prof. Roland Goecke, Prof. Tom Gedeon and Prof. Simon Lucey for their guidance through out the duration of my PhD journey. A special thanks to Roland Goecke for his patience, advises and positive attitude. Even when things were gloomy at the research front, his motivation has helped me in overcoming the ups and downs of this PhD journey. My gratitude to the Australian Government's Australian Aid for International Development agency for supporting my PhD program. Thanks to the great team of AusAid at the ANU: Gina, Molly and Shian, for helping at each and every step. A special thanks to the administration team at CECS: Jadon Radcliffe, Diane Wellach-Smith, Elspeth Davies and Bindi Mamouney for making all the conference and international university visits possible.

A big thanks to Dr. Gwen Littlewort and Prof. Marian Bartlett at the University of California San Diego for hosting me at their lab during my visit to San Diego. The time in San Diego would have been unproductive and boring, if Karan Sikka was not there! Long discussions during the mid-night at the lab have evolved into great friendships and resulted into interesting publications! A sincere thanks to Dr. Stefanous Zaferious and Prof. Maja Pantic for hosting me at the Imperial College London. It was a great learning experience to work in a large prestigious research group. Thanks to Prof. Michel Wagner at the University of Canberra, for his advices and suggestions through out the PhD.

There are no words to explain, how much support and love my wife Jyoti has given me during this period. Specially, handling my grumpy mood when things did not seem so bright. A big thanks to my father Er. Tarsem Lal Dhall, who is my role model and the biggest motivation in my life. His positive attitude towards life and zeal to bounce back whenever there is a setback in life is amazing. My heart filed thanks to my mother Mrs. Reena Dhall for her unconditional love and blessings. This thesis would have been impossible without the support from my brother Er. Abhishek Dhall. With him handling the family affairs back in India, it was easier for me to concentrate on my work here in Australia. Thanks to my second family, Mr. V.K. Joshi, Mrs. Kiran Joshi and Deepak for their blessings and love.

"It is good to have an end to journey toward, but it is the journey that matters in the end." – Ursula K. Le Guin

This journey was made special due to friends at ANU and Canberra. A big thanks to fellow researchers at CECS: Akshay Asthana, Cong Phuoc Huynh, Vivek Kumar, Florian Poppa and Hajar Sadeghi for listening to me patiently, while discussing the PhD troubles. Thanks to my team mates and friends at the University of Canberra: Shyam Rajagopalan, Sharifa Algohemein, Ramana Murthy, Ibrahim Radwan, Laura Fernandez and David Vandyke, for making my time memorable at UC.

Chicken curry played an important role during this doctoral journey, thanks to Johnny Valbuena, Babu Dallakoti, Tushar, Munira, Alex Solntsev, Olga Solntsev, Arjuna Mohotolla and Erandi Mohotolla. Thanks to Tulika Saxena, Sachin Marwaha, Amit Bajaj, Kavita Bajaj, Neety Kaur, Ashish Dubash, Divya Das and Aman Singh for awesome hangouts, discussions and food.

Abstract

The progress in computing technology has increased the demand for smart systems capable of understanding human affect and emotional manifestations. One of the crucial factors in designing systems equipped with such intelligence is to have accurate automatic Facial Expression Recognition (FER) methods. In computer vision, automatic facial expression analysis is an active field of research for over two decades now. However, there are still a lot of questions unanswered. The research presented in this thesis attempts to address some of the key issues of FER in challenging conditions mentioned as follows: 1) creating a facial expressions database representing real-world conditions; 2) devising Head Pose Normalisation (HPN) methods which are independent of facial parts location; 3) creating automatic methods for the analysis of mood of group of people.

The central hypothesis of the thesis is that extracting close to real-world data from movies and performing facial expression analysis on movies is a stepping stone in the direction of moving the analysis of faces towards real-world, unconstrained condition. A temporal facial expressions database, *Acted Facial Expressions in the Wild (AFEW)* is proposed. The database is constructed and labelled using a semi-automatic process based on closed caption subtitle based keyword search. The use of movies facilitates the creation of rich meta-data. Currently, AFEW is the largest facial expressions database representing challenging conditions available to the research community. For providing a common platform to researchers in order to evaluate and extend their state-of-the-art FER methods, the first *Emotion Recognition in the Wild (EmotiW)* challenge based on AFEW is proposed. An image-only based facial expressions database: *Static Facial Expressions In The Wild (SFEW)* extracted from AFEW is proposed.

Furthermore, the thesis focuses on Head Pose Normalisation (HPN) for real-world images. Earlier methods were based on fiducial points. However, as fiducial points detection is an open problem for real-world images, HPN can be erroneous. A HPN method based on pose normalisation of response maps generated from part detectors is proposed. A shape prior is applied on the pose normalised response maps, guarantying a facial kinematic shape constraint. The proposed method does not require fiducial points and head pose information, which makes it suitable for real-world images.

Data from movies and the internet, representing real-world conditions poses another major challenge of the presence of multiple subjects to the research community. This defines another focus of this thesis where a novel approach for modeling the perception of mood of a group of people in an image is presented. A new database: HAppy People Images (HAPPEI) is constructed from Flickr based on keywords related to social events such as *marriage*, *convocation* and *party* etc. The images are labelled for each person's happiness intensity, face clarity and pose. Three models are proposed: *Group Expression Model (GEM)*, *Weighted Group Expression Model (GEM_w)* and *Augmented Group Expression Model* (GEM_{LDA}). GEM_w is based on social contextual attributes, which are used as weights on each person's contribution towards the overall group's mood. Further, feature augmentation is proposed by fusing a Bag-of-Words (BoW) based histogram with social contextual features and a topic model (GEM_{LDA}) is trained. The proposed framework is applied to applications of *group candid shot selection* and *event summarisation* and has shown promising outcomes.

The application of Structural SIMilarity (SSIM) index metric is explored for finding similar facial expressions. The proposed framework is applied to the problem of creating image albums based on facial expressions. Further, the framework is applied for finding similar facial expression pairs for training facial performance transfer algorithms.

List of Publications

Publications by the Candidate Relevant to the Thesis

All publications are available from the CD included with this thesis or from the candidate's website as a single file:
http://users.cecs.anu.edu.au/~adhall/PhdPublications.zip

Peer-Reviewed Publications

1. **Abhinav Dhall**, Karan Sikka, Gwen Littlewort, Roland Goecke and Marian Bartlett, A Discriminative Parts Based Model Approach for Fiducial Points Free and Shape Constrained Head Pose Normalisation In The Wild, In Proceedings of the 2014 IEEE Winter Conference on Applications of Computer Vision (**WACV**), Steamboats Springs, USA, January 2014.

2. **Abhinav Dhall**, Roland Goecke, Jyoti Joshi, Michael Wagner and Tom Gedeon, Emotion Recognition In The Wild 2013, In Proceedings of the 2013 ACM International Conference on Multimodal Interaction (**ICMI**), Sydney, Australia, December 2013.

3. **Abhinav Dhall**, Context based Facial Expressions In The Wild, In Proceedings of the Fifth Biannual Humaine Association Conference on Affective Computing and Intelligent Interaction (**ACII 2013**), Pages 636-641, Geneva, Switzerland, September 2013.

4. **Abhinav Dhall**, Expression Recognition In The Wild: From Individual to Groups, Doctoral Consortium, In Proceedings of the 2013 ACM International Conference on Multimedia Retrieval **ICMR 2013**, Pages 325-328, Dallas, USA, April 2013. **[Best Doctoral Consortium Paper Award]**

5. **Abhinav Dhall**, Roland Goecke, Simon Lucey and Tom Gedeon, A Semi-Automatic Method for Collecting Richly Labelled Large Facial Expression Databases from Movies, **IEEE MultiMedia**, Pages 34-41, 2012.

6. **Abhinav Dhall**, Jyoti Joshi, Ibrahim Radwan and Roland Goecke, Finding Happiness In Social Context, In Proceedings of the 2012 Asian Conference on Computer Vision, (**ACCV 2012**), Pages 613-626, Daejeon, Korea, November 2012.

7. **Abhinav Dhall** and Roland Goecke. Group Expression Intensity Estimation in Videos via Gaussian Processes. In Proceedings of 2012 International Conference on Patter Recognition (**ICPR 2012**), Pages 3525-3528, Tsukuba, Japan, November 2012.

8. **Abhinav Dhall**, Roland Goecke, Simon Lucey and Tom Gedeon, Static Facial Expressions In Tough Conditions: Data and Evaluation Protocol And Benchmark, In Proceedings of 2011 International Conference on Computer Vision (**ICCV 2011**) Workshop BEFIT, Pages 2106-2112, Barcelona, Spain, November 2011.

9. Akshay Asthana, Miles Delahunty, **Abhinav Dhall** and Roland Goecke, Facial Performance Transfer via Deformable Face Models and Parametric Correspondence, In IEEE Transactions on Visualisation and Computer Graphics (**TVCG**), Regular Paper, Pages 1-9, IEEE Computer Society, 2011.

10. **Abhinav Dhall**, Akshay Asthana and Roland Goecke, Emotion recognition using PHOG and LPQ features, In Proceedings of 2011 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (**FG 2011**) Worksshop Facial Expression Recognition & Analysis, Pages 878-883, Santa Barbara, California, USA, March 2011.

11. **Abhinav Dhall**, Akshay Asthana and Roland Goecke, A SSIM-Based Approach for Finding Similar Facial Expressions, In Proceedings of 2011 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (**FG 2011**) Workshop on Emotion Synthesis Representation and Analysis in Continuous Space, Pages 815-820, Santa Barbara, California, USA, March 2011.

12. **Abhinav Dhall**, Akshay Asthana and Roland Goecke, Facial Expression Based Automatic Album Creation, In Proceedings of 2010 International Conference on Neural Information Processing (**ICONIP 2010**), Pages 485–492, Sydney, Australia, November 2010.

Other Publications Relevant to but not Forming Part of the Thesis

1. Karan Sikka, **Abhinav Dhall** and Marian Bartlett, Weakly Supervised Pain Localization and Classification with Multiple Segment Learning, Image & Vision Computing (**IVC**) journal Best of Faces & Gesture Recognition 2014.

2. Nandita Sharma, **Abhinav Dhall**, Tom Gedeon and Roland Goecke, Thermal spatio-temporal data for stress recognition, EURASIP Journal on Image and Video Processing (**JVIP**) 2014.

3. Shyam Rajagopalan, **Abhinav Dhall** and Roland Goecke, Self-Stimulatory Behaviours in the Wild for Autism Diagnosis, In Proceedings of the 2013 IEEE International Conference on Computer Vision (**ICCV 2013**) Workshop Decoding Subtle Cues from Social Interactions, Sydney, Australia, December 2013.

4. Ibrahim Radwan, **Abhinav Dhall** and Roland Goecke, Monocular Image 3D Human Pose Estimation under Self-Occlusion, In Proceedings of the 2013 IEEE International Conference on Computer Vision (**ICCV 2013**), Pages 1888-1895, Sydney, Australia, December 2013.

5. Orungati Venkatesh Ramana Murthy, Ibrahim Radwan, **Abhinav Dhall** and R. Goecke, On the Effect of Human Body Parts in Large Scale Human Behaviour Recognition. in Proceedings of the International Conference on Digital Image Computing: Techniques and Applications (**DICTA 2013**), Hobart, Australia, November 2013.

6. Jyoti Joshi, Roland Goecke, **Abhinav Dhall**, Sharifa Alghowinem, Michael Wagner, Michael Breakspear, Julien Epps, Gordon Parker, Multimodal Assistive Technologies for Depression Diagnosis and Monitoring, Springer Journal of MultiModal User Interfaces (**JMUI**), September 2013.

7. Nicholas Cummins, Jyoti Joshi, **Abhinav Dhall**, Vidhyasaharan Sethu, Roland Goecke and Julien Epps. Diagnosis of Depression by Behavioural Signals: A Multimodal Approach. In Proceedings of the 2013 International Audio/Visual Emotion Challenge and Workshop (**AVEC 2013**), 21st ACM International Conference on Multimedia (**MM 2013**), Barcelona, Spain, October 2013.

8. Jyoti Joshi, **Abhinav Dhall**, Roland Goecke and Jeffery F. Cohn, Relative Body Parts Movements, In Proceedings of the Fifth Biannual Humaine Association Conference on Affective Computing and Intelligent Interaction (**ACII 2013**), Pages 492-497, Geneva, Switzerland, September 2013.

9. Nandita Sharma, **Abhinav Dhall**, Tom Gedeon and Roland Goecke, Modeling Stress Using Thermal Facial Patterns: A Spatio-Temporal Approach, In Proceedings of the Fifth Biannual Humaine Association Conference on Affective Computing and Intelligent Interaction (**ACII 2013**), Pages 387-392, Geneva, Switzerland, September 2013.

10. Karan Sikka, **Abhinav Dhall** and Marian Bartlett, Weakly supervised pain localization using multiple instance learning, 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Pages 1-8, Shanghai, China, April 2013. **[Best Student Honorable Mention Paper Award]**

11. Jyoti Joshi, **Abhinav Dhall**, Roland Goecke, Michael Breakspear and Gordon Parker, Neural-net classification for spatio-temporal descriptor based depression analysis, In Proceedings of 2012 International Conference on Patter Recognition (**ICPR 2012**), Pages 2634-2638, Tsukuba, Japan, November 2012.

12. Ibrahim Radwan, **Abhinav Dhall** and Roland Goecke. Correcting Pose Estimation with Implicit Occlusion Detection and Rectification, In Proceedings of 2012 International Conference on Patter Recognition (**ICPR 2012**), Pages 3496-3499, Tsukuba, Japan, November 2012.

13. Ibrahim Radwan*, **Abhinav Dhall***, Jyoti Joshi and Roland Goecke, Regression Based Pose Estimation with Automatic Occlusion Detection and Rectification, In Proceedings of 2012 IEEE International Conference on Multimedia & Expo (**ICME 2012**), Melbourne, Australia, July 2012. [***Equal first authors**] **[Nominated for Best Paper Award]**

List of Abbreviations

AAM	Active Appearance Model Cootes et al. [2002]
AFEW	Acted Facial Expressions In The Wild Dhall et al. [2012a]
ASM	Active Shape Model Cootes et al. [1995]
AU	Action Unit Ekman and Friesen [1978]
AVOZES	Audio-Video OZstralian English Speech database Goecke and Millar [2004]
CK	Cohn-Kanade database Kanade et al. [2000]
CK+	Extended Cohn-Kanade database Lucey et al. [2010]
CLM	Constrained Local Model Saragih and Goecke [2009]
CMU-PIE	Carnegie Mellon University (CMU) PIE database Sim et al. [2003]
EI	Expression Image
EmotiW	Emotion Recognition In The Wild Dhall et al. [2013]
FACS	Facial Action Coding System Ekman and Friesen [1978]
FAP	Facial Animation Parameters
FER	Facial Expression Recognition
GPR	Gaussian Process Regression Rasmussen [2006]
HAPPEI	HAPpy PEople Images database Dhall et al. [2012b]
HOG	Histogram of Oriented Gradients Dalal and Triggs [2005a]
PHOG	Pyramid of Histogram of Oriented Gradients Bosch et al. [2007]
IEBM	Iterative Error Bound Minimisation Saragih and Göcke [2006]
LDA	Latent Dirichlet Allocation Blei et al. [2001]
LBP	Local Binary Pattern Ojala et al. [2002]
LBP-TOP	Local Binary Pattern - Three Orthogonal Planes Zhao and Pietikainen [2007]
LGBP	Local Gabor Binary Pattern Zhang et al. [2005b]

LGBP-TOP	Local Gabor Binary Pattern - Three Orthogonal Planes Almaev and Valstar [2013]
MFCC	Mel-Frequency Cepstral Coefficients
Multi-PIE	Carnegie Mellon University (CMU) Multi-PIE database Gross et al. [2008a]
PAW	Piecewise Affine Warping
PCA	Principal Component Analysis
RGB	Red, Green and Blue
RMSE	Root Mean Squared Error
RSE	Residual Shape Error
RTE	Residual Texture Error
SIC	Simultaneous Inverse Compositional
SFEW	Static Facial Expressions In The Wild database Dhall et al. [2011c]
SSIM	Structural SIMilarity index metric Wang et al. [2004b]
SVR	Support Vector Regression Chang and Lin [2001]
SVM	Support Vector Machine Chang and Lin [2001]

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The recent advancement of social media has given users a platform to socially engage and interact with a larger population. Millions of images and videos are being uploaded every day by users on the web from different events and social gatherings. There is an increasing interest in designing systems capable of understanding human manifestations of emotional attributes and affective displays. For inferring the affective state of the users in such images and videos, captured in real-world conditions, methods which can perform facial expression analysis 'in the wild' are required. Here, the term **'in the wild'** signifies different environments/scenes and backgrounds, illumination conditions, head pose, occlusion *etc.* (as illustrated in Figure 1.1). Automatic facial expression analysis has made a significant progress in last two decades. However, such developed frameworks have been strictly employed to the data collected in controlled laboratory settings with frontal faces, perfect illumination and posed expressions. On contrary, the images and videos on the web have been captured in different, unconstrained environments and this poses a big challenge to automatic facial expression analysis methods.

Facial expressions are the facial changes in response to a person's internal emotional states, intentions, or social communications Tian et al. [2005]. Facial expression analysis is an effortless task for human beings. Along with head movement, facial expressions are a major parts of non-verbal communication. Automatic Facial Expression Recognition (FER) is an active field of research for over two decades now. FER has a pivotal role in developing interfaces, which can understand affective displays and emotional responses of a user. However, the trivial task (for humans) of facial expression understanding 'in the wild' becomes non-trivial for computers due to various challenges posed by the real-world conditions.

FER finds applications in affective computing (autism syndrome diagnosis, depression analysis, pain detection, stress classification, drowsiness detection etc.), intelligent environments (smart homes), lie detection, psychiatry, emotion and paralinguistic com-

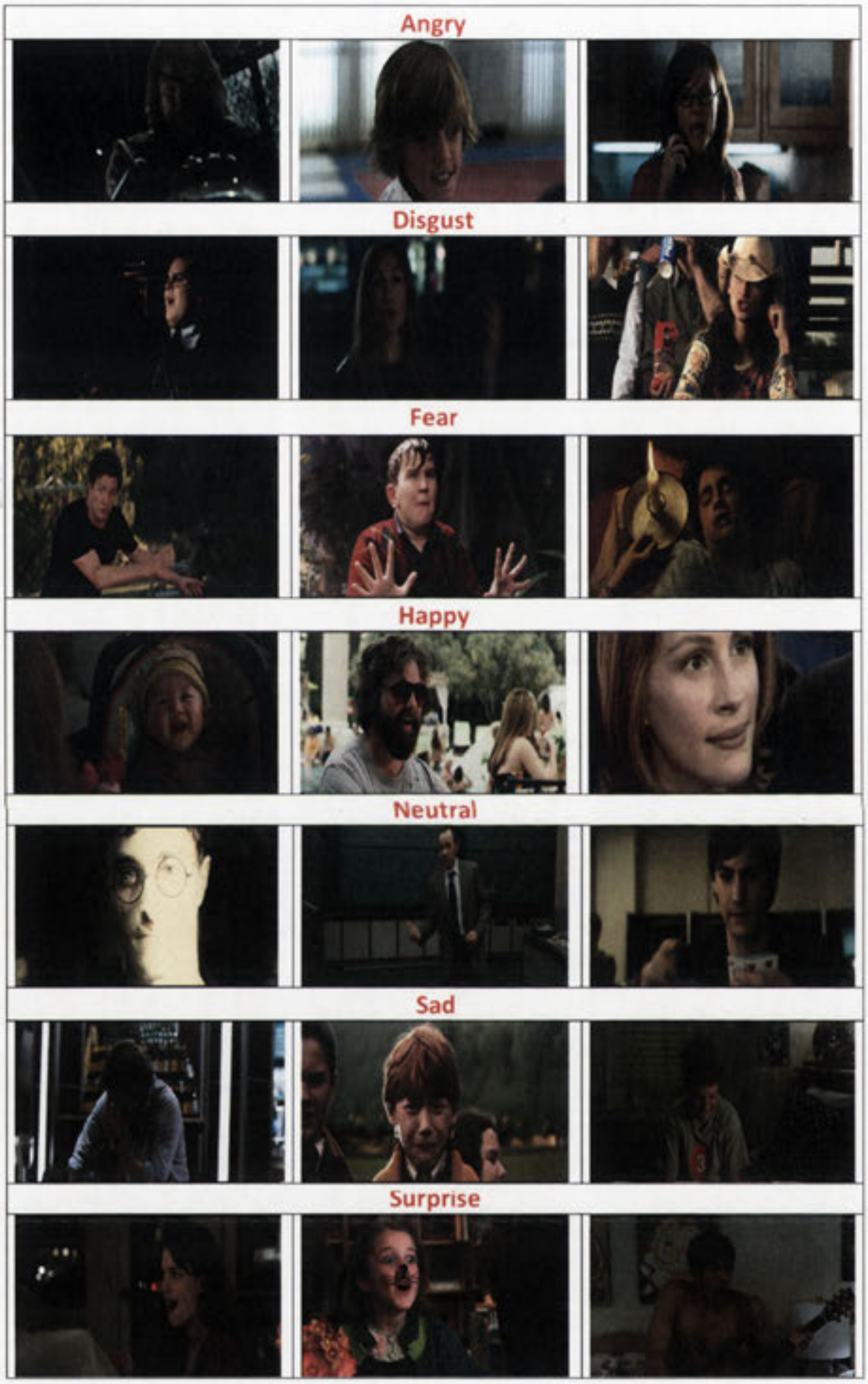


Figure 1.1: The frames in the figure are from the AFEW database Dhall et al. [2012a]. The figure depicts the challenges of FER 'in the wild' - out-of-plane head movement, illumination, different backgrounds and occlusion.

munication, computer games (e.g. Microsoft Kinect), expression driven facial animation (e.g. facial movements in the movie Avtaar) and multimodal Human Computer Interfaces (HCI). Therefore, it is of strong interest to both academia and industry and this makes research in FER in real-world conditions a worthwhile goal to pursue.

1.1 Challenges

There are several challenges which need to be addressed for FER 'in the wild'. Consider an illustrative example of categorization i.e. assigning an expression label to a video clip of a subject(s) protesting at the Tahrir square in Egypt during the 2011 protests. In order to learn an automatic system which can infer the label representing the expression, labelled data containing video clips representing different expressions in diverse settings is required. There exists a lot of standard expression databases such as the Cohn-Kanade (CK) Kanade et al. [2000], Multi-PIE Gross et al. [2008a], FEEEDTUM Wallhoff [2006] and RU-FACS Bartlett et al. [2006], which include both static and dynamic data of subjects displaying a fixed set of expressions. However, all of these databases contain samples with posed/spontaneous expressions under lab-controlled conditions. Ideally, one would like to collect spontaneous data in real-world conditions. However, as anyone working in the emotion research community will attest, collecting spontaneous data in real-world conditions is a tedious task. Therefore, new methods which can speed up the process of creating databases representing real-world conditions are required.

Once the data is available, the next challenge is face and facial parts detection, followed by head pose normalisation. As the subject in the Tahrir square video clip example generally moves his/her head, it poses another challenge of out-of-plane head movements. This effects face and facial parts detection, which are required for Head Pose Normalisation (HPN). If a face has a non-frontal head pose, alignment becomes a non-trivial task and this can introduce noise/error during the feature extraction step. During spontaneous expressions, subjects move their arms and hands as part of the non-verbal communication, this leads to the problem of occlusion. Occlusion needs to be detected and localised for finding accurate fiducial points.

The complexity of such video clips (like the one in the example above) increases with the presence of multiple subjects. Research in this field has been focussing on the recognition of a single subject's expression i.e. given a video clip or image, only a single subject is present in it. However, data being uploaded on the web, specially revolving around social events such as the illustrated example contains group of people. Group mood analysis finds its application in opinion mining, image and video album creation, image visualisation and early violence prediction among others. There has been work in psychology on the analysis of emotion of group of people, cues from this can be taken on creating models for handling group emotions. The major challenges for group mood analysis are: 1) labelled data representing various social scenarios; 2) robust face and fiducial points detector (this is relevant for a single subject scenario as well, though not for lab-controlled scenario databases given the current state-or-art detectors Zhu and Ramanan [2012]) and 3) models which can take into consideration the affective compositional effects and the affective context. A simple solution to group mood analysis is emotion averaging. However, in real-world conditions averaging is not ideal. This motivates the research for models, which accommodate various attributes that effect the perception of group mood and their interaction.

1.2 Objective

The first objective of this thesis is to devise methods for collecting databases representing real-world like conditions. A trivial way to collect real-world data is to manually record data. However, this is a time consuming and error-prone process. Recent human action recognition databases such as the Hollywood database Laptev et al. [2008] have been collected from movies. The Hollywood database has provided a new platform to researchers and has contributed to the progress in the field of automatic human action recognition in real-world settings. Inspired by this approach, movies can also be used for developing facial expression databases. A simple approach for using movies for collecting data is manual parsing. However, manual segmentation of movies is time consuming and automated methods need to be formulated.

The second objective of this thesis is to propose a HPN method, which can be used for real-world images. Generally, HPN methods are based on fiducial points, which can induce noise during HPN, when there is an error in fiducial points detection. HPN methods also require head pose information, which itself is an open problem. Therefore, a method, which does not require head pose information and is not dependent on fiducial points, is proposed.

The third objective of this thesis is to model the expression (mood)[1] of a group of people in an image. Current, FER methods are based on single subject per sample only. The research question here is what is the effect of social features (e.g. where people are standing in a group, with whom are they standing etc.) on the perception of the mood of a group. How can social features be integrated in a model for inferring the mood of a group of people? Interesting ideas can be taken from the latest work in age and gender inference in group photographs Gallagher and Chen [2009]. In their work, a group information based prior is used for inferring the age and gender of subjects of a group. Similar group related information can be used for inferring the contribution

[1] In this thesis the terms 'expression' and 'mood' are being used interchangeably when referring to a group of people in an image.

of each subject of a group towards the perception of mood of that group in an image.

The fourth objective of this thesis is to devise methods based on structural similarity for finding similar facial expressions. It would be interesting to find out if shape alone can be used for finding similar facial expressions in controlled settings. This further can be applied to practical applications such as facial expressions based automatic album creation.

1.3 Contributions

In the following, a list of the novel contributions of this thesis is presented -

1. A novel semi-automatic method is proposed for collecting a temporal facial expressions database (*Acted Facial Expressions In The Wild (AFEW)* Dhall et al. [2012a]) representing real-world conditions. This attempts to solve the problem of how to speed up the process of database creation and labelling. Moreover, the approach also proposes a solution for collecting temporal facial expressions database representing real-world scenarios. The method recommends short video clips and labels to the labeller based on the presence of emotion related keywords in the closed caption subtitles. Further, based on the AFEW database, an image based facial expression database (*Static Facial Expressions In The Wild (SFEW)* Dhall et al. [2011c]) is proposed. Strict experimental protocols based on the presence or absence of the same subjects in the train and the test set are proposed for both AFEW and SFEW to facilitate the comparison of results by different researchers.

2. A discriminative response map based HPN method is proposed. Traditional HPN methods are based on fiducial points as input. However, fiducial points detection is an open-problem for real-world images. Therefore, the proposed HPN method is based on texture (sparse response maps) normalisation. An extension is proposed to make the HPN method invariant to head pose information.

3. A method is proposed for finding similar facial expressions based on structural similarity comparison Dhall et al. [2011a]. An automatic image album creation method Dhall et al. [2010] is proposed on the basis of finding similar facial expressions. The proposed method is further applied for finding similar facial expressions for creating training sets for learning a facial performance transfer model.

4. Group mood inference model is proposed based on social context. The method takes into consideration both global and local factors for deciding the contribution of each subject towards the perception of mood of the group an image. A new

database (HAPpy PEople Images (HAPPEI) Dhall et al. [2012b]) is proposed based on Flickr search.

1.4 Structure of the Thesis

The background study of facial expression analysis system is presented in Chapter 2. A typical FER flow is presented in Figure 2.1. Various stages of a FER methods and various state-of-art FER methods are discussed. FER methods are divided into different categories based on different criteria like image or video signal, geometric or appearance feature descriptor. A literature survey is presented describing these various FER methods.

Chapter 3 - Facial Expressions In The Wild Database: The chapter proposes a semi-automatic technique for collecting a temporal facial expressions database which mimics the real-world conditions. The details of the data extraction and labelling process are discussed. The proposed database AFEW is an audio-video dataset collected from movies. An image based facial expression database SFEW is manually extracted from AFEW. Experimental protocol is proposed for AFEW and SFEW. AFEW further forms the base of the Emotion Recognition In The Wild (EmotiW) grand challenge Dhall et al. [2013] and its details are presented.

Chapter 4 - Head Pose Normalisation: This chapter proposes a HPN method based on the Mixture of Pictorial Structures (MoPS) Zhu and Ramanan [2012] approach. Rather than using fiducial points for HPN, the proposed HPN method is based on the part-detector's confidence maps, which are used for generating the virtual frontal head pose. The proposed method and its implementation details are discussed in detail. The technique is experimented on real-world data and outperforms the state-of-art HPN algorithms.

Chapter 5 - Similar Facial Expressions and Applications: This chapter proposes a novel structural similarity based approach for finding similar facial expressions. The proposed technique is applied to the problems of 'facial expressions based automatic album creation', 'album by similar facial expression' and for finding similar facial expression images for learning models for facial performance transfer. The details of the proposed method, its application and experiments are discussed in detail in the chapter.

Chapter 6 - Analysis of the Mood of a Group of People: In this chapter, a novel framework for inferring the mood of a group of people in an image is proposed. A graphical model based on social contextual features is proposed. A new database HAPPEI is collected from Flickr. The details of the mood inferring model, social features are discussed in detail. The group mood analysis is applied to two applications of 'event summarisation' and 'candid photo shot selection'.

Chapter 7 - Conclusion and Future Work: The chapter summarises the contribution and important learnings from the work done in the thesis. The possible future directions based on the work presented in the thesis are thoroughly discussed.

Chapter 2

Background

This chapter discusses the current state-of-the-art in facial expression analysis. The structure of this chapter is as follows: the chapter introduces typical facial expression analysis method's flowchart in Section 2.1. Various face and fiducial points detection techniques are discussed in Section 2.2. Facial descriptors are discussed in Section 2.3. Classification techniques used in FER literature are discussed in Section 2.4. The literature review for facial expression databases (Chapter 3) is discussed in Section 2.5. The literature review for HPN (Chapter 4) is discussed in Section 2.6. The literature review for group mood analysis (Chapter 6) is presented in Section 2.7. Some of the prominent FER applications are discussed in Section 2.8.

2.1 Facial Expression Analysis Pipeline

A detailed survey of FER methods is presented in Fasel and Luettin [2003], Cohen et al. [2003] , Pantic and Rothkrantz [2004] and Zeng et al. [2009]. A typical facial expression analysis system has the following main components: face and fiducial points detection, feature extraction, and classification. Figure 2.1 displays the main blocks of a typical FER system. Once the image has been captured, face detector such as the popular Viola-Jones (VJ) detector Viola and Jones [2001] is used to locate the face. Further, facial parts location (fiducial points) are inferred using parametric models such as the Active Appearance Models (AAM) Cootes et al. [1998], which have gained a lot of popularity in the recent years. Once the location of the facial parts such as eyes, mouth are known, facial features are computed. Features can be extracted either on a holistic level or on individual parts of the face. The individual features computed in the latter case are concatenated to construct the final feature vector. Further, dimensionality reduction methods such as Principal Component Analysis (PCA) can be applied to the feature vector. This gives a compact, discriminating and less noisy feature representation. The classifier then classifies a face's state into either fixed

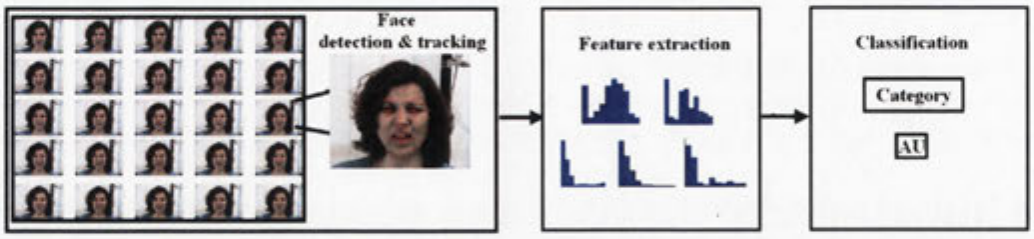


Figure 2.1: A typical FER system Dhall et al. [2011c].

emotion classes or in the the FACS Ekman and Friesen [1978] format. The components of a FER are system are discussed in detail in the following sections.

2.2 Face and Fiducial Points Detection

Given an image the face detection algorithm computes the location of the face in the image. Early methods based on skin colour segmentation, edge detection and other heuristics have been discussed in Hjelmås and Low [2001]. A recent survey on face detection literature can be found in Zhang and Zhang [2010].

One of the classic and most used face detector is the VJ Viola and Jones [2001] face detector. It is based on a cascade based boosting framework Freund and Schapire [1995] in which haar-like features are extracted quickly using an integral image. The use of integral image leads to near real-time execution of the face detection method. The cascade classifier scans an image using a sub-window at different scales and localises the patches which are labelled as faces by the classifier. The Adaboost algorithm is also applied for selecting discriminative haar-features during the training phase of the classifier. The cascade classifier consists of various weak classifiers which together act as a strong classifier. The use of weak classifiers early on helps in fast rejection of patches which do not resemble to faces at all. The open source OpenCV computer vision library contains an implementation of the VJ object detector. For multi-view face detection, multiple models are learnt for different head pose Jones and Viola [2003]. The VJ detector has long training time, Wu et al. [2008] propose a greedy feature selection approach. They select features using forward feature selection before training the cascade classifier.

There are several other face detection methods such as based on energy-based models which infers the face location and head pose simultaneously Osadchy et al. [2007] and vector boosting based methods where a tree representation is proposed for dividing the face space into smaller subspaces Huang et al. [2005].

Once the location of the face is known for finding the facial parts location non-rigid deformable models have been proposed (Cootes et al. [1995], Cootes et al. [1998], Saragih et al. [2009] etc.). The power of the non-rigid deformable models stems from

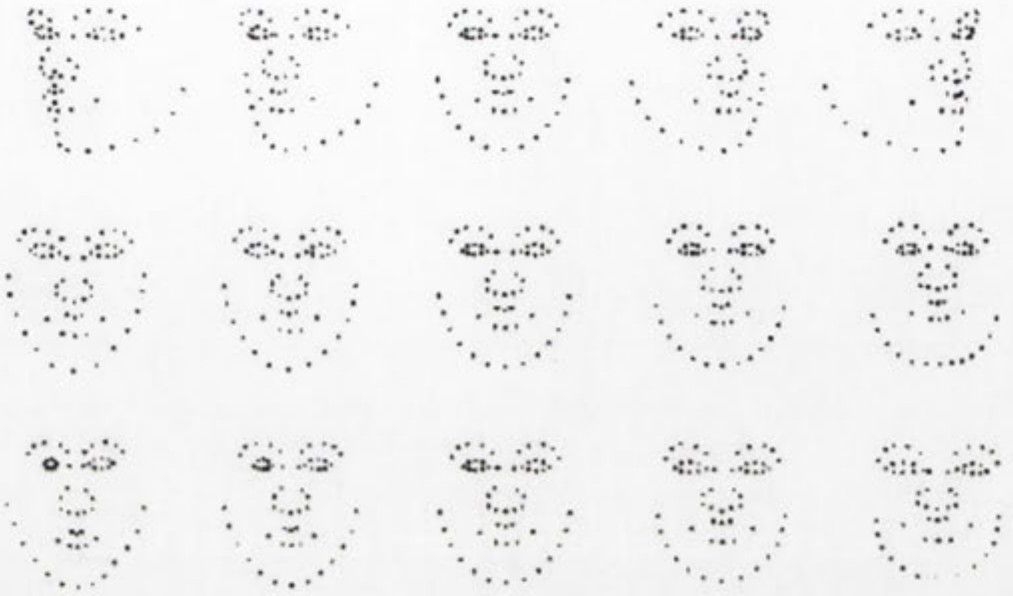


Figure 2.2: Visualisation of top three modes of shape variation Asthana [2013] generated using the DeMoLib library Saragih and Goecke [2007].

their low-dimensional representation of the shape and texture of a face. One of the earliest deformable model method is the Active Contour Model proposed by Kass et al. [1988]. The Active Shape Model (ASM) algorithm Cootes et al. [1995] models the shape of an object and has been used extensively in face tracking. During the training process, the landmark points of all the input samples are aligned into a common co-ordinate frame by Procrustes analysis. Post this the model is computet by applying PCA over the shapes. The shape of the model can be controlled/changed via parameters of deformation. Then fitting of the model can be performed on a new image using an iterative method which calculates the best match for the model boundary and hence decides the new location for the model points.

2.2.1 Active Appearance Model

The Active Appearance Models (AAM) are an extension of the ASM, they not only model the shape but also consider the grey-level appearance. During the training process the grey-level appearance is modeled by warping each training sample image using a triangulation algorithm which aligns it to the mean shape. The grey level information is then sampled and PCA is applied to the samples. Hence, this model can then be fitted to a new image using an optimisation approach which uses the difference between intensities of the learnt model and the reference image.

For constructing a shape model, each training image is represented by a $2n$-dimensional

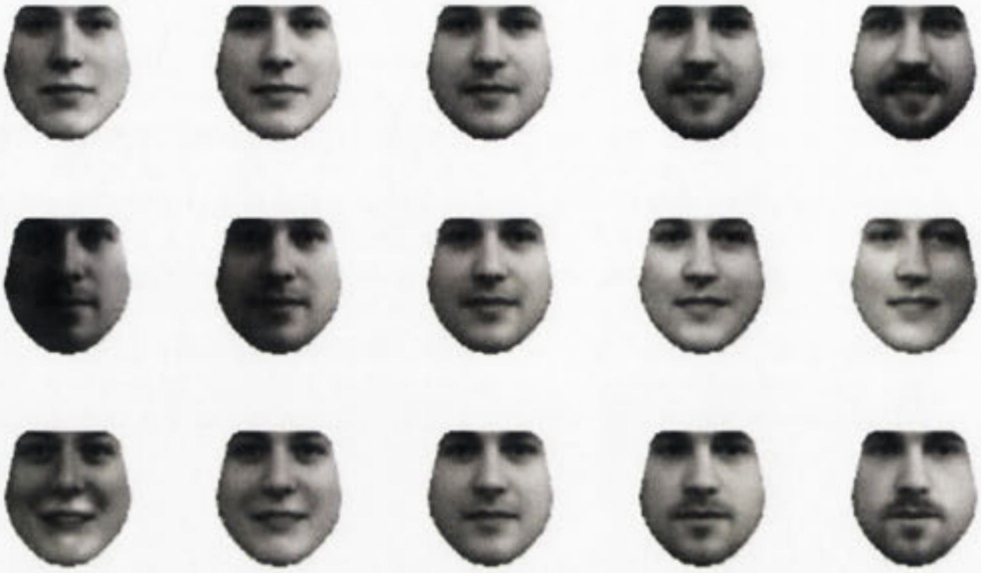


Figure 2.3: Visualisation of top three modes of shape variation Asthana [2013] generated using the DeMoLib library Saragih and Goecke [2007].

shape vector:

$$\mathbf{s} = [x_1; y_1; \ldots; x_n; y_n]^T \tag{2.1}$$

where n is the number of landmark points describing the shape of the face. Next, PCA is applied to the aligned shapes. This shows various modes representing the variation in the training data. Figure 2.2 shows the top three modes of variation. The top row (Mode 0) represents the variation in shape introduced by pose change across the yaw direction. The middle row (Mode 1) represents the variation in shape introduced by pose change across the pitch direction. The bottom row (Mode 3) represents the shape variation around the mouth area. These modes are data-dependent and can be different for different sets of training data.

Next, piecewise affine warp is applied to the texture of all the the training samples in order to warp the texture to a canonical frame. This produces shape normalised texture. The texture of a training sample can be represented as:

$$\mathbf{t} = [r_1, g_1, b_1, \ldots, r_m, g_m, b_m]^T \tag{2.2}$$

where $\mathbf{t}$ is a three-dimensional vector and m is the number of pixels in a texture. Figure 2.3 describes the top three modes of texture variation.

There are a number of AAM fitting algorithms and they can be broadly classified into two classes: generative and discriminative fitting. In generative fitting class of methods (*Fixed Jacobian* Cootes et al. [1998], *Project Out Inverse Compositional* Baker and Matthews [2001], *Simultaneous Inverse Compositional* Baker et al. [2003]), minimisation/maximisation of some measure of fitness between the model's texture and warped image region. In discriminative fitting class of methods (*Iterative Error Bound Minimisation* Saragih and Göcke [2006], *Haar-like Feature Based Iterative Discriminative Method* Saragih and Göcke [2007]) a relationship is learnt between the features and the parameters updates, by using the features extracted from parameters settings which are perturbed from their optimal setting in each image.

The disadvantage of AAM is their limitation in generalization to unseen subjects. AAM models can be classified as subject-dependent or subject-independent. Subject-dependent AAM is for the scenario when the train and test images have the same subjects. Subject-independent is for scenarios where the subjects in the train and test set are different. Subject-independent performance is important for face analysis on the web. Constrained Local Models (CLM) Saragih et al. [2009] are an extension of the AAM algorithms. The texture is divided into blocks. This helps in generalisation and better subject-independent performance. Subject-dependent AAM methods perform better than subject-independent CLM Chew et al. [2012]. However, the current state-of-art descriptors compensate for small errors introduced by subject-independent CLM Chew et al. [2012].

Another limitation of AAM/CLM is their requirement of large amount of labelled data representing different scenarios such as illumination, pose and expression during training. Labelling fiducial points is a manually laborious and erroneous task. Recent work by Asthana et al. [2009a] proposes a database augmentation by generating synthetic data representing different pose and expressions. Asthana et al. [2009a] learn regression models (over fiducial points and texture parameters) for *eg.* from frontal view to a non-frontal view and generate new samples during test inference. More recently, Sagonas et al. [2013] have used AAM for semi-automatically labelling fiducial points in databases. They infer fiducial points on test images and manually select accurately inferred fiducial point face samples. The AAM is then re-trained using the existing data and the new correctly localised fiducial points. This active learning approach is repeated till all faces in the test set are labelled.

2.2.2 Pictorial Structure

In the Pictorial Structure (PS) framework Felzenszwalb and Huttenlocher [2005], an object is represented as a graph with n vertices $V = \{v_1, ..., v_n\}$ for the parts and a set of edges E, where each $(v_i, v_j) \in E$ pair encodes the spatial relationship between parts i and j. For a given image I, PS learns two models. The first one learns the evidence of each part as an *appearance model*, where each part is parameterised by its location (x, y), orientation θ, scale s, and foreshortening. All of these parameters (together referred to as $\mathcal{D}$) are learned from exemplars and produce a likelihood model for I. The second model learns the kinematic constraints between each pair of parts in a prior *configuration model*. $\mathcal{L}$ is the parts configuration. Given the two models, the posterior distribution over the whole set of part locations is

$$p(\mathcal{L}|\mathcal{I},\mathcal{D}) \propto p(\mathcal{I}|\mathcal{L},\mathcal{D})p(\mathcal{L}|\mathcal{D}) \tag{2.3}$$

where $p(\mathcal{I}|\mathcal{L},\mathcal{D})$ measures the likelihood of representing I in a particular configuration and $p(\mathcal{L}|\mathcal{D})$ is the kinematic prior configuration. A major problem of this framework is the low contribution of the occluded parts, resulting in either erroneous or missing detections of these parts, leading to an inaccurate pose estimation. Everingham et al. [2006] proposed a PS based fiducial point detector, which is initialised using VJ face detector. Recently, Zhu and Ramanan [2012] proposed an extension to the PS framework by adding mixtures representing different face poses. Zhu and Ramanan [2012] performs face and fiducial point detection and head pose inference in one framework. The face detector performs better than the VJ face detector. The disadvantage of PS based method like Everingham et al. [2006] is that it requires initialisation from a face detector like VJ. Zhu and Ramanan [2012] overcomes this limitation by using multiple pose as mixture detectors (The method is discussed in detail in Section 4.2).

Selecting the appropriate face and fiducial point detector is problem driven. For example in the case of affect analysis problems such as depression analysis, it is desirable that the system should generalise over subjects. Joshi et al. [2012] use PS of Everingham et al. [2006] for fiducial points detection. Even though the MoPS framework performs better than PS of Everingham et al. [2006], Joshi et al. [2012] prefer the use of Everingham et al. [2006]. They argue that the inference time for MoPS is way longer than Everingham et al. [2006], which matters for analysing long duration depression video clips. Furthermore, they use spatio-temporal descriptors (LBP-TOP Zhao and Pietikainen [2007]) that compensate for errors generated by PS of Everingham et al. [2006]. On the other hand, application such as facial performance transfer Asthana et al. [2012], require accurate fiducial points. Asthana et al. [2012] use subject dependent AAM models as they are more accurate as compared to CLM and subject-independent AAM.

A very recent approach Asthana et al. [2013b], computes response maps from patch experts. These response maps are generated by discriminative patch experts. The AAM fitting is performed on top of these response maps. Asthana et al. [2013b] found that this gave better results as compared to the state-of-the-art methods Saragih and Goecke [2009], Zhu and Ramanan [2012]. In another new work Zhou et al. [2013], a graph matching based approach is proposed. Interest points are generated using RANSAC and an affine-invariant shape constraint is learnt online using similar exemplars. Fiducial point detection is thus computed by solving a graph matching problem.

2.3 Facial Descriptors

Once the face and facial parts location is identified, feature descriptors are computed on the aligned face for extracting information for learning classifiers. FER techniques can be segregated on the basis of type of descriptors used. Generally, facial descriptors can be broadly classified as geometric and appearance. Geometric features Pantic and Rothkrantz [2000], Dhall et al. [2010], Zhang et al. [2011] correspond to facial points and to the location of different facial parts. Appearance features generally correspond to the face texture information Zhao and Pietikainen [2007], Dhall et al. [2011b], Tariq et al. [2012], Sikka et al. [2012], Almaev and Valstar [2013].

Popular geometric feature are based on Facial Animation Parameters[1] (FAP). FAP are defined in the MPEG-4 video coding standard. Lavagetto and Pockaj [1999] presents method for synthesizing facial animations using FAP and Facial Definition Parameters[2] (FDP).

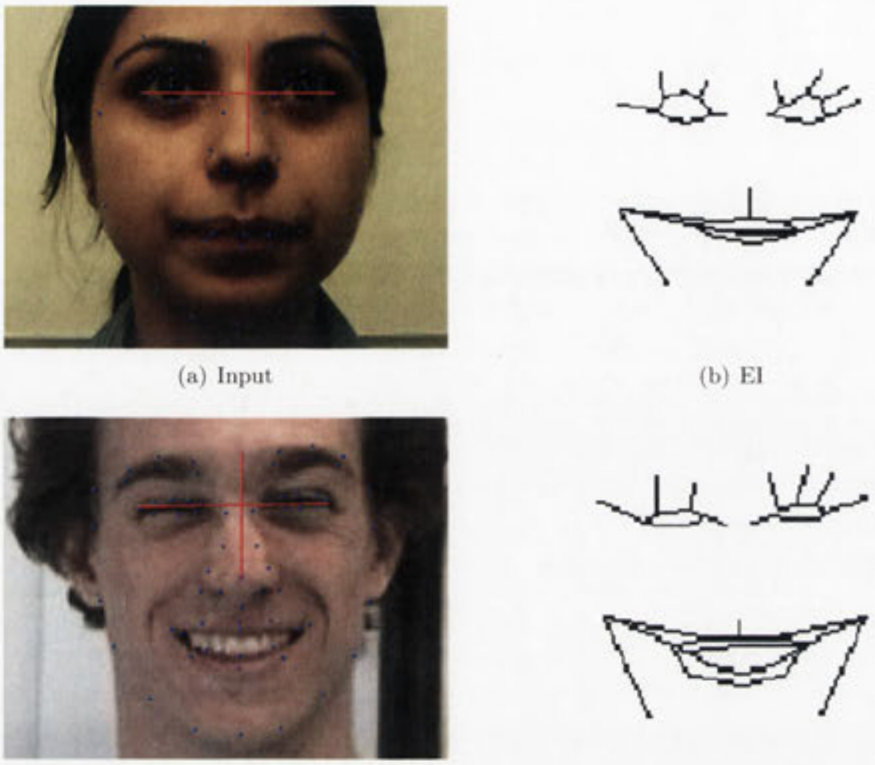
Valstar et al. [2005], Valstar and Pantic [2006] extracted geometric features based on analysis of tracked facial points. They compute: a) euclidian distance between the points of a face between two consecutive frames; b) euclidian distance based on the increase in the distance between two points in a frame with respect to the distance between the same points in frame *1*; c) vertical distance between the same point in frame *1* and frame n; and d) vertical distance between the same point in frame *1* and frame n.

Sebe et al. [2007] use Piecewise Beizier Volume Deformation (PBVD) tracker for tracking the facial parts. The motion information between two consecutive frames is measured using template matching. Further, various classifiers are compared. Asthana et al. [2009b] compute geometric features by fitting AAM models on input faces. They compared various AAM fitting techniques and experimented on CK+ database. One of the limitations of this work is that it required manual initialisation of facial parts.

In Dhall et al. [2010], a geometric descriptor called Emotion Image (EI) is proposed (For details see Section 5.2.2). This feature constructs a visual map based on an undirected map based on input from a facial points detector. EIs of two faces (images) is compared using Structural Similarity Index Metric (SSIM) Wang et al. [2004b] (For details see Section 5.2.3) for computing their similarity and is applied to the problem expression based album creation. Figure 2.4 shows EI for two face samples. The discriminative ability of EI is dependent on fiducial point detection quality, this may introduce some errors when the fiducial points detection is not very accurate.

[1] http://www.dsp.dist.unige.it/~pok/RESEARCH/MPEG/fapspec.htm

[2] http://www.dsp.dist.unige.it/~pok/RESEARCH/MPEG/fdpspec.htm



(a) Input

(b) EI

(c) Input

(d) EI

Figure 2.4: (a) and (c) are the input faces with red lines representing the landmark points, which are static with respect to local facial moment and donot contribute to the expression. (b) and (d) are the corresponding Expression Images (EI), which are compared for their structural similarity.

Valstar et al. [2005], Valstar and Pantic [2006] show that the performance of geometric features is similar to that of the appearance features. However, the limitation of geometric feature comes from their dependence on accurate facial parts location information. Facial parts detection is relatively accurate on lab-controlled scenario data, however, it is still an open problem for images in real-world conditions. If there is an error in the facial parts detection, the error generally propagates in the geometric feature representation. Appearance features deal better with mis-alignment. Chew et al. [2012] argue that appearance descriptors are able to compensate error produced by facial parts detectors upto some extent. Popular appearance descriptors are described below:

Local Binary Patterns (LBP): The LBP family of descriptors has been extensively used in computer vision for texture and face analysis Ojala et al. [2002], Zhao and Pietikainen [2007], Ojansivu and Heikkilä [2008]. The LBP descriptor assigns binary labels to pixels by thresholding the neighbourhood pixels with the central value.

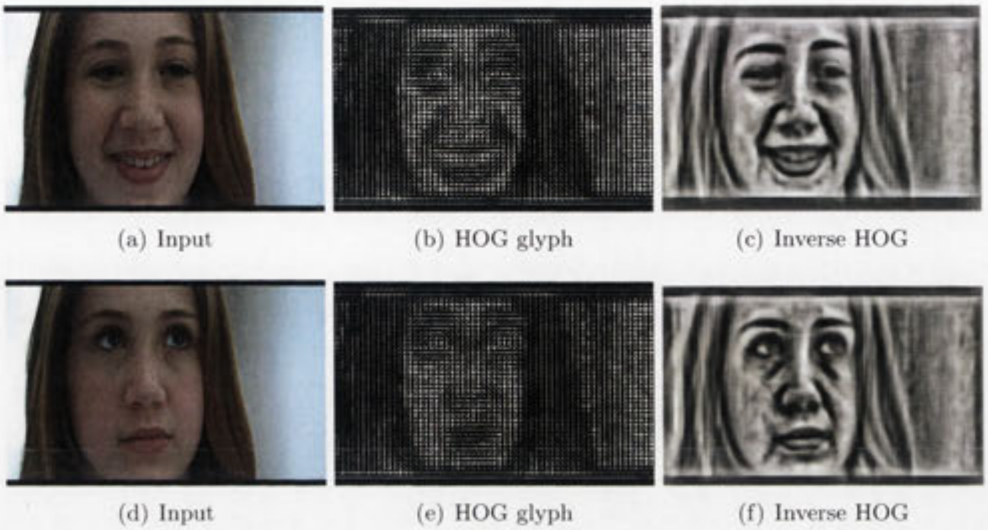


Figure 2.5: (a) & (b) are input frames from SFEW database Dhall et al. [2011c] describing *happy* and *neutral* expressions respectively. (b) & (e) are the HOG gylph representations of (a) & (b). (c) & (f) are the Inverse HOG Vondrick et al. [2013] visualisations for (a) & (b) respectively. (c) & (f) show what the classifiers learn.

Therefore for a centre pixel p of an image I and its neighbouring pixels N_i, a decimal value d is assigned to it.

$$d = \sum_{i=1}^{k} 2^{i-1} I(p, N_i) \tag{2.4}$$

$$where\ \ I(p, N_i) = \begin{cases} 1 & \text{if } c < N_i \\ 0 & \text{otherwise} \end{cases}$$

Further, histogram is created using the binary value d for the image.

Local Phase Quantisation (LPQ): In the literature, local phase quantisation (LPQ), an extension of LBP has been shown to perform better Ojansivu and Heikkilä [2008] than LBP and found to be invariant to blur and illumination to some extent. LPQ is based on computing the Short-Term Fourier transform (STFT) on a local image window. At each pixel, the local Fourier coefficients are computed for four frequency points. Then, the signs of the real and the imaginary part of the each coefficient is quantised using a binary scalar quantiser, for calculating the phase information. The resultant 8-bit binary coefficients are then represented as integers using binary coding. Further, the histogram is calculated in a similar manner to LBP. LPQ is computed block-wise and the block specific LPQ is concatenated to create a single vector.

Pyramid of Histogram of Gradients (PHOG): The histogram of oriented gradients (HOG) descriptor Dalal and Triggs [2005a] counts occurrences of gradient orientation in localised portions of an image and has been used extensively in computer vision. For computing HOG, Canny edge detector is applied to the region of interest. Then the face is divided into spatial grids at all pyramid levels. After this a 3 × 3 Sobel mask is applied to the edge contours for calculating the orientation gradients. Then the gradients of each grid are joined together at each pyramid level. There is an option for two orientation ranges: [0-180] and [0-360]. In LI et al. [2009], [0-360] orientation range perform better then [0-180]. Figure 2.5 describes the information captured by the HOG descriptor. The glyph based and inverse HOG Vondrick et al. [2013] representation are shown for frames corresponding to *happy* and *neutral* expressions, respectively. HOG is clearly able to extract discriminative information for FER. The pyramid of histogram of oriented gradients descriptor Bosch et al. [2007] is an extension of HOG, which has shown good performance in object recognition Bosch et al. [2007].

Local Binary Pattern-Three Orthogonal Planes (LBP-TOP): Local Binary Pattern - Three Orthogonal Planes (LBP-TOP) Zhao and Pietikainen [2007] is a popular descriptor in computer vision. It considers patterns in three orthogonal planes: XY, XT and YT, and concatenates the pattern co-occurrences in these three directions. The local binary pattern (LBP-TOP) descriptor assigns binary labels to pixels by thresholding the neighborhood pixels with the central value. Therefore for a centre pixel $\mathcal{O}_p$ of an orthogonal plane $\mathcal{O}$ and its neighboring pixels N_i, a decimal value d is assigned to it:

$$d = \sum_{\mathcal{O}}^{XY,XT,YT} \sum_{p} \sum_{i=1}^{k} 2^{i-1} I(\mathcal{O}_p, N_i) \tag{2.5}$$

More recently, in Jiang et al. [2011], authors extended LBP-TOP by replacing LBP by LPQ Ojansivu and Heikkilä [2008] and achieved better results then LBP-TOP. Further, Almaev and Valstar [2013] extended Local Gabor Binary Pattern (LGBP) Zhang et al. [2005b] in the TOP format to create Local Gabor Binary Pattern - Three Orthogonal Planes (LGBP-TOP). LGBP introduced for face recognition is constructed by applying Gabor filters to an input face at various orientations and scales. LBP is applied to the filter output and the LBPs obtained from different Gabor filters are concatenated to create a single vector. Almaev and Valstar [2013] show that the performance of LGBP-TOP is superior to that of LPQ-TOP and LBP-TOP for AU detection. However, features based on TOP format perform well when the location of the apex of an expression is known in a video clip. In real-world data Dhall et al. [2011d] it is difficult to find the apex. Moreover it cannot be assumed that the sample videos in an experiment will exhibit similar facial expression stages (*onset*, *apex*, *offset*).

Bag-of-Words (BoW): BoW based representations come from the domain of document processing. A BoW feature represents a document (image/video) as an unordered set of frequencies of words (In the context of this thesis words can be local features like LBP, HOG etc computed on patches.). LI et al. [2009] were the first to use a BoW for FER, they fused PHOG and BoW based histogram constructed from dictionary based on Scale Invariant Feature Transform (SIFT). Xu and Mordohai [2010] learn BoW dictionary from motion features computed from optical flow. Sikka et al. [2012] present a detailed analysis of different dictionary creation and word assignment strategies and their effect on FER performance. As the frames/patches are handled in an unordered manner in a BoW visual dictionary, the BoW framework is able to overcome the limitation of TOP class of features (i.e. dependency on information of apex of an expression). Even though BoW based vector can represent the frequency of presence of different stages of an expression, the temporal sequencing information is still missing. To overcome this, recently a data-driven technique Bettadapura et al. [2013] is proposed to explicitly encode the temporal information using n-grams. Bettadapura et al. [2013] perform experiments on human action recognition and activity analysis and show that adding temporal sequencing information based on their method increases the accuracy of the BoW based techniques.

2.4 Facial Expression Classification

Over the time various machine learning techniques for eg. SVM Valstar et al. [2005], Asthana et al. [2009b], Dhall et al. [2011b], Meng et al. [2011], Bayes Network Cohen et al. [2003], Boosted decision trees Day [2013], Hidden Markov Model Sebe et al. [2006], Meng and Bianchi-Berthouze [2011], Nearest Neighbour Dhall et al. [2011b], voting based classification Hayat et al. [2013] and deep neural networks Ebrahimi et al. [2013] etc. The choice of classifier can be based on the number of frames in a sample. Face expression analysis methods can be classified into another two broad classes: image based and video based. Below FER methods from both the categories are discussed with the type of classifier chosen.

Image based methods Shan et al. [2005], Pantic and Rothkrantz [2004] deal with single images which are usually the apex images. Whitehill et al. [2009] evaluate the existing descriptors and classification methods for smile detection in static images. They collected a large dataset GENKI consisting of 6300 images from the internet. For descriptors they evaluated gabor filters, box filters, edge orientation histograms and LBP. They also experimented the effect of manual marking vs automatic system based eye landmark marking on smile detection. For classification they used GentleBoost and SVM.

Tian et al. [2001] propose an automatic facial expression analysis system which classifies the expression based on the FACS coding. In Wang et al. [2004a], authors classify expressions into seven classes via boosting haar feature based Look-Up-Table types week classifier. They use VJ face detector to locate the face in the image and then compute haar features over the face. The experiments were conducted on the JAFFE database Lyons et al. [1998]. In Sohail and Bhattacharya [2007], authors use k-Nearest Neighbor (KNN) algorithm to classify expressions into categories. Distance features are extracted from the normalised landmark point positions of the face. Like Wang et al. [2004a], they also used the JAFFE facial expressions database.

LI et al. [2009], divided the face into four parts. SIFT descriptor is computed on interest points in the parts. Further BOW based vector quantization is applied to the calculated SIFT features. This decreases the dimension of the system. To encode the shape information authors computed PHOG Bosch et al. [2007] over the facial parts. They apply PCA for dimensionality reduction on both the features individually and then use SVM for classification. Later, decision fusion is calculated for generating the final result. They report robust performance on the Cohn-Kanade Kanade et al. [2000].

In real world, human facial expressions are dynamic in nature. They constitute an onset, one or more apex (peaks) and an offset. Studies Bassili [1979], Ambadar et al. [2005], have proven the effectiveness of video based facial expression analysis over the static analysis. In Bassili [1979], authors suggests that motions cues from a face image sequence are enough to recognise an expression even with minimal spatial information. Video based facial expression analysis in realistic environments is a challenging task. Several factors such as dynamic head movement, non-frontal pose, varied illumination, subtle facial actions and high variation in the temporal scale of facial actions contribute to its complexity.

In one of the early pioneering works by Yacoob and Davis [1994], facial parts are tracked and optical flow is calculated at high gradient values of the image sequence. Here the head were static. The direction of the flow is quantised to eight levels in order to have a mid-level representation for high level facial expression classification.

Black et al. [1998] use parametric models to extract parameters from facial features and use nearest neighbor classifier for FER. Different parametric models are used to differentiate between facial features relative to the head. It assumes that facial feature points are given beforehand. The rigid motion between two consecutive frames is calculated using a planar model. Here eyes, mouth and brows are excluded. Later the same process is repeated for facial features. Unlike Yacoob and Davis [1994], the mid-level representation is calculated by deducting the motion parameter estimation with a threshold value, this reduces the minute motions due to head pose. The classification accuracy for 70 image sequences from 40 speakers is 92%. Rosenblum et al. [1996] use the feature extraction and optical flow calculation similar to Black et al. [1998] but apply radial basis function neural network to classify facial expressions.

In another work, Kaiser and Wehrle [1992] tracked facial dots in image sequences for FER. But as the points move due to deformity in the face parts, it becomes difficult to track the points and hence, effects the accuracy. In Otsuka and Ohya [1996] Hidden

Markov Model(HMM) were used. In the first step, wavelet transform is applied to the image sequence. For expression recognition HMM are used on the feature vector. The main disadvantage of the method is that it cannot be used on a continuous sequence since the HMM are trained on image sequences starting with a neutral expression face. Secondly, it is designed for a single subject. Therefore in Otsuka and Ohya [1997], they propose the use of Fourier transform over the velocity vectors of image sequence as the feature vector for HMM and detect apex of an expression using squared sum of the feature vector as a parameter. Further in Lien [1998], three methods are suggested for AU detection using HMM.

In Bartlett et al. [2004] authors compared various machine learning classification techniques for real time FER from video sequences. They experimented with SVM and boosting methods on the CK dataset Kanade et al. [2000]. They experimented different SVM kernels and combination of features using boosting followed by classification using SVM. They also observed that feature selection using AdaBoost gave better performance then feature selection using Linear Discriminant Analysis (LDA) and PCA.

Pantic et al. [2002] propose an automatic AU detection method for profile face image sequences. Face tracking is dealt with as a segmentation problem where the profile face is the foreground. It finds largest connected component in HSV color space and then use the watershed segmentation algorithm to finally segment the face. Using contour based method, 20 points are extracted which are then used to identify AU using a rule based method.

Pantic et al. [2005a] propose two methods: first for automatic recognition of AU in video sequences and second for classifying AU coded expressions into learned emotion categories. The method is suitable for analysing temporal sequence pattern. Post face registration, temporal template called Motion History Image (MHI) Bobick and Davis [2001] are constructed from the image sequence. Then temporal rules are used to identify AUs in the CK Kanade et al. [2000] and the MMI Valstar et al. [2005] databases. The method achieves 90% recognition rate when detecting 27 AUs. Further the work is extended in Valstar and Pantic [2006], a wavelet based Gentleboost template is used to track 20 fiducial points, which further are used to construct spatio-temporal features. A subset of features is selected using AdaBoost and SVM is used for checking presence of AUs.

Yang et al. [2008] propose the computation of similarity features based on kernel methods. K-Means clustering is applied on the apex images and the cluster centres are used as reference for computing a L2 distance based similarity score. Further haar features based dynamic patterns are used to model the sequence dynamics. Further, Adaboost learning is performed on the temporal pattern for classification on the CK database. The system performs well on the dataset but has a prerequisite of knowledge of apex frames for computing the similarity features.

Yang et al. [2008] extended their work in Yang et al. [2009] and compute rankboost for expression classification and expression intensity estimation. Here the intensity estimation problem is converted into a ranking problem. For making Rankboost more robust L1 regularisation is integrated in the Rankboost algorithm. Further Yang et al. [2010] divide the face into local patches which roughly correspond to FACS AU locations. Haar features are calculated on the local patches and feature vector is computed via minimising error based on optimisation. This is integrated in a boosted framework for classification.

In their challenge entry Dhall et al. [2011b] in the First Emotion Recognition & Analysis (FERA) 2011 challenge Valstar et al. [2012], Dhall et al. [2011b] proposed a frame selection based FER method. Fiducial points are extracted using CLM and clustering is performed on the normalised shape points of all the frames of a video clip. The cluster centres are then chosen as the key-frames on which texture descriptors are computed. On analysing visually, the cluster centres corresponded to various stages of an expression i.e. onset-apex-offset. Two classifiers: SVM and Largest Margin Nearest Neighbour (LMNN) Weinberger and Saul [2009] are compared. The method performed well on the both task (subject independent and dependent emotion classification) in the FERA 2011 challenge.

The GEMEP−FERA database contains spontaneous data recorded in lab controlled conditions. State-of-the-art methods such as Yang et al. [2010], Dhall et al. [2011b], Sikka et al. [2012], Almaev and Valstar [2013] have been experimented on either posed or spontaneous facial expressions databases recorded in lab-controlled setting. Only recently, methods have been proposed for FER in real-world conditions. Liu et al. [2012], proposed a transfer learning based method, where unlabelled facial expression data is used along with the SFEW database. Various methods for emotion recognition on AFEW database are proposed in the EmotiW challenge Dhall et al. [2013]. The methods are discussed in Section 3.8.2.

2.4.1 Expression Labelling

The classification method chosen also depends on the type of expression labelling. Classic FER methods majorally divide facial expressions into a set of defined classes. These method follow the early work of Darwin Darwin [1998] and further proposed by Ekman [1993]. In Ekman [1993], facial expressions have been divided into six universal categories (*joy*, *surprise*, *anger*, *sadness*, *fear* and *disgust*). In real scenarios these basic expressions occur relatively infrequently. Facial expression change is displayed by humans by subtle movements in one or more parts of the face. The Facial Action Coding System (FACS) Ekman and Friesen [1978] is an expression coding standard developed by behavioral scientists. It decomposes the human faces into 46 component movements, which are referred to as Action Unit (AU). It has been used extensively in both automatic and manual facial expression analysis research. Recent databases McKeown et al. [2010] use continuous labelling in the *Valence*, *Arousal* and *Dominance* scales. Detailed survey discussing continuous emotion labelling can be found in Gunes and Pantic [2010] and Gunes et al. [2011]. It is argued that using a continuous representation it is possible to represent complicated expression and emotions. However, it is non-trivial to label frames as it is a frame by frame labelling. For labelling *Dominance*, the context is very important, for example: screaming can be both due to pain or pleasure. With a database representing real-world conditions, labelling becomes trickier. In this thesis the discrete expression labels are used for FER in challenging conditions, for their simplicity.

2.5 Facial Expression Databases

Over the past decade, many databases have been published. One of the earliest is the widely used Cohn-Kanade database Kanade et al. [2000]. It contains 97 subjects, who posed in a lab controlled condition for the six universal expressions and the neutral expression. Its extension, CK+ Lucey et al. [2010], contains 123 subjects but the new videos were shot in a similar environment. The CMU Multi-PIE Gross et al. [2008a] data corpus is a new facial expression database which contains 337 subjects imaged under 15 view points and in 19 illumination conditions. In total, it has around 375,000+ images with 6 different facial expressions. The method of construction of these databases is purely manual where the subjects posed sequentially. The MMI Pantic et al. [2005b] database is a searchable temporal database with 75 subjects. All of these are posed lab-controlled environment databases. The subjects display various acted (not spontaneous) expressions. The recording environment is nowhere close to real-world conditions and the data collection and labelling process is manually.

The RU-FACS database Bartlett et al. [2006] collected at the Rutgers university is a FACS-coded temporal database exhibiting spontaneous facial expressions, but it is proprietary and unavailable to other researchers. The Belfast database Douglas-Cowie et al. [2000] consists of a combination of studio recordings and TV programme grabs labelled with particular expressions. The number of TV clips in this database is small.

The *JAFFE database* Lyons et al. [1998] is one of the earliest static facial expression datasets. It contains 219 images of 10 Japanese females. However, it has a limited number of samples, subjects and has been created in a lab controlled environment. In one of the first experiments on close-to-real data, Paleari et al. [2010] proposed a bimodal, audio-video features based system. The database has been constructed from TV programs. However, the size of database is fairly small, with 107 clips only. Recently, Richter et al. [2012], proposed an image based FER dataset based on Google

Database	**Construction Process**	**Environment**	**Age Range**	**Illumination**	**Occlusion**	**Subjects**	**Searchable**	**Subject Details**	**Multiple Subjects**
AFEW Dhall et al. [2012a]	S-A	CTR	1-70	CTN	Yes	330	Yes	Yes	Yes
Belfast Douglas-Cowie et al. [2000]	M	CTR/Lab	?	C	Yes	100	No	No	No
CK Lucey et al. [2010]	M	Lab	18-50	C	No	97	No	No	No
CK+ Lucey et al. [2010]	M	Lab	18-50	C	No	123	No	No	No
F.TUM Wallhoff [2006]	M	Lab	?	C	No	18	No	No	No
GEMEP Bänziger and Scherer [2010]	M	Lab	?	C	Yes	10	No	No	No
GENKI Whitehill et al. [2009]	S-A	CTR	?	CTR	Yes	10	No	No	No
HAPPEI Dhall et al. [2012b]	S-A	CTR	?	C	Yes	NA	Yes	No	Yes
M-PIE Gross et al. [2008a]	M	Lab	27.9	C	Yes	337	No	No	No
MMI Pantic et al. [2005b]	M	Lab	19-62	C	Yes	29	Yes	No	No
Paleari Palcari et al. [2010]	M	CTR	-	CTN	Yes	-	No	No	No
PIE Sim et al. [2003]	M	Lab	?	C	Yes	68	No	No	No
RU-FACS Bartlett et al. [2006]	M	Lab	18-30	C	Yes	100	No	No	No
Semaine McKeown et al. [2010]	M	Lab	?	C	Yes	75	Yes	No	No
USTC-NVIE Wang et al. [2010b]	M	Lab	17-31	C	Yes	215	No	No	No
UT-Dallas O'Toole et al. [2005]	M	Lab	18-25	C	Yes	284	No	No	No
UVa-Nemo Dibeklioğlu et al. [2012]	M	Lab	8-76	C	Yes	400	No	No	No
VAM Grimm et al. [2008]	M	CTR	?	C	Yes	20	No	No	No
Web Image Yu et al. [2013]	S-A	CTR	?	C	Yes	NA	No	No	No

Table 2.1: Comparison of facial expression databases. C = Controlled, CTN = Close To Natural and CTR = Close To real. M = Manual and S-A = Semi-automatic

Image search. They manually filter from a set of 80,000 images and labelled 4761 images.

The Semaine database McKeown et al. [2010] has been collected while subject converse with an artificial listener. The authors have done multiple emotion labelling types: discrete and continuous. It has spontaneous emotions in lab-controlled scenarios. Compared to the manual method used to construct and annotate these databases Kanade et al. [2000], Lucey et al. [2010], Bartlett et al. [2006], Douglas-Cowie et al. [2000], McKeown et al. [2010], a semi-automatic process based on movie subtitles has been used in the construction process of the Acted Facial Expressions In The Wild (AFEW) Dhall et al. [2012a] (Please refer to the details in Section 3.2). A meta-data schema based on XML is developed, this makes the database easily searchable and accessible from a variety of languages and platforms. In contrast, CK, CK+, Multi-PIE, RU-FACS and Belfast need to be searched manually. The MMI and Semaine databases have a searchable web interface and annotation.

AFEW database is similar in spirit to the Labeled Faces in the Wild (LFW) database Huang et al. [2007] and the Hollywood Human Actions (*HOHA*) dataset Laptev et al. [2008]. These contain varied pose, illumination, age, gender and occlusion. However, LFW is a *static* face recognition database created from single face images found on the WWW specifically for face recognition and HOHA is an action recognition database created from movies.

Recently, Yu et al. [2013] proposed an image based facial expressions dataset collected from Google Images using keywords related to facial expressions labels (*'anger'*, *'disgust'*, *'fear'*, *'happy'*, *'neutral'*, *'sad'* and *'surprise'*). A face detector is applied on the downloaded images and images with no faces are rejected. Further, a SVM based FER classifier is learnt on a small set of labelled images. The classifier returns the class probabilities and images with least probability are chosen to be manually labelled. The newly manually labelled images are added to the already existing labelled images set and the FER based SVM classifier is retrained. This active learning method assists in fast labelling of the images. As a face detector is used to filter out the downloaded images which may not have a face, images with faces with non-frontal head pose can be easily missed. This way challenging real-world data depicting non-frontal faces is difficult to collect.

Table 2.1 shows a detailed comparison of facial expression databases. The comparison is based on the following properties: '*Construction Process*' - Semi-automatic or Manual; '*Environment*' - 'Close to Real' or 'lab'; 'Age Range'; '*Illumination*' - 'Controlled' or 'Close to Natural'; 'Occlusion'; 'Subjects'; 'Searchable'; 'Subject Details'; 'Multiple Subjects';

2.6 Head Pose Normalisation

During conversation human beings tend to move their head which causes head motion. Handling pose is a classical problem in facial analysis. Traditionally, pose-affected face analysis problems (recognition, expression analysis, etc.) can be broadly divided into two categories: a) top-down and b) bottom-up. In the former, the head pose is estimated first and then pose-specific classification models are used for inference Hu et al. [2008], Moore and Bowden [2011]. In the latter, the head pose is normalised first and then a frontal pose-specific classification model is used Rudovic et al. [2010a,b], Asthana et al. [2011].

In one of the first works, Blanz and Vetter [1999] proposed *3D Morphable Models* for constructing 3D facial points from a single image. Asthana et al. [2011] proposed a 3D HPN method using *view-based* AAM Cootes et al. [2002] and 3D model warping. The biggest drawback of these approaches is that the 3D models are computationally very expensive. 2D deformable model based approaches Cootes et al. [2002] overcome the computational problem. Facial landmark points are extracted using a 2D AAM and frontal pose points are computed using a linear regression model. However, such approaches only deal with expressionless faces.

Asthana et al. [2009a] proposed a regression-based method for generating faces at various poses. This method generates faces at different poses by learning a mapping from frontal to non-frontal facial landmark points, leading to an augmented labeled face dataset for further learning AAMs. Rudovic et al. [2010b] proposed a *Gaussian Process Regression (GPR)* Rasmussen [2006] based multi-view facial expression recogniser. Given a non-frontal pose and facial landmark points, the authors compared various regression methods for HPN. The pose normalised landmark points were further used to train a SVM based FER model.

In an interesting work, Rudovic and Pantic [2011] extended their earlier work Rudovic et al. [2010b] to propose a Shape Constrained Gaussian Process (SC-GP) regression based HPN method. The authors argue that without any explicit face shape constraints, the normalised points may not adhere to the face shape. An explicit shape constraint is added by learning an ASM. However, ASM have limitations in the case of subject independent experiments. ASM require accurate initialisation, which is non-trivial for non-frontal faces.

A top-down approach was proposed by Hu et al. [2008], which learns view-specific facial expression recognition (FER) models. During the inference, a head pose estimator is used to select the view-specific FER model. Moore and Bowden [2011] presented an extensive comparison of texture descriptors for multi-view FER and experimented on the BU-3DFE Yin et al. [2006] and CMU Multi-PIE Gross et al. [2008b] databases. The experiments in Hu et al. [2008] Moore and Bowden [2011] Rudovic et al. [2010b] Rudovic and Pantic [2011] are performed on lab-controlled databases.

A drawback of approaches such as Asthana et al. [2009a], Rudovic et al. [2010b] is that they either are AAM-based or require manual labels during inference. As Asthana et al. [2009a], Rudovic et al. [2010b] are based on points based regression, they require accurate facial landmark points for normalisation. This is in contrast to our parts-based approach, which does not require facial landmark points. Moore and Bowden Moore and Bowden [2011] require head pose information for selecting a pose-specific FER model. These approaches assume accurate results from the face detection and head pose estimation steps, which are both non-trivial tasks in real-world images. To remove the prerequisite of head pose estimation, Hu et al. [2008] proposed to learn separate FER models for each pose. However, this complicates the problem as increasing the number of non-frontal poses would increase the number of models to be learnt.

Another limitation of prior work Rudovic et al. [2010a,b] is that they use fewer landmark points (39 in Rudovic et al. [2010a,b]). This can be a problem, for example, in FACS-based facial action unit recognition such as AU20 (with no chin information). In contrast, our method generates a detailed 68-point annotation.

2.7 Analysis of Group of People

The analysis of a group of people in an image or a video has recently received a lot of attention in the domain of computer vision. Methods can be broadly divided into two categories: a) **bottom-up** methods: when the subject's attributes are used to infer information at the group level Ge et al. [2012], Hernandez et al. [2012], Dhall et al. [2012a]; b) **top-down** methods: when the group/sub-group information is used as a prior for inference of subject level attributes.

2.7.1 Bottom-up Techniques

Tracking groups of people in a crowded has been of particular interest lately Ge et al. [2012]. Based on trajectories constructed from the movement of people, Ge et al. [2012] propose a hierarchical clustering algorithm which detects sub-groups in crowd video clips. In an interesting experiment, Hernandez et al. [2012] installed cameras at four locations on the MIT campus and tried to estimate the mood of people looking into the camera and compute a mood map for the campus. They used the Shore framework Küblbeck and Ernst [2006] for face analysis. The Shore framework detects multiple faces in a scene in real-time. The framework also generates attributes such as age, gender and pose. In Hernandez et al. [2012], the scene level happiness is an averaging of an individual person's smile. However, in reality group mood is not an averaging model Kelly and Barsade [2001]. There are attributes which affect the perception of a group's mood and the mood of the group itself.

In another interesting bottom-up method, Murillo et al. [2012] proposed group classification for recognising urban tribes (a group of people part of a common activity). They used low-level features such as color histogram and high-level features as attributes such as age, gender, hair and hat etc (using Face.com API) to learn BoW based classifier. To add the group context, a histogram describing the distance between two faces and number of overlapping bounding boxes is computed. Fourteen classes depicting various groups such as 'informal club', 'beach party' and 'hipsters' etc. are used. The experiments show that a combination of attributes can be used to describe a type of group.

2.7.2 Top-down Techniques

As an interesting top-down approach, Gallagher and Chen [2009] proposed contextual features based on the group structure for computing the age and gender of individuals. The global attributes described in chapter 5 are similar to Gallagher and Chen [2009]'s contextual features of social context. However, the problem which Gallagher and Chen [2009] are trying to solve is inverse to the problem of inferring the mood of a group of people in an image, which is discussed in this thesis later (Chapter 6). Gallagher and Chen [2009] compute contextual features based on the structure of the group for better age and gender recognition. Their experiments on images obtained from the web, show an impressive increase in performance when the group context is used. In another top-down approach, Wang et al. [2010a] model the social relationship between people standing together in a group for aiding recognition. The social relationships is inferred in unseen images by learning them from weakly labelled images. They learn a graphical model based on social-relationships such as '*father-child*', '*mother-child*' etc and social-relationship features such as relative height, height difference and face ratio etc. In another interesting work, Lee and Grauman [2011] propose a face discovery method based on exploring social features such as on social event images.

In object detection and recognition work by Torralba and Sinha [2001] context information of the scene and its relationship with the objects is described. Moreover, works such that by Parikh et al. [2008] acknowledge the benefit of using global spatial constraint for scene analysis. In face recognition Stone et al. [2008], social context is being employed to model relationship between people such as between friends on Facebook. The relationship between connected people is modelled using a Conditional Random Field (CRF) Manyam et al. [2011].

Recently, Fiss et al. [2011] proposed a framework for selecting candid images from a video of a single person. They conducted a physiological study, 150 subjects were shown images of a person. They were asked to rate the attractiveness of the images and mention attributes which influenced their decision. They also asked professional photographers to label the images. Further, a regression model was learnt based on various attributes such as eye blink, clarity of face, face pose etc. The limitation of this approach is that the samples contain a single subject only.

Zhang et al. [2009] proposed an affect based movie/video clip browsing system. They learnt two regression models, which predict the valence and arousal values, describing the affect of movies. The regression models learnt on ensemble of audio-video features such as motion, shot switch rate, frame brightness, pitch, bandwidth, roll off, spectral flux etc. However, they did not use the expression information for individual or groups in scenes.

For analysing a single subject's happiness/smile detection there is not a lot of work in literature. One prominent work is by Whitehill et al. [2009], they proposed a new image based database labelled for smiling and non-smiling images and evaluated several state of the art methods for smile detection. But in all these works the faces are considered independent of each other. For computing the contribution of each subject. There are several factors which effect group level mood analysis. Local factors (individual subject level): age, gender, face visibility, face pose, eye blink etc. Global: where people are standing, with whom people are standing. In Chapter 6, the focus is on face visibility, smile intensity and relative face size and relative face distance. Further, labelled data of image containing groups of people is required which is collected from Flickr.

2.8 Applications of Facial Expression Recognition

FER has been applied to various problems, some are discussed below:

1. **Depression analysis:** Facial dynamic analysis inspired from FER method Zhao and Pietikainen [2007] is used for analysis facial activity in depressed subjects Joshi et al. [2012]. LBP-TOP is computed on aligned faces in a piece-wise manner and a BoW based dictionary is learnt. This method was found to be discriminative for depressed vs healthy controls. Joshi et al. [2012] confirmed the hypothesis of Ellgring [2008] that during severe depressive episode psycho-motor activity decreases i.e. the intensity and occurrence of facial expression are dampened.

2. **Pain classification:** Facial expression analysis in the form of AU detection has been used for pain classification on the McMasters pain dataset Lucey et al. [2011]. Subject dependent AAM are used to localise the fiducial points and AU detection is performed. Pain classification is done based on presence of AUs. Sikka et al. [2013a] proposed a pain classification method based on video clip segmentation. BoW based feature are computed on each sub-sequence. Classification is performed using Multiple Instance Learning Zhang et al. [2005a] algorithm.

3. **Movie affect classification:** Joho et al. [2011] observed subjects watching movies clips. PBVD based tracker Sebe et al. [2007] is used to compute motion features. The affect induced in subjects is used for generating video highlights.

4. **Similar Facial Expressions:** Kemelmacher-Shlizerman et al. [2010] proposed a method for finding similar facial expressions. LBP is computed on aligned faces and Chi-Square distance metric is used to find the most similar facial expression from a dictionary of faces.

5. **Human Robot Interface:** For better human robot interaction, it is important for the robot to understand the user's state of mind. There have been several attempts at adding facial expression analysis in robots Wimmer et al. [2008].

6. **Assessing affect of advertisements:** Recent startups like Affectiva, Emotient, Realeyes are offering solution for analysing the affect induced in users, while watching advertisements. Users watch an advertisement over the web on their computers and a webcam records their facial dynamics. Data analytic is later performed on the recorded data and FER is used to infer the affect of the advertisement on large demographics. This is a major example of FER application in a real world problem.

7. **Image management and viewing:** Dhall et al. [2010] propose the use of facial expression analysis for creating expression based image albums. This process is detailed in Chapter 5. This enable easy summarisation of events. Dhall et al. [2012b], use the mood of group to perform event summarisation, the details are in Section 6.10.5. Moreover, a group's mood is used to recommend candid photo from a set of images captured in burst shot mode (Section 6.10.6).

Chapter 3

Facial Expressions In The Wild Database

Image analysis is inherently data-driven. In the domain of automatic human face analysis, realistic data plays a very important role. Much progress has been made in the fields of face recognition Huang et al. [2007] and human activity recognition Laptev et al. [2008] in the past years due to the availability of realistic databases as well as robust representation and classification techniques. However, in the case of human facial expression analysis, there is a lack of databases representing real-world scenarios. Even though there are a number of popular facial expression databases (Lyons et al. [1998], Pantic et al. [2005b] and Lucey et al. [2010]), the majority of these have been recorded in tightly controlled laboratory environments, where the subjects were asked to 'act' certain expressions. These 'lab scenarios' are in no way a (close to) true representation of the real world. Ideally, a dataset[1] should represent spontaneous facial expressions in challenging real-world environments. However, as anyone in the face analysis community will attest to, such datasets are extremely difficult to obtain.

Generally, the traditional datasets are constructed by recording subjects while they pose for a particular expression Lucey et al. [2010] or perform a task (for example filling a form Hoque and Picard [2011]). This construction process is time consuming and error prone. The task of dataset construction becomes even more difficult when trying to capture different scenarios (out of the lab) representing real-world scenarios (as illustrated in Figure 1.1). As an important stepping stone on this path for solving these two challenges (time-consuming construction process and need for capturing real-world scenarios), a semi-automatic method is proposed for collecting data representing close-to-real-world scenarios. Movies represent real-world conditions and are chosen as the data source for constructing the database in this chapter. A video clip recommender system based on subtitle parsing is proposed. The labellers do not have to scan the

[1] The word database and dataset are used interchangeably through out the chapter.

full movie manually but use the recommender system, which suggests only those video clips, which have a high probability of a subject showing a meaningful expression. This method helps in collecting and annotating large amounts of data quickly. Based on the availability of detailed information regarding the movies and their contents on the WWW, the labellers annotate the video clips with dense information about the subjects. The work presented in this chapter has been published in Dhall et al. [2011c], Dhall et al. [2012a] and Dhall et al. [2013].

The proposed dataset is called *Acted Facial Expressions In The Wild* **(AFEW)** Dhall et al. [2012a]. From AFEW, a static dataset called *Static Facial Expressions In The Wild* **(SFEW)** Dhall et al. [2011c] is extracted. The remainder of this chapter is structured as follows. AFEW and SFEW are compared with earlier databases in Section 2.5. Section 3.1 discusses the contributions of the proposed databases. The subtitle extraction process is described in Section 3.2.1. The recommender system is detailed in Section 3.2.2. The meta-data annotations are discussed in Section 3.2.3. Section 3.3 defines the SFEW database. The experimental protocols and database versions are discussed in Sections 3.4 and 3.5. Section 3.6 discusses the quantitative comparison of AFEW and SFEW with existing databases. AFEW and SFEW baselines are presented in Section 3.7. The EmotiW challenge and its baselines are described in Section 3.8. The chapter findings are concluded in Section 3.9.

3.1 Contributions of the database

The databases AFEW and SFEW have the following novelties:

- AFEW is the largest temporal facial expression data corpus available to the research community, consisting of short video clips of facial expressions in close to real-world environments.

- The subjects have a wide age range (1-70yr), which makes the databases very generic in terms of age, unlike other facial expression databases (See Section 2.5 for detailed comparison). The databases have a large number of clips depicting children and teenagers, which can be used to study facial expressions in younger subjects. The datasets can also be used for both static and temporal facial age research.

- AFEW is currently the only facial expression database, which has multiple labelled subjects in the same frame. This enables an interesting study on the 'theme/group' expression of a scene with multiple subjects, which may or may not have the same expression individually at a given time.

- The database exhibits 'close-to-real' illumination conditions. The clips stem from scenes with indoor, night-time and outdoor natural illumination. While it is clear that movie studios use controlled illumination conditions even in outdoor settings, it is intuitive that these are closer to natural conditions than lab-controlled conditions and, therefore, valuable for facial expression research. The diverse nature of the illumination conditions in the dataset makes it useful for not just facial expression analysis but potentially also for face recognition, face alignment, age analysis and action recognition.

- The movies in AFEW, have been chosen to cover a large set of actors. Many actors appear in multiple movies in the dataset, which will enable to study how their expressions have evolved over time, whether they differ for different genres and characters, etc.

- The design of the database schema is based on XML. This enables further information about the data and its subjects to be added easily at any stage without changing the video clips. This means that detailed annotations with attributes about the subjects and the scene are possible and can be extended in the future.

- The database download website: *http://cs.anu.edu.au/few* contains information regarding the experiment protocols and train and test splits for both temporal and static FER experiments.

3.2 Database Creation

In this section, the details of the database construction are discussed. A semi-automatic approach is followed during the creation of the database. The process is divided into two parts. In the first step, the subtitles are extracted and parsed in the recommender system. In the second step, the labeler annotates the recommended clips based on the information available on the internet. Figure 3.1, describes the steps of database creation.

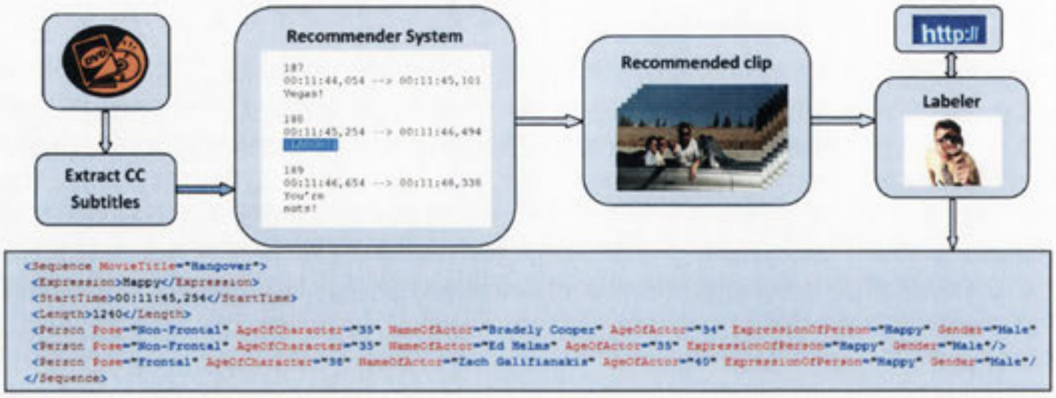
3.2.1 Subtitle Extraction

In its current form, seventy-four movie DVDs [2] have been analysed for AFEW. *Subtitles for Deaf and Hearing impaired (SDH)* and *Closed Caption (CC)* subtitles are extracted from the DVDs, these types of subtitles contain information about the audio and non-audio context such as emotions, information about the actors and scene, e.g. '[CHEERING]', '[SHOUTS]', '[SURPRISED]', etc. The subtitles are extracted from the movies using the VSRip tool[3]. For the movies where VSRip can not extract subtitles, SDH subtitles are downloaded from the web[4]. The extracted subtitle images

[2] See Appendix A for the list of movies in AFEW.

[3] http://www.videohelp.com/tools/VSRip

[4] http://subscene.com/ and http://www.opensubtitles.org/



```
187
00:11:44,054 --> 00:11:45,103
Vegas!

188
00:11:45,254 --> 00:11:46,494
[illegible]

189
00:11:46,654 --> 00:11:48,338
You're
nuts!
```

```
<Sequence MovieTitle="Hangover">
 <Expression>Happy</Expression>
 <StartTime>00:11:45,254</StartTime>
 <Length>1240</Length>
 <Person Pose="Non-Frontal" AgeOfCharacter="35" NameOfActor="Bradely Cooper" AgeOfActor="34" ExpressionOfPerson="Happy" Gender="Male"
 <Person Pose="Non-Frontal" AgeOfCharacter="35" NameOfActor="Ed Helms" AgeOfActor="35" ExpressionOfPerson="Happy" Gender="Male"/>
 <Person Pose="Frontal" AgeOfCharacter="38" NameOfActor="Zach Galifianakis" AgeOfActor="40" ExpressionOfPerson="Happy" Gender="Male"/
</Sequence>
```

Figure 3.1: From the left, a subtitle is extracted from the DVD and then parsed by the recommender system. In the example here, when the subtitle contains the keyword 'LAUGH', the corresponding clip is played by the tool. The human labelers can chose or reject the recommended video clip. The human labeler then annotates the subjects in the scene, using a GUI tool, based on the information about the subjects in the clip available on the WWW. The resulting annotation is stored in the XML schema shown at the bottom of the diagram. Note the structure of the information about a movie scene containing multiple subjects. The frame in the figure is from the movie 'The Hangover'.

are parsed using Optical Character Recognition (OCR) and converted into *.srt* subtitle format using the *Subtitle edit tool* [5]. The *.srt* format contains the start time, end time and text content with milliseconds accuracy.

3.2.2 Video Recommender System

Once the subtitles are extracted, the subtitles are parsed and searched for expression related keywords (for example: [HAPPY], [SAD], [SURPRISED], [SHOUTS], [CRIES], [GROANS], [CHEERS], [LAUGHS], [SOBS], [SILENCE], [ANGRY], [WEEPING], [SORROW], [DISAPPOINT], [AMAZED] etc.). If found, the system recommends video clips to the labeler. The start and end time of the clip is extracted from the subtitle information. The system plays the video clips sequentially and the labeller enters information about the clip and its characters / actors from the WWW. In the case of clips with multiple actors, the sequence of labeling is based on two criteria. For actors appearing in the same frame, the order of annotation is left to right. If the actors appear at different time stamps, then it is in the order of appearance. The dominating expression in the video is labeled as the 'theme' expression. The labeling is then stored in an XML metadata schema. Finally, the labeler enters the age of the character or, where this information was unavailable, estimated the age.

[5] www.nikse.dk/se

For AFEW 2.0 Dhall et al. [2012a], in total fifty-four DVDs (the various AFEW versions are discussed in Section 3.5) contain 77666 individual subtitles. Out of these, 10327 clips corresponding to subtitles containing expressive keywords are suggested by the recommender system. The labelers chose 1426 clips from these on the basis of criteria such as the visible presence of subjects, at least some part of the face being visible, and the display of meaningful expressions. Subtitles are manually created by humans and can contain errors. This may lead to a situation where the recommender system may suggest an erroneous clip. However, such a recommendation can be rejected by the labelers. The labellers annotated the clips based on the video, audio and subtitle information, so that they could make a more informed decision. The proposed recommender system can be used to easily add more clips to the database and scale it up to web scale.

3.2.3 Database Annotations

The database contains metadata about the video clips in an XML-based schema, which enables efficient data handling and updating. The human labelers densely annotated the subjects in the clips.

- *Expression* - This specifies the *theme expression* conveyed by the scene. The expressions were divided into six expression classes (*angry*, *disgust*, *fear*, *happy*, *neutral*, *sad* and *surprise*) plus neutral. The default value is based on the search keyword found in the subtitle text, for example for 'smile' and 'cheer' it is *Happiness*. The human observer can change it based on their observation of the audio and scene of the clip.

- *StartTime* - This denotes the start timestamp of the clip in the movie DVD and is in the hh:mm:ss,zzz format.

- *Length* - Duration of the clip in milliseconds.

- *Person* - This contains various attributes describing the actor / character in the scene.

  - *Pose* - This denotes the head pose based on the labeler's observation. In the current version, the head pose is classified manually as frontal or non-frontal.

  - *AgeOfCharacter* - Where the age of the character is available from the WWW, this information is used. Frequently, this is only the case for the characters of the lead actors. Otherwise, the labeller estimates the age based on his/her perception.

  - *NameOfActor* - Real name of the actor.

Attribute	Description
Length of sequences	300-5400 ms
No. of sequences	1832 (AFEW 3.0) EmotiW: 1088
No. of annotators	2
Expression classes	Angry, Disgust, Fear, Happy, Neutral, Sad and Surprise
Total No. of expressions (some sequences have multiple subjects	2153 (AFEW 3.0) EmotiW: 1088
Video format	AVI
Audio format	WAV

Table 3.1: Attributes of the AFEW database Dhall et al. [2013].

- *AgeOfActor* - Real age of the actor. The labeler extracted the information from www.imdb.com. In a very few cases, the age information is missing, in this case the labeller estimates it, based on his/her perception.
- *ExpressionOfPerson* - This denotes the expression class of the character as labelled by the human observer. This may be different from the 'Expression' tag as there may be multiple people in the frame showing different expressions with respect to each other and the scene/theme.
- *Gender* - Gender of the actor.

This XML-based metadata schema has three advantages. First, this information can be used to introduce context, for example, it is interesting to see the effect of attributes such as age and gender on FER systems. Secondly, it is easy to use and search using any standard programming language on any platform that supports XML. Secondly, the structure makes it simple to add new attributes about the video clips, such as pose of the person in degrees and scene information, in the future, while keeping the existing data and ensuring that the already existing tools can take advantage of this information with minimal changes. Details of the database are in the Table 3.1.

3.3 Static Facial Expressions in the Wild

Static facial expression analysis databases such as Multi-PIE and JAFFE Lyons et al. [1998] are lab-recorded databases in tightly controlled environments. For creating a static image database that represents the real world more closely, frames were extracted from AFEW. This static database was named *Static Facial Expressions in the Wild (SFEW)*. The *Strictly Person Independent* version of SFEW is described in Dhall et al. [2011c] and is posted as a challenge on the BEFIT website (http://fipa.cs.kit.edu/511.php). The three versions of SFEW, which are based on the level of subject dependency for evaluating facial expression recognition performance of systems in different scenarios are described in Section 3.4.

Protocol	**AFEW/SFEW (Train-Test sets contain:)**
Strictly Person Specific (SPS)	same single subject
Partial Person Independent (PPI)	a mix of common and different subjects
Strictly Person Independent (SPI)	different subjects Dhall et al. [2011c]

Table 3.2: Experimentation protocol scenarios for SFEW and AFEW.

3.4 Experimental Protocol

Over the years, many FER methods have been proposed based on experiments on various databases following different protocols, making it difficult to compare the results fairly. Therefore, a strict experimentation protocol is defined for the database. The different protocols are based on the level of person dependency present in the sets. AFEW and SFEW data is divided into two partitions: Set1 and Set2. The two sets are used alternatively for train and test. Experiments on AFEW and SFEW are divided into three categories:

1. *SPS* - Strictly Person Specific,
2. *PPI* - Partial Person Independent, and
3. *SPI* - Strictly Person Independent.

SPS describes scenario when there is a same single subject in the two sets. This protocol finds importance for FER methods which are intended to be person specific. PPI defines scenario where some subjects appear in both the sets. In SPI protocol, both the sets have mutually exclusive subjects. SPI protocol represents the data on the web, where the chances are less that the subject in the train will ever appear in the test set. Table 3.2 explains the three protocols for both the databases. The protocols represent different practical scenarios.

3.5 Database Versions

The first version of AFEW contained 979 video clips. Later, the number of video clips was increased based on the process defined in Section 3.2. AFEW 2.0 Dhall et al. [2012a], contains 1426 video clips. AFEW 3.0 forms the base of the first Emotion Recognition in the Wild (EmotiW) grand challenge at the ACM International Conference on MultiModal Interaction (ICMI) 2013. AFEW 3.0 has been extracted from seventy-four movies and contains 1832 video clips in total. EmotiW is a multimodal emotion recognition challenge where researchers are competing by testing their state-of-art method on the AFEW 3.0 database. Audio information is extracted along with the video clip using the XML based meta-data created during the AFEW construction

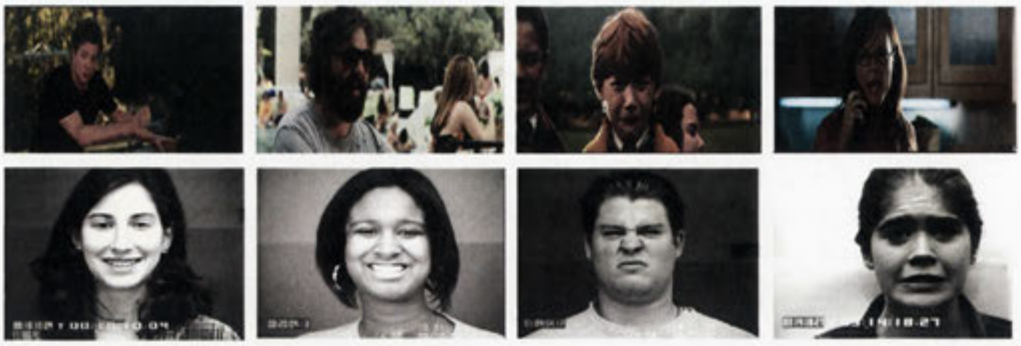


Figure 3.2: Row 1 shows frames from the AFEW database and Row 2 shows frames from the CK+.

process (Section 3.2). The multimodal baseline details for EmotiW is discussed in Section 3.8. SFEW contains a total of 700 images. SFEW_SPS contains 76 images of the actor Daniel Radcliffe for five expression classes (*anger*, *fear*, *happiness*, *neutral* and *surprise*). SFEW_PPI contains 700 images and SFEW_SPI contains 700 images in two sets. The images are chosen which clearly show the facial expression and the face in the frame is detectable.

3.6 Quantitative Comparison of Databases

Experimentally, AFEW is compared to the CK+ database, which was introduced by Lucey et al. [2010] as an extension of the CK database. A basic facial expression comprises of various temporal dynamic stages: *onset*, *apex* and *offset* stage. In CK+, all videos follow the temporal dynamic sequence: *neutral* → *onset* → *apex*, which is not a true reflection of how expressions are displayed in real-world situations as the data about the offset phase is missing. Earlier systems trained on existing databases such as CK+, have learnt on the above mentioned stages. It is intuitive that the complete data containing all the temporal stages may not be always available in real-world settings. For example, a person entering a scene may already be happy and close to the highest intensity of happiness (onset). In AFEW, this is not fixed due to its close-to-natural settings. For extracting the face, the VJ face detector is used over the CK+ sequences. For the comparison, six common classes from both the AFEW and CK+ databases: (*anger*, *fear*, *disgust*, *happiness*, *sadness* and *surprise*) are used.

SFEW is compared with the JAFFE and Multi-PIE databases in two experiments: (1) a comparison of SFEW, JAFFE and Multi-PIE on the basis of four common expression classes (*disgust*, *neutral*, *happiness* and *surprise*) and (2) a comparison of SFEW and JAFFE on all seven expression classes. Various descriptors are computed for the comparison.

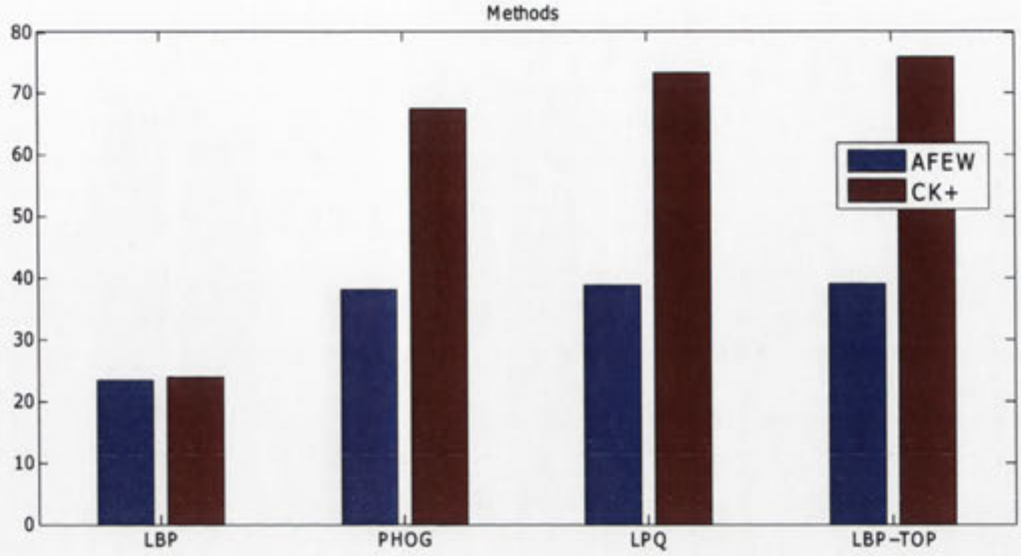


Figure 3.3: The graph shows the performance of LBP, PHOG, LPQ and LBP-TOP on the CK+ and AFEW databases.

The cropped faces detected using the VJ face detector are divided into 4 × 4 blocks for computing LBP Huang et al. [2011], LPQ Huang et al. [2011] and PHOG Bosch et al. [2007]. For LBP and LPQ, the neighbourhood size is set to 8. For PHOG the parameters are: bin length = 8, pyramid levels L=2 and angle range = [0,360]. PCA is applied on the extracted features and 98% of the variance is kept. For classification, a SVM model based on RBF kernel is trained. A five-fold cross validation is performed for selecting the parameters. For AFEW, the static descriptors are concatenated. LBP-TOP performs best out of all the methods. For all the methods, the overall expression classification accuracy is much higher for CK+ (see Figure 3.3).

For SFEW's ***four expression class*** experiment, the classification accuracy on the Multi-PIE subset is 86.3% for LPQ and 88.3% for PHOG, respectively. For JAFFE, it is 83.3% for LPQ and 90.8% for PHOG. For SFEW, it is 53.1% for LPQ and 57.2% PHOG. Figure 3.4 shows the classification accuracy comparison for LPQ and PHOG on SFEW, JAFFE and Multi-PIE. For the ***seven expression class*** experiment, the classification accuracy for JAFFE is 69.0% for LPQ and 86.4% for PHOG. For SFEW, it is 43.7% for LPQ and 46.3% for PHOG. LPQ and PHOG achieve high classification accuracy on JAFFE and Multi-PIE, but have significantly lower classification accuracy on SFEW.

It is evident from the graph (Figures 3.3 and 3.4) that the descriptors perform better on CK+ as compared to AFEW. This gives an insight into the performance of the current state-of-the-art descriptors when applied to close to real-world scenarios. Figure 3.2 compares frames from AFEW and CK+. From the results in the Figure 3.2,

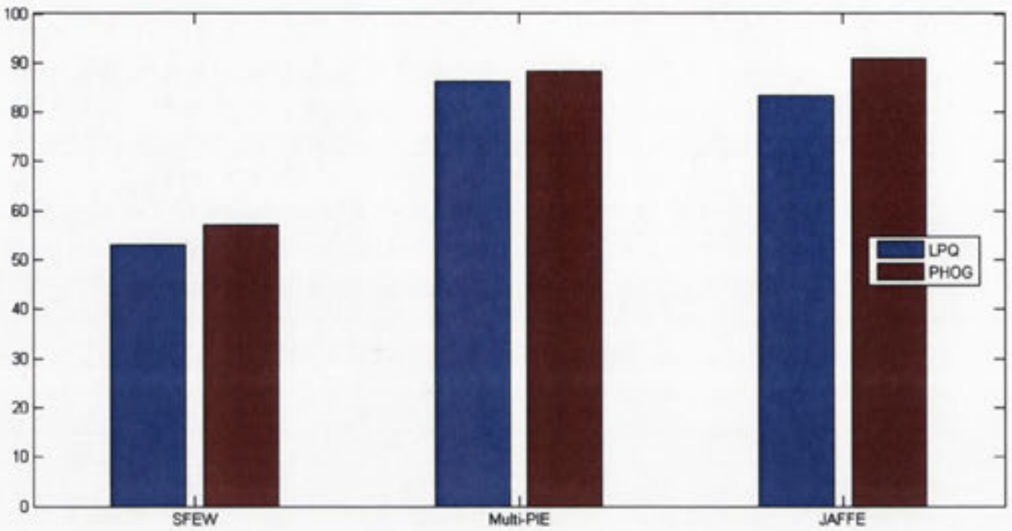


Figure 3.4: Four expression class accuracy comparison of SFEW, JAFFE and Multi-PIE based on LPQ and PHOG descriptors, and SVM classification.

it is evident that out of lab scenario make the problem of FER challenging. It can be argued that expression analysis in (close to) real-world situations is a non-trivial task and requires more sophisticated methods at all stages of the approach, such as robust face localisation/tracking, illumination and pose invariance. Further, experiments in Section 3.8 show that using robust face localisation/tracking method improves the performance of the system.

3.7 Baselines

LBP-TOP performs the best out of the four descriptors (Figure 3.3) for AFEW. This is due to the fact that LBP-TOP encodes the temporal information as well. It is chosen for creating baseline for the three AFEW protocols. The system performs better for PPI protocol as compared to SPI. This is similar to the performance of methods in the FERA challenge Valstar et al. [2011]. The features should ideally discard identity specific information and only capture the facial dynamics. However, as LBP-TOP is a texture descriptor, it does encode identity information up to some extent. For all the protocols for SFEW, the baselines based on the method defined in Dhall et al. [2011c] are computed. PHOG Bosch et al. [2007] and LPQ Huang et al. [2011] features are computed on the cropped face. The features are concatenated together to form a feature vector. For dimensionality reduction, PCA is computed and 98% of the variance is kept. Further, a non-linear SVM is used to learn and classify expressions. Parameter selection is performed using five-fold cross validation. Table 3.3 describes the expression

Protocol	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise	Avg.
AFEW_PPI	32.5	12.3	14.1	44.2	33.8	25.2	21.8	26.3
AFEW_SPI	40.1	7.9	14.5	37.0	40.1	23.5	8.9	24.5
AFEW_SPS	-	-	0.0	50.0	0.0	-	50.0	25.0
SFEW_PPI	29.5	43.5	48.5	35.5	33.0	12.0	35.0	33.8
SFEW_SPI	23.0	13.0	13.9	29.0	23.0	17.0	13.5	18.9
SFEW_SPS	35	-	45.8	0.0	7.1	-	0.0	17.5

Table 3.3: Average classification accuracies (in %) of different protocols on AFEW 2.0 and SFEW the three protocols (Section 3.4)

class-wise and overall classification accuracy for the three protocols for both AFEW and SFEW.

3.8 Emotion Recognition In The Wild

Emotion recognition traditionally has been been based on databases where the subjects posed for a particular emotion Lucey et al. [2010], Pantic et al. [2005b]. With recent advancements in emotion recognition various spontaneous databases have been introduced Valstar et al. [2011], McKeown et al. [2010]. For providing a common platform for emotion recognition researchers, challenges such as the Facial Expression Recognition & Analysis (FERA) Valstar et al. [2011] and Audio Video Emotion Challenges 2011 Schuller et al. [2011], 2012 Schuller et al. [2012] have been organised. These are based on spontaneous emotion databases Valstar et al. [2011], McKeown et al. [2010].

Emotion recognition methods can be categorised on the type of environment: lab-controlled and 'in the wild'. Traditional databases and methods proposed on them are based in a lab-controlled environment. This generally means uncluttered (and generally static) backgrounds, controlled illumination and minimal subject head movement. This is not a correct representation of real-world scenarios. Databases and methods which represent close-to-real-world environments (such as indoor, outdoor, different color backgrounds, occlusion and background clutter) have been recently introduced.

For moving emotion recognition systems from labs to the real-world, it is important to define a platform, where researchers can verify their methods on data representing close-to-real-world scenarios. The EmotiW[6] Dhall et al. [2013] challenge aims to provide a platform for researchers to create, extend and verify their methods on real-world data.

[6] http://cs.anu.edu.au/few

Database	Challenges	Natural	Label	Environment	Subjects Per Sample	Construction Process
AFEW Dhall et al. [2012a]	EmotiW	Spontaneous (Partial)	Discrete	Wild	Single & Multiple	Semi-Automatic
CK+ Lucey et al. [2010]	-	Posed	Discrete	Lab	Single	Manual
GEMEP-FERA Valstar et al. [2011]	FERA	Spontaneous	Discrete	Lab	Single	Manual
MMI Pantic et al. [2005b]	-	Posed	Discrete	Lab	Single	Manual
Semaine Pantic et al. [2005b]	AVEC	Spontaneous	Continous	Lab	Single	Manual

Table 3.4: Comparison of AFEW database which forms the basis of the EmotiW 2013 challenge Dhall et al. [2013].

AFEW also forms the bases of the EmotiW challenge being organised as part of The ACM International Conference of MultiModal Interaction ICMI 2013. The challenge seeks participation from researchers working on emotion recognition intend to create, extend and validate their methods on data in real-world conditions. There are no separate video-only, audio-only, or audio-video challenges. Participants are free to use either modality or both. Results for all methods will be combined into one set in the end. Participants are allowed to use their own features and classification methods. Table 3.4 compared EmotiW with other emotion recognition challenges.

The challenge data is divided into three sets: '*Train*', '*Val*' and '*Test*'. *Train*, *Val* and *Test* set contain 380, 396 and 312 clips respectively. The data is subject independent and the sets contain clips from different movies. The motivation behind partitioning the data in this manner is to test methods for unseen scenario data, which is common on the web. For the participants in the challenge, the labels of the testing set are unknown. The subject distribution among the various sets is described in Table 3.5.

3.8.1 Audio Features

In this challenge, a set of audio features similar to the features employed in Audio Video Emotion Recognition Challenge 2011 Schuller et al. [2011] motivated from the

Set	Number of Subjects	Max. Age	Avg. Age	Min. Age	Males	Females
Train	99	76y	32.8y	10y	60	39
Val	126	70y	34.3y	10y	71	55
Test	90	70y	36.7y	8y	50	40

Table 3.5: Subject description of the three sets.

(a) **Set of functionals applied to Low Level Descriptors (LLD).**

Functionals
Arithmetic Mean
standard deviation
skewness, kurtosis
quartiles, quartile ranges
percentile 1%, 99%
percentile range
Position max./min
up-level time 75/90
linear regression coeff.
linear regression error(quadratic/absolute)

(b) **Audio feature set - 38 (34 + 4) low-level descriptors**

LLD	
Energy/Spectral LLD	PCM Loudness MFCC [0-14] log Mel Frequency Band [0-7] Line Spectral Pairs (LSP) frequency [0-7] F0 F0 Envelope
Voicing related LLD	Voicing Prob. Jitter Local Jitter consecutive frame pairs Shimmer Local

Table 3.6: *Audio* descriptors in the EmotiW baseline.

INTERSPEECH 2010 Paralinguistic challenge (1582 features) Schuller et al. [2010] are used. The features are extracted using the open-source Emotion and Affect Recognition (openEAR) Eyben et al. [2009] toolkit backend openSMILE Eyben et al. [2010].

The feature set consists of 34 energy & spectral related low-level descriptors (LLD) × 21 functionals, 4 voicing related LLD × 19 functionals, 34 delta coefficients of energy & spectral LLD × 21 functionals, 4 delta coefficients of the voicing related LLD × 19 functionals and 2 voiced/unvoiced durational features. Table 3.6 describe the details of LLD features and functionals.

3.8.2 EmotiW Experiments

For computing the baseline results, openly available libraries are used. Pre-trained face models (Face_p146_small, Face_p99 and MultiPIE_1050) available with the MoPS Zhu and Ramanan [2012] are applied for face and fiducial points detection. The models are applied in hierarchy. With the MoPS framework, face and fiducial points are detected in a single framework. It is the current state-of-the-art face and facial parts detector (along with Asthana et al. [2013a], Xiong and De la Torre [2013], however these two require accurate face localisation) and hence is preferred over VJ which is used for initial experiments on AFEW (Section 3.6). The fiducial points generated by MoPS is used for aligning the face and the face size is set to 96 × 96. Post aligning LBP-TOP features are extracted from non-overlapping spatial 4×4 blocks. The LBP-TOP feature from each block are concatenated to create one feature vector. Non-linear RBF kernel based SVM is learnt for emotion classification. The video only baseline system achieves 27.2% classification accuracy. The audio baseline is computed by extracting features using the OpenSmile toolkit. A linear SVM classifier is trained. The audio only based system gives 19.5% classification accuracy. Further, a feature level fusion is performed, where the audio and video features are concatenated and a non-linear SVM is learnt. The performance drops here and the classification accuracy is 22.2%. On the test set which contains 312 video clips, audio only gives 22.4%, video only gives 22.7% and feature fusion gives 27.5%.

Figure 3.5, compares the performance of various participating teams with the baselines. Ebrahimi et al. [2013], proposed a deep learning based emotion recognition method and their system performed the best (41.02%) out of all others. The second best performance is from the team of Sikka et al. [2013b], who proposed a multiple kernel learning based approach where different modalities are fused as multiple kernels. The second runner up is from the team of Liu et al. [2013], who proposed a method based on partial least square regression on Grassmannian manifolds and early noisy aligned face removal using PCA. Almaev et al. [2013] proposed a distribution based pairwise SVM based classification method. They used a variant of LBP-TOP called

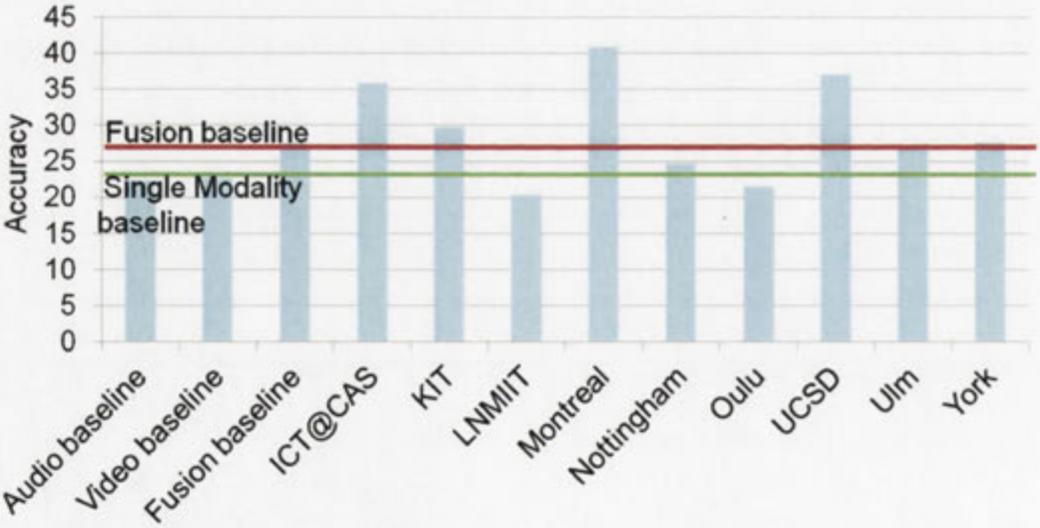


Figure 3.5: Comparison of the results on the EmotiW data of participating teams. The red line represents the late fusion EmotiW baseline. The green line represents the single modality EmotiW baseline.

Protocol	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise	Average
$\textbf{Val}_{\textbf{audio}}$	42.37	12.00	25.93	20.97	12.73	14.06	9.62	19.95
$\textbf{Test}_{\textbf{audio}}$	44.44	20.41	27.27	16.00	27.08	9.30	5.71	22.44
$\textbf{Val}_{\textbf{video}}$	44.00	2.00	14.81	43.55	34.55	20.31	9.62	27.27
$\textbf{Test}_{\textbf{video}}$	50.00	12.24	0.00	48.00	18.75	6.97	5.71	22.75
$\textbf{Val}_{\textbf{audio-video}}$	44.07	0.00	5.56	25.81	63.64	7.81	5.77	22.22
$\textbf{Test}_{\textbf{audio-video}}$	66.67	0.00	6.06	16.00	81.25	0.00	2.86	27.56

Table 3.7: Classification accuracy for *Val* and *Test* sets for *audio*, *video* and *audio-video* modalities.

Local Gabor Binary Pattern in Three Orthogonal Planes (LGBP-TOP) Almaev and Valstar [2013], which is more robust to alignment errors. Meudt et al. [2013] proposed an audio feature selection based method which uses SVM for classification. Day [2013] proposed a boosted decision trees based audio only classifier for their entry in EmotiW. Krishna et al. [2013] proposed an audio visual approach by fusing various audio and video features and classification using SVM.

Table 3.7, describes the classification accuracy for the *Val* and *Test* for audio, video and audio-video systems. For the *Test* set the feature fusion increases the performance of the system. However, the same is not true for the *Val* set. In Appendix A, Table A.1, describes three tables displaying the confusion matrix for $\textbf{Val}_{\textbf{audio}}$, $\textbf{Val}_{\textbf{video}}$, $\textbf{Val}_{\textbf{audio-video}}$. Table A.2, describes three tables displaying the confusion matrix for $\textbf{Test}_{\textbf{audio}}$, $\textbf{Test}_{\textbf{video}}$, $\textbf{Test}_{\textbf{audio-video}}$.

3.9 Summary

The central hypothesis of this chapter is that movie data can be used for constructing facial expression databases representing challenging real-world conditions in a semi-automatic framework. Collecting databases in computer vision is one of the most time consuming and laborious task. For the problem of facial expression analysis, databases have been recorded in lab-conditions. For the advancement of the field, database representing the real-world scenarios are required. Moreover capturing spontaneous expressions in close-to-real-world scenarios is a non-trivial problem. To overcome this, short video clips from movies are used for creating a new facial expressions database representing real-world settings. To expedite the data collection and labelling process a semi-automatic method for database creation and labelling is proposed. A clear experimental protocol is defined in Section 3.4. The database is compared to a classic facial expressions database CK+. The results in Section 3.6 (Figure 3.2, 3.3) show that the current state-of-the-art methods do not perform well on 'in the wild' data. A challenge is being organised based on AFEW. Its details and baseline is discussed in Section 3.8.

A future direction for this research is to use state-of-the-art face detectors for discarding video clips where the actors are not facing the camera. Further, online learning based subject re-identification can be incorporated into the recommendation system. This will save time in labelling the identity of the subjects.

In this chapter, the MoPS framework performs fairly well 3.8.2 for face and fiducial points detection. The next step for a FER method is head pose normalisation, for computing feature descriptors in a canonical frame. In the next Chapter 4, a head pose normalisation technique is discussed based on MoPS framework.

Chapter 4

Head Pose Normalisation

In everyday situations and natural conversations, humans generally tend to move their head while expressing themselves. This leads to several challenges, such as out-of-plane head rotations, (self-) occlusion and illumination variations (Figure 1.1). Facial landmark localisation and head pose handling play a vital role as pre-processing steps for facial analysis in research areas such as HCI, biometrics, and affective computing, and have been an active fields of research (e.g. Cootes et al. [1995], Zhu and Ramanan [2012]). Particularly for face recognition and spontaneous facial expression analysis in real-world conditions, the head pose needs to be normalised to the frontal pose for robust inference. For dealing with non-frontal faces, due to out-of-plane rotations, a face needs to be registered accurately. In Section 3.8.2, MoPS is used for detecting facial parts for the EmotiW challenge data's baseline (Section 3.8). Post robust fiducial points detection the next step is to normalise the head pose, originally for the EmotiW challenge, the non-frontal pose is only partially handled by applying an affine warp. However, this does not handles strictly non-frontal head poses well. This chapter proposes a view-invariant Head Pose Normalisation (HPN) framework based on the MoPS framework. Given a non-frontal face image, the proposed framework normalises the head pose and reconstructs the facial points for the input face in its frontal pose (referred to as virtual pose). An example is shown in Figure 4.1. The contents of this chapter have been published in Dhall et al. [2014].

A PS Felzenszwalb and Huttenlocher [2005] based framework is proposed for generating shape-constrained virtual frontal face points from a non-frontal face image. In particular, the proposed HPN method builds upon the MoPS framework Zhu and Ramanan [2012], where the deformation of a part is learnt with respect to its mother part as a mixture. MoPS framework jointly infers the location of the face, facial parts and the head pose. In contrast, the proposed HPN method normalises a non-frontal head pose within the MoPS inference framework.

The proposed approach employs the response maps generated from discriminatively

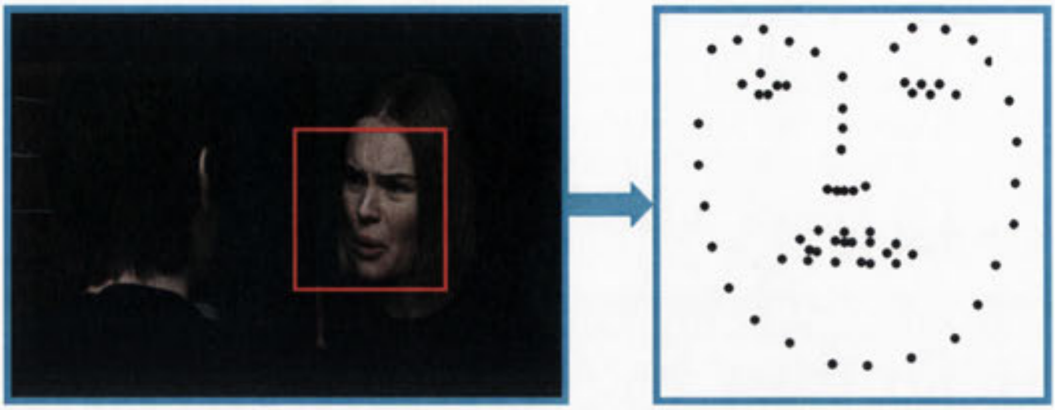


Figure 4.1: **Automatic Head Pose Normalisation (HPN):** Given a non-frontal face Dhall et al. [2011c], the proposed framework reconstructs the input face's corresponding facial points in the virtual frontal pose. Image source: SFEW

trained facial part detectors. These confidence score maps are then normalised from non-frontal to frontal head pose using block-wise structure regression. A shape model is further applied on the virtual pose normalised confidence score maps to generate the virtual frontal landmark points. The entire framework is embedded within the MoPS framework Zhu and Ramanan [2012] for achieving robust performance on real-world images.

The rest of the chapter is organised as follows: The key contributions of the chapter are discussed in Section 4.1. The MoPS framework is described in Section 4.2. Points based HPN methods are detailed in Section 4.3. The point wise and all points based HPN methods are explained in Section 4.3.1 and 4.3.2 respectively. Twin-GPR and Rudovic and Pantic [2011] are discussed in Section 4.4. The proposed HPN method based on confidence maps is discussed in Section 4.5. HPN method when the head pose is known is described in Section 4.5.1. An extension for invariance from head pose information is described in Section 4.5.2. The experiments on Multi-PIE Gross et al. [2008a] and SFEW Dhall et al. [2011c] databases are described in Section 4.6. The chapter is concluded in Section 4.7.

4.1 Contributions of the Chapter

The **contributions** of the chapter are as follows:

1. Previous methods Asthana et al. [2009a, 2011], Rudovic et al. [2010a,b] require facial landmark points as a prerequisite for HPN. However, robust facial landmark detection itself is still an active, open field of research, particularly when dealing with real-world images, leading to errors in the results. In contrast, the confidence maps based HPN methods use response maps generated by parts based detectors,

thereby not requiring any inputs from facial point detectors, resulting in more robust and accurate HPN.

2. The virtual frontal points returned by the proposed HPN methods are generated with a shape constraint. This overcomes the problem of standard regression based methods Asthana et al. [2009a, 2011], Rudovic et al. [2010b,a] where there is no implicit constraint on the shape of the object among the input and output data.

3. Previous methods Asthana et al. [2009a, 2011], Rudovic and Pantic [2011], Rudovic et al. [2010a,b] require head pose information for selecting a pose-specific regression model, which is certainly error-prone on real-world images. In contrast, the proposed method is head pose invariant.

4. Many researchers have pointed out that texture contains a strong signal for face analysis problems such as face tracking Cootes et al. [2002], identity Asthana et al. [2011], and expression recognition Moore and Bowden [2011]. Many studies such as Asthana et al. [2009a, 2011], Rudovic and Pantic [2011], Rudovic et al. [2010a,b] perform regression on facial points directly and, thus, lack texture information. The frontal normalised confidence maps being generated by the proposed HPN method can also be used directly as texture descriptors (e.g. Gabor filter response).

4.2 Mixture of Pictorial Structures

The **Mixture of Pictorial Structure** framework Felzenszwalb and Huttenlocher [2005] represents the parts of an object as a graph with n vertices $V = \{v_1, \ldots, v_n\}$ and a set of edges E. Here, each edge $(v_i, v_j) \in E$ pair encodes the spatial relationship between parts i and j. A face is represented as a tree graph here. Formally speaking, for a given image I, the MoPS framework computes a score for the configuration $L = \{l_i : i \in V\}$ of parts based on two models: an *appearance model* and a *spatial prior model*. These two models will be discussed now using the tree-based pictorial structures formulation similar to Zhu and Ramanan Yang and Ramanan [2011], Zhu and Ramanan [2012]. In particular, the formulation of Zhu and Ramanan [2012] is followed.

The **Appearance Model** scores the confidence of a part-specific template w_p applied to a location l_i. Here, p is a view-specific mixture corresponding to a particular head pose. $\phi(I, l_i)$ is the histogram of oriented gradient descriptor Dalal and Triggs [2005b] extracted from a location l_i. Thus, the appearance model calculates a score for configuration L and image I as:

$$App_p(I, L) = \sum_{i \in V_p} w_i{}^p.\phi(I, l_i) \tag{4.1}$$

The advantage of the part templates (detectors) in the appearance model is that less amount of data is required for training for each part detector. The response maps generated by these discriminative part detectors are sparse, which makes their reconstruction in frontal view (for HPN) simpler.

The **Shape Model** learns the kinematic constraints between each pair of parts. The shape model (as in Zhu and Ramanan [2012]) is defined as:

$$Shape_p(L) = \sum_{ij \in E_p} a^p_{ij} dx^2 + b^p_{ij} dx + c^p_{ij} dy^2 + d^p_{ij} dy \tag{4.2}$$

Here, dx and dy represent the spatial distance between two parts. The parameters a, b, c and d correspond to the location and rigidity of a spring, respectively. From Eqs. 4.1 and 4.2, the scoring function S is:

$$Score(I, L, p) = App_p(I, L) + Shape_p(L) \tag{4.3}$$

During the inference stage, the task is to maximise Eq. 4.3 over the configuration L and mixture p (which represents a pose). Therefore, if the pose of the face is known, then the inference is equivalent to finding the configuration L^*, which maximises the score for a given pose p[1]:

$$L^* = \max_L(Score(I, L, p)) \tag{4.4}$$

When the pose is unknown, all models learnt for different values of p are applied (Eq. 4.4) and the configuration specific to the highest scoring mixture is chosen as the facial parts locations. The use of a graph structure allows the application of dynamic programming and a distance transform, while finding the part location on the parts based model's scores.

4.3 Points Based Head Pose Normalisation

The points based HPN methods being discussed now are based on the idea of applying regression Asthana et al. [2009a], Rudovic et al. [2010a,b] over non-frontal points to obtain frontal points. For an image I containing a non-frontal face, the fiducial points locations are computed using the parts based model discussed in the previous section. For HPN, a mapping function (regressor) $F : L^i_p \rightarrow L^i_f$ is learnt that maps point

[1] It is assumed here that the head pose is known.

locations in the non-frontal view to locations in the frontal view. L_p^i and L_f^i are the $2D$ coordinates of part i in non-frontal and frontal pose, respectively. It should be highlighted that Rudovic et al. [2010b] also learnt a similar mapping; however, during the test phase, *manually* defined landmark points were used as input. In contrast, in the proposed approach (Section 4.5), the part locations are computed *automatically*. Therefore, the results (Section 4.6) are closer to a real-world scenario and account for error due to face detection and facial parts localisation. Two different variants of Points based HPN methods used in the experiments section (Section 4.6) are discussed below.

4.3.1 Part Wise Points (PWP) Based Normalisation

In PWP based HPN methods, frontal points are generated by regressing one point at a time using a point specific regression model. Based on univariate regression, n models corresponding to each part ($2D$ location) are trained for each non-frontal pose. Thus, the total number of models learnt is $n * P$, where P is the number of non-frontal poses in the training data. The mapping function is then the regression function $\mathcal{R}_l : L_p \rightarrow L_f^i$, i.e. the frontal location L_f^i of each part is learnt from its corresponding non-frontal locations L_p. The major limitation with this method is that the outputs from different regression models are treated individually. As pointed out by Rudovic and Pantic [2011], in such a case, there is no guarantee that the regressed part locations will adhere to the anatomical shape constraint of the face. The next model (4.3.2) addresses this limitation.

4.3.2 All Parts Points (APP) Based Normalisation

The limitation of PWP is overcome by learning a multi-variate regression model. In APP based HPN frontal points are generated by learning a single regression model. In other words, a function $\mathcal{R}_l : L_p \rightarrow L_f$ that maps all parts in the non-frontal pose to all parts in the frontal pose is learnt. In the classic GPR framework Rasmussen [2006], a multi-variate regression model is generally computed by mapping independent input points to single output dimension models. There is no explicit constraint, which models the relationship between the output dimensions. Thus, the model is made more robust by posing the APP based normalisation as a structured regression problem and using the Twin Gaussian process regression (Twin-GPR) Bo and Sminchisescu [2010] framework. In the next section Twin-GPR and its limitation when used for HPN are discussed.

4.4 Twin-GPR

The Twin-GPR framework models the relationship between the input and output variables. It uses Gaussian process priors on both covariances and responses, both multivariate. The Kullback-Leibler divergence between the input and output data distributions, modelled as a Gaussian process, is minimised for capturing the correlation between the output dimensions. Bo and Sminchisescu [2010] proposed this method for regressing 2D human pose, which is a structured regression problem, where the output dimensions are correlated by the human body kinematics. Similar to their problem, the intent in this chapter is to reconstruct the facial points, where the points adhere to the anatomical face shape-constraint. See Bo and Sminchisescu [2010] for details of the method.

Twin-GPR assumes that the input and output distributions are Gaussian. For images in real-world conditions, this assumption may not be satisfied due to the error induced during the face alignment step Rudovic and Pantic [2011]. This drawback is addressed in Rudovic and Pantic [2011] by learning shape models based on ASM Cootes et al. [1995]. Shape parameters are applied during the GPR inference to maintain the face shape constraint. To calculate the shape parameters, facial points in the frontal view are required. To synthesise the constrained shape in the frontal view, the shape parameters are required, which creates a chicken-and-egg problem. Two methods are proposed here to overcome this situation: (1) Shape parameters are estimated from frontal view points synthesised using a normal GPR regression. These shape parameters are then used in Shape Constrained Gaussian Process (SC-GP) regression. (2) GPR regression is used to synthesise frontal view shape points and shape parameters together. The regressed shape parameters are then used to reconstruct the shape. Next, a parameter search is performed, which reduces the error between the SC-GP output and the shape reconstructed using the parameters.

A limitation of deformable models, such as ASM, is that they perform very well for subject-dependent data (i.e. the subject in the training and test images is the same), but their performance on subject-independent data is not robust. Ideally, for a face analysis problem such as FER, the face alignment method should be invariant to the subject's identity for making it work in real-world conditions Chew et al. [2012]. ASM is also sensitive to initialisation and requires accurate face detection. To overcome this limitation, a **Confidence map based HPN** (CM-HPN) that exploits the advantage of parts based detectors and performs HPN on the parts detector response is proposed. Facial landmark points are not required for HPN when it is performed within the PS inference framework. This is the main **benefit** of the proposed method CM-HPN over the prior work Asthana et al. [2009a], Rudovic et al. [2010a,b]. In the experiments (Section 4.6), the performance of two CM-HPN methods (discussed below in Sections

4.5) is compared with the points based methods Asthana et al. [2009a], Rudovic et al. [2010b].

4.5 Confidence Map Based HPN

The primary idea in this work is to learn a mapping from the raw outputs of parts based detectors (confidence maps) for non-frontal faces to their frontal counterpart. This step would normalize the head-pose. A Confidence map is a 2D matrix whose each element's value is the detector's inference score describing the probability of presence of a part. Further, a shape constraint by exploiting properties of parts based models is applied. The mathematical formulation for cases with known pose is discussed in Section 4.5.1 and later extended to unknown poses (Section 4.5.2).

4.5.1 Pose Specific Confidence Map Based HPN

Recall that Eqs. 4.1 and 4.2 are the appearance and shape components of the overall score (Eq. 4.3) optimized by the PS model. Given a non-frontal face image for pose p, part-specific filters are applied via Eq. 4.1. This produces part-specific response maps (denoted by App_p). For simplicity, the response of the appearance model for a particular part i is denoted as a function θ, which is defined as:

$$C_i^p = \theta(I, i, p) \tag{4.5}$$

The response C_i^p will be a matrix of the size of the image I. Component (x, y) of this matrix represents the probability of part i being present at location (x, y) in the image. The task is to reconstruct the response map at the frontal pose for part i, referred to as C_i^f, from its response map C_i^p at pose p. This is achieved using a structured regression model (Twin-GPR, Section 4.4) as discussed below. C_i^f can be considered as a (synthesised) *virtual frontal view response map* for a part i. The motivation of using Twin-GPR is to maintain the relationship between neighbouring points being inferred in the frontal view response map.

The response maps C_i^p is divided into blocks and a *block-by-block* Twin-GPR regression is learnt. This idea is motivated by the work of Biederman and Kalocsais [1997], who discuss the importance of maintaining location information for facial parts when dealing with faces in a holistic manner. Thus, each response map C_i^p is first divided into k equal sized non-overlapping blocks $B = \{B_1^p B_2^p....B_k^p\}$ as shown in Figure 4.2 and a separate regression function is learnt for mapping each block. Thus, the big problem of mapping an entire confidence map is transformed into many smaller problems of mapping individual blocks with the aim of maintaining a structure. Non-overlapping blocks are preferred over scanning window or overlapping blocks for their computa-

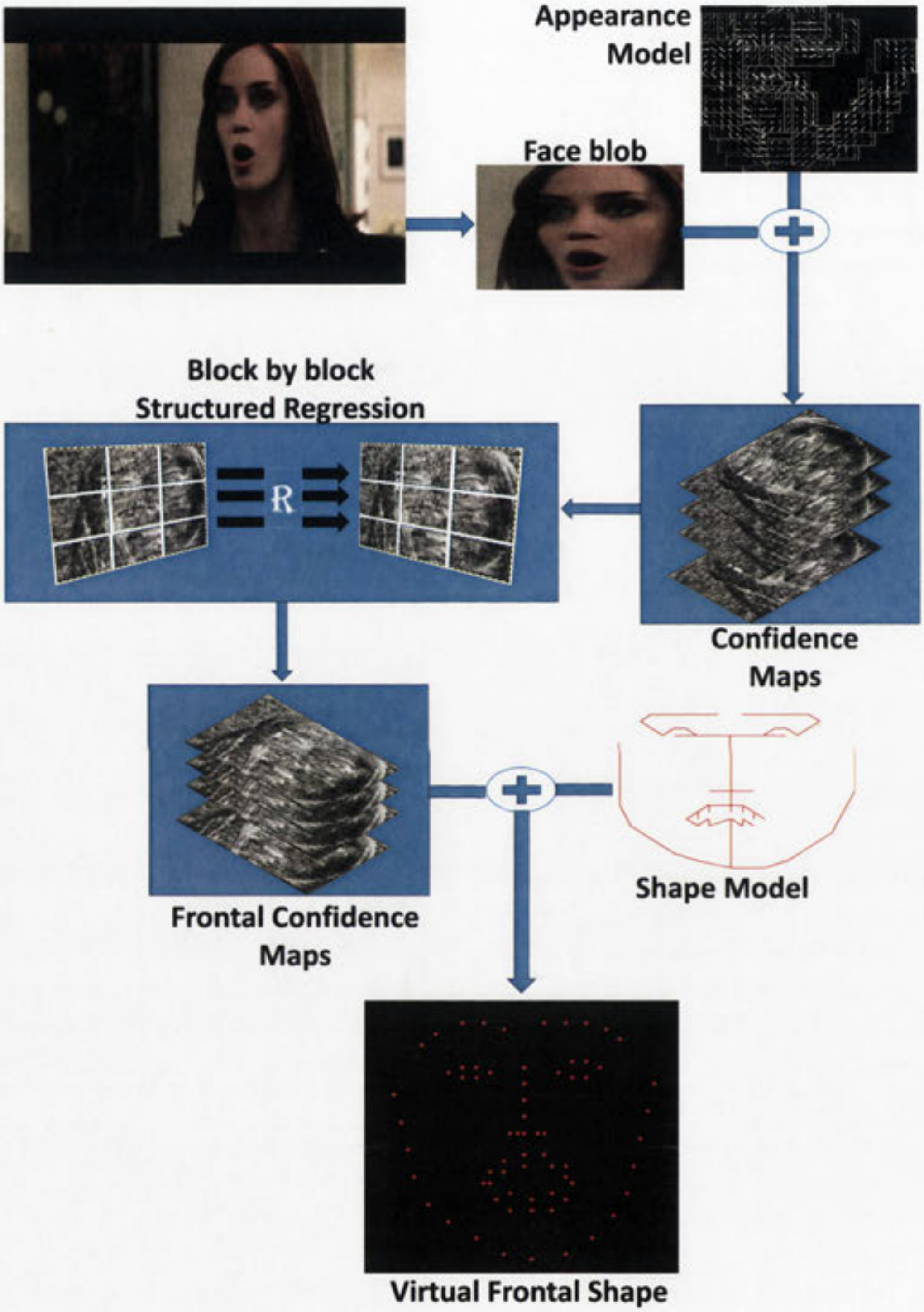


Figure 4.2: *Confidence Map Based HPN:* Steps for normalising the head pose of an input image via parts based models, structured regression and confidence maps.

Algorithm 1: Frontal virtual points reconstruction using pose-specific confidence map regression

```
Input: Image $I$ and pose $p$
Output: $Score'$ and $L_f^*$
for $part\ i \in V$ do
    Compute part wise confidence maps, $C_i^p = \theta(I, i, p)$ (Eq. 4.5) ;
    Divide $C_i^p$ into $k$ blocks $B$
    $B = \{B_1^p B_2^p ....B_k^p\}$;
    for $a = 1 : k$ do
        Reconstruct $B_a^f \leftarrow B_a^p$ using corresponding model from $\mathcal{R}_i$
    end
    Rejoin reconstructed blocks $C_i^f \leftarrow \{B_1^f, B_2^f ...B_k^f\}$ ;
end
$FC \in \sum_{i \in V} C_i^f$ ;
Compute frontal $Shape_f(L)$ (Eq. 4.2) and maximise $Score'$ (Eq. 4.8)
$L_f^* = \max_L(Score'(I, L, p)$
```

tional simplicity. Mathematically, during **training** a set of models is trained for each part i, denoted by $\mathcal{R}_i$. Each set $\mathcal{R}_i$ comprises of a regression model $\mathcal{R}_i^j$ that maps block B_j^p to its frontal counterpart written as B_j^f. For each pose, the models are learnt independently. The training models can be represented as:

$$\mathcal{R} = \{\mathcal{R}_i | i \in (1, 2, 3, ..., n)\} \tag{4.6}$$

$$\mathcal{R}_i = \{\mathcal{R}_i^j | j \in (1, 2, 3, ..., k)\} \tag{4.7}$$

On a big picture level, the process of reconstructing frontal maps for each block and concatenating them produces virtual frontal view response maps for each part i. Virtual response maps for all parts together shall be referred to as $Virt_p$. $Virt_p$ is generated from App_p and is referred to as set of initial response maps at pose p.

Shape Constraint: Further the shape constraint is applied on the virtual frontal pose maps $Virt_p$ to generate the virtual frontal shape as shown in Figure 4.2. This is accomplished by jointly maximising a modified score function where $Virt_p$ is used as appearance response and the shape model (denoted as $Shape_f$) corresponding to the frontal pose is fixed:

$$Score'(I, L, p) = Virt_p(I, L) + Shape_f(L) \tag{4.8}$$

The intuition behind fixing the shape model to the frontal pose is to constrain the framework to output (virtual) fiducial point locations in the frontal pose only. The point locations are then obtained by solving the above optimisation problem using dynamic programming.

Since the head pose is known *a priori*, this method is referred to as the ***pose-specific*** Confidence Map based HPN (**CM-HPN**$_{PS}$). Algorithm 1 describes the reconstruction process in detail. CM-HPN$_{ps}$ is limited in that as it requires head pose information (similar to Asthana et al. [2009a], Rudovic et al. [2010b,a], Asthana et al. [2011]) and, hence, in the next section (4.5.2), a technique to extend **CM-HPN**$_{PS}$ for unknown head poses is presented.

4.5.2 Pose Invariant Confidence Map Based HPN

As discussed in Section 4.5.1, when the head pose is unknown, the configuration L^* of the highest scoring mixture p is chosen as the best facial parts location. Based on this model, CM-HPN$_{PS}$ can be computed in a pose-invariant manner by simply enumerating over the $Score'(I, L, p)$ of each pose. Substituting $Score'$ (from Eq. 4.8) into Eq. 4.4 and maximising over all poses in the training data, the **CM-HPN**$_{PI}$ based $Virt_p$ based inference maximises:

$$L_f^* = \max_p [\max_L (Score'(I, L, p))] \tag{4.9}$$

Here, L_f^* is the highest scoring 'virtual' frontal head pose configuration. Basically, Algorithm 1 is computed for all poses p in the training set and L is the 'virtual' frontal head pose configuration generated with a regression model specific to a pose p.

4.6 Experiments

The CMU Multi-PIE Gross et al. [2008b] is a static facial expression database. The dataset for the experiments in this chapter is based on the protocol used by Rudovic et al. [2010b]. The experiment dataset contains images from four pan angles (0°, $-15°$, $-30°$, $-45°$), with 200 images per pose. There are a total of 74 subjects. A five-fold cross validation over the samples is performed. To compare the points regression based HPN methods Asthana et al. [2009a], Rudovic et al. [2010b], the methods are implemented. PWP and APP (Section 4.3) are based on Asthana et al. [2009a], Rudovic et al. [2010b], with the difference that during inference, the facial points are located using the MoPS framework. Using MoPS (for PWP and APP) gives an edge to the methods (PWP and APP) in terms of their performance due to good initialisation as compared to initialisation methods used in Asthana et al. [2009a], Rudovic et al. [2010b]. It is shown that the performance of Twin-GPR is better than both GPR and SVM Rudovic et al. [2010b]. Therefore, for learning PWP and APP, Twin-GPR is used. However for PWP, the output dimension is a single variate only (separate models are trained for x and y positions).

Grid	2×2	3×3	4×4	5×5
NMSE	.081±.007	.076±.006	.061±.006	.059±0.0002

Table 4.1: NMSE comparison for 4 grid configurations for CM-HPN$_{PS}$. 5×5 has the smallest NMSE and, hence, performs the best.

Implementation details: The face area is located using the VJ face detector Viola and Jones [2001]. The faces **are not aligned** as a pre-processing step before the HPN step, as face alignment is a non-trivial problem for real-world images. Next, the detected face areas are rescaled to 320×240 pixels for consistency. The parameters for Twin-GPR are tuned empirically. The range of parameters experimented for the RBF kernel size is $[0.1 - 2]$ and for lambda is $[1.0e^{-1} - 1.0e^{-5}]$. The MoPS framework from Zhu and Ramanan [2012] is used and all experiments are based on independent models. Zhu and Ramanan [2012] reports that the performance difference of independent and parts sharing models is not high. However, independent MoPS models are more accurate. Moreover, the models are learnt at three different pyramid scales for achieving scale invariance. For the training details of the MoPS, see the original paper Zhu and Ramanan [2012].

The performance of the points (PWP, APP) and confidence maps based HPN methods (CM-HPN$_{PS}$, CM-HPN$_{PI}$) is compared on the Multi-PIE database Gross et al. [2008b] using the error in the location of the reconstructed landmark points w.r.t. the frontal landmark points in the ground truth. Zhu and Ramanan Zhu and Ramanan [2012] normalised the landmark location using the inter-occular distance and the average of height and width of faces. Similar to Zhu and Ramanan [2012], *Normalised Mean Square Error* (**NMSE**), which describes the landmark localisation error (the L2 distance between the virtual frontal points and the ground truth) normalised by the face size is used. This facilitates a fair comparison of the proposed methods (CM-HPN$_{PS}$, CM-HPN$_{PI}$) with others in the future.

As discussed in Section 4.5, both CM-HPN$_{PS}$ and CM-HPN$_{PI}$ (Section 4.5) are computed block-by-block in a grid, whose configuration is chosen empirically. Different grid structures for non-overlapping blocks: $[2 \times 2, 3 \times 3, 4 \times 4, 5 \times 5]$ are compared. Table 4.1 summarises the performance comparison in terms of NMSE. The models with 25 blocks performed the best with the lowest NMSE. This supports the method's hypothesis that dividing maps into blocks reduces the complexity of the learnt model and maintains a spatial constraint. Similar results have been described in Lazebnik et al. [2006] w.r.t. natural image analysis and the bag-of-words framework. In this case, as the number of blocks increases, the performance also increases until it saturates. For further comparison of the confidence maps based methods with the point-based ones, the highest performing grid configuration 5×5 blocks is used. Table 4.2 shows the

Pose	15°	30°	45°	Avg.
PWP Asthana et al. [2009a] Rudovic et al. [2010b]	0.098 ±0.002	0.089 ±0.001	0.100 ±0.007	0.095 ±0.05
APP Rudovic and Pantic [2011]	0.062 ±0.001	0.087 ±0.003	0.100 ±0.005	0.084 ±0.02
CM-HPN$_{PS}$	**0.059** ±0.001	**0.058** ±0.003	**0.059** ±0.002	**0.059** ±0.0002
CM-HPN$_{PI}$	0.076 ±0.009	0.082 ±0.005	0.088 ±0.005	0.082 ±0.006

Table 4.2: NMSE comparison for the four HPN methods: PWP Asthana et al. [2009a], Rudovic et al. [2010b], APP Rudovic and Pantic [2011], CM-HPN$_{PS}$, CM-HPN$_{PI}$.

NMSE based performance for PWP, APP, CM-HPN$_{PS}$ and CM-HPN$_{PI}$. The 5 × 5 blocks (i.e. 25 blocks in total in a confidence map) grid configuration was chosen for CM-HPN$_{PS}$ and CM-HPN$_{PI}$. CM-HPN$_{PS}$ has the smallest NMSE, performing the best. APP and CM-HPN$_{PI}$ perform on par with each other (even though no prior head pose information is used in CM-HPN$_{PI}$). For the pose angle 45°, the NMSE is high for both points based methods. This is due to the fact that as the head pose deviates more from the frontal view, computing the facial points on the non-visible side is error prone. The reconstruction error is highest for PWP. This is primarily due to the lack of a relationship between the output of the different regressions models.

Ideally, the performance of CM-HPN$_{PS}$ and CM-HPN$_{PI}$ should be similar. However, CM-HPN$_{PI}$ has higher NMSE when the maximum score is achieved by an incorrect pose. For example: if for a face with original head pose 45°, CM-HPN$_{PI}$ will apply HPN with all models in $\mathcal{R}$ (for example: $45° \rightarrow 0°$, $30° \rightarrow 0°$). If HPN with 30° model scores higher than the 45° model's score, than the method will assume that the face has default head pose of 30° and will choose the corresponding incorrect reconstruction.

Expression wise NMSE is shown in Figure 4.3. For Surprise expression, the error is large for all methods but CM-HPN$_{PS}$. CM-HPN$_{PS}$ performs consistently best across all expressions. The biggest variation is for the points based methods PWP and APP, which are based on Asthana et al. [2009a], Rudovic and Pantic [2011]. **SFEW**: To test the performance of the methods (CM-HPN$_{PS}$ and CM-HPN$_{PI}$) on 'in the wild data', the SFEW database Dhall et al. [2011c], which contains a set of video frames depicting facial expressions from movies is used. Qualitative, visual comparison is performed as there is no frontal ground truth for SFEW unlike for Multi-PIE. Images where the pose is roughly similar to the Multi-PIE training set's pose range are chosen manually. Here, the performance of the HPN method is tested for: a) images 'in the wild', b) unseen pose, and c) unseen expressions.

Figure 4.4 shows the performance of the proposed methods on SFEW images. Results of point based regression methods are shown in columns 3 and 4, while those for

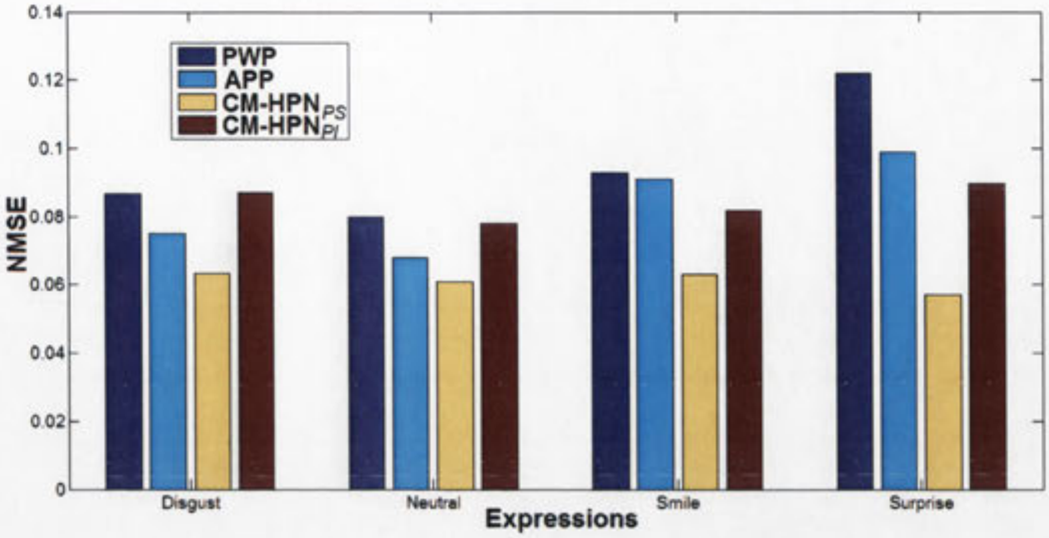


Figure 4.3: *Comparison of the performance of Point Wise Point (PWP) Asthana et al. [2009a], Rudovic et al. [2010b], All Part Point (APP) Rudovic and Pantic [2011], Pose Specific Confidence Map based Head Pose Normalisation (CM-HPN$_{PS}$) and Pose Invariant Confidence Map based Head Pose Normalisation (CM-HPN$_{PI}$) by facial expression.*

the confidence maps based regression methods are shown in columns 5 and 6 of Figure 4.4. It is evident from these images that the reconstruction of the overall shape for method PWP is not as accurate as the other methods and that APP is not able to reconstruct the mouth correctly in some cases. Among all four methods, CM-HPN$_{PS}$ generally performed the best, but is unable to reconstruct eyes clearly in some cases. This can be addressed by employing denser grids. The reader should note that the initialisation for the confidence score based methods is done by the VJ face detector. If a more accurate face detector such as MoPS itself is used, the reconstruction quality is expected to improve. It is also interesting to note that the jaw line of the reconstructed faces for the outputs of CM-HPN$_{PS}$ and CM-HPN$_{PI}$ shows a high similarity to the jaw line shape of the subjects in the corresponding non-frontal images as compared to the output of APP, where the jaw line seems to be 'averaged out'.

It is worth noting that Asthana et al. [2009a], Rudovic et al. [2010b], Rudovic and Pantic [2011] either use manually defined points or AAM. In contrast the performance of APP can be attributed to robust landmarks detection by the MoPS framework. As discussed earlier (Section 4.4), Rudovic and Pantic [2011] proposed SC-GP to apply a shape-constraint and to overcome the problems, which may arise due to inaccurate facial landmarks detection while regressing using Twin-GPR. Therefore, the performance of parts based methods can benefit by using a MoPS model.

In the last row of Figure 4.4, the reconstruction of the CM-HPN$_{PI}$ method is not

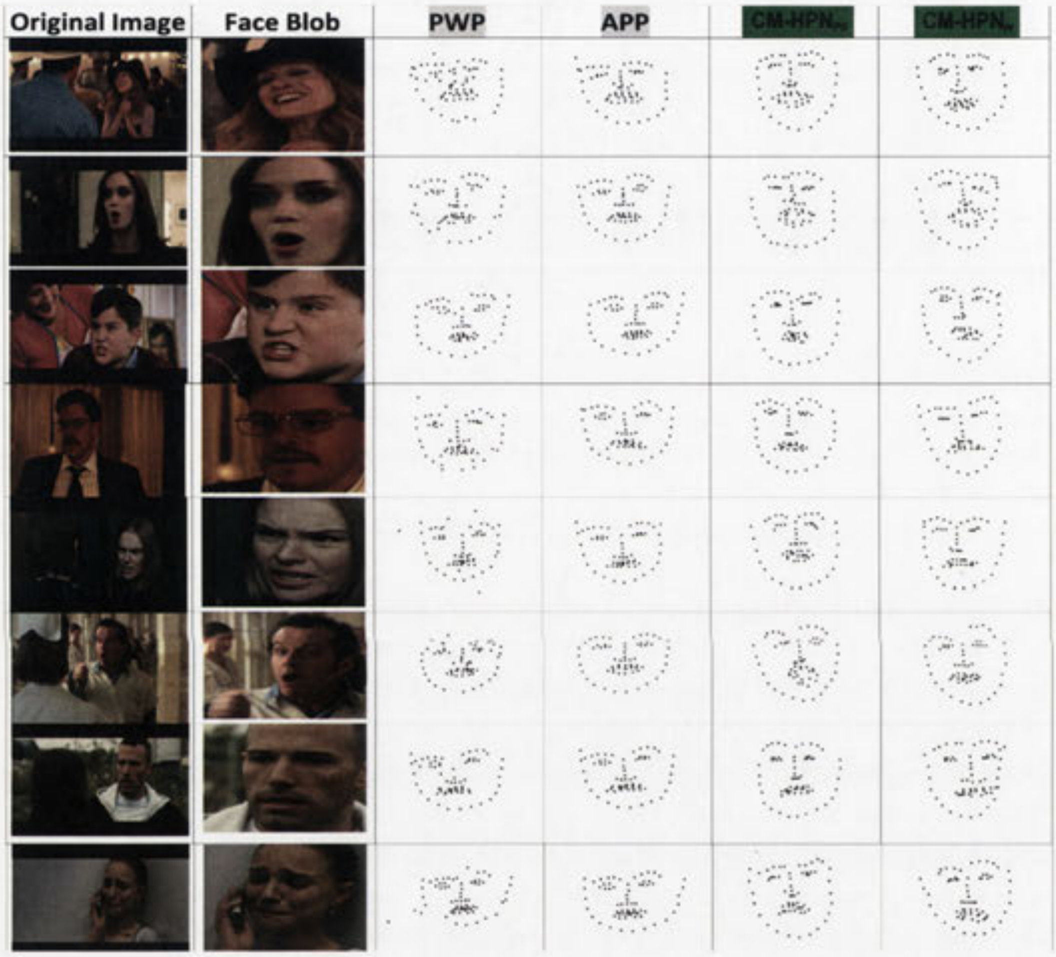


Figure 4.4: *Performance on selected SFEW images of the points and confidence maps based methods: PWP Asthana et al. [2009a], Rudovic et al. [2010b], APP Rudovic and Pantic [2011], CM-HPN$_{PS}$, CM-HPN$_{PI}$ regression.* The left most column is the input image, second column is the face area found by VJ face detector. Columns 3-6 show the reconstructed landmark points for the face. Please note that for SFEW, only qualitative experiments are possible as there is no ground truth for frontal facial landmark points.

accurate for the eyebrows. On further investigation, it is found that these errors are due to the error induced by the regression method, when the score of a non-frontal model's frontal reconstruction is higher than the original non-frontal model's reconstruction score. This could be corrected by applying efficient normalisation (for example: setting the mean to 0 and variance to 1) to data before regression. Twin-GPR is a generic structured regression model, the performance of the framework can be improved by using the class of structured SVM regression algorithms, which are problem specific. Further, based on the part sharing formulation, the method can be easily extended to **continuous pose normalisation** by sharing regression models among parts in neighbouring poses. As continuous pose normalisation is not the focus of this chapter, it is only briefly described as an algorithm in **Appendix** B, to show that it is feasible.

A user survey is performed on SFEW, where 15 subjects were asked to rate the expression preserving ability of HPN for the 4 methods (PWP, APP, CM-HPN$_{PS}$ and CM-HPN$_{PI}$) on a scale of 1 (poor) - 5 (excellent). CM-HPN$_{PS}$ and CM-HPN$_{PI}$ achieved mean values of 3.2 and 2.7 and standard deviation of 1.2 and 1.1, respectively. This is better than the ratings of PWP and APP, whose mean values are 1.9 and 2.5 and standard deviation values are = 1.1 and 1.2), respectively. Performing an ANOVA confirms that the result is statistically significant with $p < 0.0001$.

4.7 Conclusions and Future Work

In this chapter, a new HPN method called Confidence Map based HPN is proposed. The method is embedded in the pictorial structure inference step and has no explicit dependency on facial parts location. This makes the method especially suitable for images in real-world conditions as facial parts detection in real-world images is an open problem. The use of a shape prior on reconstructed maps by applying a facial shape constraint is further proposed. Moreover, enumerating over different poses allows the proposed algorithm to work without any prior head pose information.

The results on the Multi-PIE database show the effectiveness of the methods in comparison to other state-of-the-art points based methods. The generalization capability of the methods is demonstrated on an 'in the wild' database: SFEW, by using pre-trained models from the Multi-PIE database in qualitative experiments. It is important to note that the images in SFEW are taken in different environments as compared to Multi-PIE which was recorded in controlled laboratory settings.

The points based approaches only provides geometric features, which are not appropriate for problems such as micro-expression and facial action unit analysis. The proposed methods provide both geometric and texture information. Therefore, as part of future work, the texture descriptors obtained as part of HPN will be extended and experimented with. Further, robust regression methods such as the structured support vector regression will be used for confidence map normalisation such that the error induced by generic Twin-GPR can be reduced.

To summarise till now, in Chapter 3, the focus is on how to construct facial expression databases representing different challenging real-world conditions. From the learning in Chapter 3 that MoPS framework is the current state-of-the-art face and fiducial points detector on AFEW, a HPN method based on MoPS framework is proposed in Chapter 4. In the next chapter, based on fiducial points a geometric descriptor is proposed. The descriptor of two face is compared for their structural similarity for finding similar facial expressions. The problem solution is based on fiducial points and is applied to the problems of: (a) *facial expressions based album creation*; (b) *creating album by facial expression search* and (c) *finding image candidates for facial performance transfer.*

Chapter 5

Similar Facial Expressions and Applications

Facial expression analysis has many applications in HCI systems. Gaining insight into the state of the user's mind via facial expression analysis can provide valuable information for affective sensing systems. There has been a considerable amount of research exploring various methods for facial expression analysis, which mainly revolves around classification of query images into classes either via supervised or unsupervised algorithms. As discussed in 2.4.1, that the classic universal expressions are not always observed in daily life and are some times ambiguous, hence in this thesis discrete classes are chosen. Discrete expression classes are not complicated to label as compared to continuous labelling. The focus in this thesis is analysing facial expression in challenging real-world conditions only without having to worry about the continuous labelling complexity. There are also many multimedia applications such as facial expression transfer, image search, ranking by expression, browsing and album creation, which require finding the most relevant facial expression image. In these cases, the system requirement is to find the most similar facial expression image. The work presented in this chapter has been published in Dhall et al. [2010], Dhall et al. [2011b] and Asthana et al. [2012].

This chapter proposes a novel method for finding the most similar facial expression. Instead of the regular L2 norm distance, the use of the Structural SIMilarity (SSIM) Wang et al. [2004b] metric for similarity comparison as a distance metric in a nearest neighbour unsupervised algorithm is investigated. The feature vectors are generated using AAM. The contributions are described in Section 5.1. Section 5.2 presents the proposed approach in detail. The proposed geometric descriptor Expression Image (EI) is discussed in Section 5.2.2. SSIM is described in Section 5.2.3. The proposed technique is applied to *facial expression analysis based album creation* (Section 5.3) and *album by similar facial expression* (Section 5.4). Moreover, the technique is extended and used for finding corresponding facial expression images across two or more subjects, which is useful in applications such as *facial animation* and *automatic expression transfer* (Section 5.5). Person-independent facial expression performance results are shown on the Multi-PIE Gross et al. [2008a], FEEDTUM Wallhoff [2006] and AVOZES Goecke [2004] databases. The performance of the SSIM metric versus other distance metrics in a nearest neighbour search for finding the most similar facial expression to a given image is discussed. Finally, Section 5.6 provides the conclusions and future work.

5.1 Contribution

The contributions of this chapter can be summarised as follows:

- A novel method of facial expression analysis using SSIM is proposed. SSIM is used as the distance metric for NN classification after an efficient expression feature vector is created from fitting a person-independent AAM to the face image.

- The proposed approach is compared with various distance metrics in an unsupervised algorithm and experiments are performed on the Multi-PIE Gross et al. [2008a] and FEEDTUM Wallhoff [2006] facial data corpora.

- It is described on how the proposed technique can be extended and used for locating corresponding facial expression images across two or more subjects for the task of facial animation and automatic facial expression transfer. The results are experimentally tested on the AVOZES Goecke and Millar [2004] data corpus.

In Sohail and Bhattacharya [2007], the performance of kNN and weighted kNN for facial expression recognition using the L2 norm distance is discussed.

There has been some work on finding the similarity among two images. The SSIM Wang et al. [2004b] is one such method. Recently, the SSIM has been used for applications such as face recognition, image quality matching, face range matching etc. (e.g. Khwaja et al. [2010]). Nearest neighbour (NN) classification is one of the oldest machine learning techniques but is still used extensively for its simplicity and performance. The majority of the new work on nearest neighbour is based on learning the distance metric such as Largest Margin Nearest Neighbor (LMNN) Weinberger and Saul [2009] and Nearest Component Analysis (NCA)Goldberger et al. [2004].

5.2 Approach

The facial expression analysis system starts with extracting the AAM shape vector after AAM fitting. The shape vector is then used to construct an intermediate representation EI. Then, the system classifies examples using SSIM as a distance metric in kNN. From here on, the technique is referred to as SSIM-NN. All steps are explained in detail below.

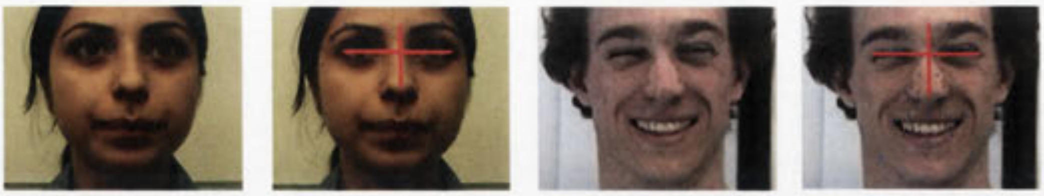


Figure 5.1: First and third are the input images. Second and fourth images show the normalization Euclidean distances taken, the vertical and horizontal red bars represent the vertical and horizontal distances.

5.2.1 Face Localisation and Fitting.

The first step is to locate the face. The VJ Viola and Jones [2001] face detector is applied, which gives the location of the face. This is used as an initialisation seed point for AAM Cootes et al. [1998] fitting. In this work, the AAM fitting method described in Saragih and Goecke [2009] is used for its speed and accuracy. The fitting process produces the shape vector s, which is then used to construct the feature vector for facial expression analysis in the system.

5.2.2 Expression Image

The EI is a visual descriptor map, which depicts the facial expression of the face in a normalised manner. In this chapter the suitability of SSIM is explored for facial expression analysis. SSIM operates on images and hence EI is proposed, which describes the facial expression. AAM fitting results in a row vector P containing location of each of the n landmark points.

$$P = [x_1; y_1 \dots x_n; y_n] \tag{5.1}$$

P is then normalised by taking the horizontal Euclidean distance between the outer eye corners on the left and right side. The vertical distance is the Euclidean distance between the tip of the nose and the midpoint between the eyebrows. Normalisation is required here to project the faces/points into a canonical frame for comparison. The choice for this normalisation is driven by the static nature of these points with respect to facial expressions. Figure 5.1 depicts the normalisation step for two sample face images. Please note that the landmark points chosen for calculating the Euclidean distances do not contribute to the motion within the face, their position is static with respect to expression change. Further, an undirected graph G is defined as

$$G = (V, E) \tag{5.2}$$

The graph observes the following properties: V is a finite, non-empty set and V

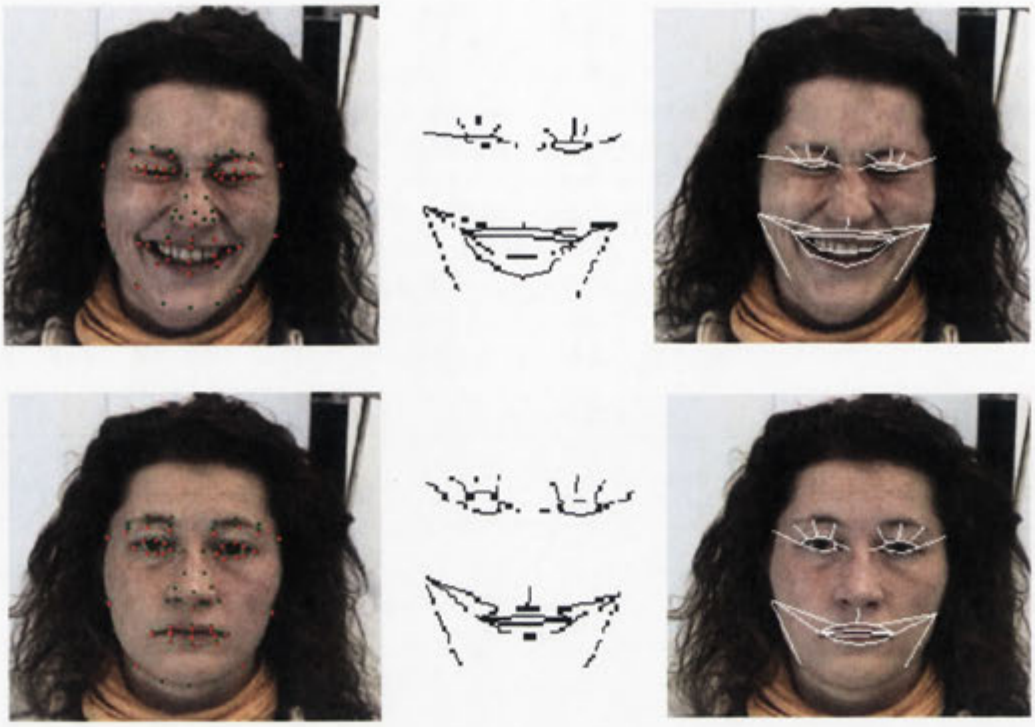


Figure 5.2: Row 1 and Row 2 depict the steps for EI formation for two images (FEEDTUM Wallhoff [2006]) of the same subject with different expressions: Happy and Sad, respectively. From left to right are the input image with AAM tracked landmark points (landmark points used for EI have been marked in red), the EI and the EI overlaid onto the original image.

represents the vertices of G. It describes m $(m < n)$ landmark points, which are used for defining the EI. E is a finite set of sets and is called edges of G. Each element of E is a set that is comprised of exactly two vertices (landmark points). From Equation 5.1 and 5.2, an undirected graph G^P can be defined as

$$G^P = (V^P, E^P) \tag{5.3}$$

V^P is a finite, non-empty set. The vertices V^P are the physical location (x, y) derived from P corresponding to the V of G. E^P is a finite set of edges derived from E of G. EI is then defined as a binary visual map derived from G^P by drawing the graph as an image. The edges of the graph are similar to distance vectors.

Graph G is prepared offline. The choice of specific landmark points and the corresponding distance vector image is derived from two motivations. Firstly, if the EI were to also include the landmark points on the face contour (chin), this would lead to a bias towards a grouping of images that have the same subject (due to similar face shape). The aim here is different. It is to find similar facial expressions rather than images of the same person. Hence, a balanced number of landmark points, which contain sufficient information for representing the facial expression, are chosen. The landmark points for constructing EI are chosen empirically, based on how person-independent the SSIM comparison could become using EI. Secondly, SSIM works on images, hence, EI are created from the chosen landmark points. Figure 5.2 describes the process of EI formation for two sample images of the same subject with different expressions. Row 1 shows a *Happy* and Row 2 is a *Sad* expression. Please note that the EI descriptor images formed on the right hand side depict the facial expressions only and have lost any subject identity information, which may bias the classifier towards similar faces, rather than similar expressions. The landmark points highlighted in red colour have been used for EI formation. The EI is not a vector of landmark positions; it is a high-level visual representation of the facial expression, which is constructed from chosen landmark points and their distance vectors.

5.2.3 SSIM

Once the EI are formed for all images, the next task is to compare them for their similarity. The SSIM Wang et al. [2004b] is used as the distance measure for expression grouping. It is a technique of calculating similarity among two images. The SSIM performs three different similarity measurements of luminance, contrast and structure, and thereafter combines them to obtain a single number. It calculates luminance and contrast first and isolates structure from these two factors. Please refer to Wang et al. [2004b] for further detail. The SSIM metric between two windows w_1 and w_2 on the same size $N \times N$ is given by:

$$SSIM(w_1, w_2) = \frac{(2\mu_{w_1}\mu_{w_2} + c_1)(2\sigma_{w_1w_2} + c_2)}{({\mu_{w_1}}^2 + {\mu_{w_2}}^2 + c_1)({\sigma_{w_1}}^2 + {\sigma_{w_2}}^2 + c_2)} \tag{5.4}$$

where μ_{w_1} and μ_{w_2} are the average of w_1 and w_2 respectively. ${\sigma_{w_1}}^2$ and ${\sigma_{w_2}}^2$ are the variance of w_1 and w_2 respectively. $\sigma_{w_1w_2}$ is the covariance between w_1 and w_2. Here, $c_1 = (k_1L)^2$ and $c_2 = (k_2L)^2$ are the variables to stabilize the division with a weak denominator. Here $k_1 << 1$ and $k_2 << 2$. L is the dynamic range of the pixel values.

The SSIM is symmetric ($S(w_1, w_2) = S(w_2, w_1)$), bounded ($-1 \leqslant S(w_1, w_2) \leqslant 1$) and has a maximum ($S(w_1, w_2) = 1$ iff $w_1 = w_2$). This value decreases as the images start to differ up to a minimum value of -1. Despite its simplicity, the SSIM index performs remarkably well across a wide variety of image and distortion types Wang et al. [2004b]. It has shown to perform better than traditional distance metrics such as L2 norm Wang et al. [2004b]. The disadvantage of SSIM is its computational complexity and sensitivity to translation and rotation.

SSIM is used as a way of comparing two EIs, which simply visualise the expression vectors. If the EI structures of normalised mouth and eye regions are similar, then the SSIM gives a higher value. The default implementation of SSIM uses a Gaussian window for matching. Empirically, it is found that the disk shaped convolution window performed better than other shapes, including the Gaussian window.

5.2.4 SSIM vs CW-SSIM

Wang and Simoncelli [2005] also presented the Complex Wavelet Structural Similarity Index (CW-SSIM), which is more robust to translation and rotation than the SSIM metric. It uses a complex version of a steerable pyramid transform for decomposing the image, which results in the translation and rotation invariant wavelet. In this work SSIM is preferred over CW-SSIM because the EI is already global translation, rotation and scaling invariant. Therefore using CW-SSIM would only increase the computational complexity of the proposed system. In Kakadiaris et al. [2006], invariance to facial muscle movement changes is mentioned as the reason for using CW-SSIM. It makes the face recognition more robust toward local changes due to the facial expression change. In contrast, the local translation due to muscle movements, which constitute a facial expression change is required for the problem tackled in this chapter.

5.2.5 Nearest Neighbour Search

The Nearest Neighbour classification algorithm is one of the oldest non-parametric instance based algorithm for classification. For every input I, the algorithm finds k most similar examples from the labelled training set and assigns the class, which is most represented by the resultant nearest neighbours. As discussed in Boiman et al. [2008], the primary reason why non-parametric classification methods, such as the NN, perform inferior in comparison to the parametric classifiers is due to the quantisation of the image descriptors, which leads to a larger error in the case of highly discriminative descriptors. The parametric classifiers compensate for this during the training phase but this leads to poor performance in non-parametric classifiers. The SSIM metric is used as the similarity distance metric between two EI, as it capitalises well on the structural information of the face. Furthermore, the high-level descriptor EI captures the highly discriminative aspects of the shape of the expression (the mouth and the eyes) well. This leads to comparable performance when the proposed technique is benchmarked against some of the state-of-the-art methods (see Sec. 5.2.6).

5.2.6 Experimental Validation

In the experiments for validating SSIM-NN, the Multi-PIE Gross et al. [2008a] and FEEDTUM Wallhoff [2006] databases are used. In the first data set, 260 images of 125 different speakers comprising of 81 males and 44 female subjects are chosen randomly from the Multi-PIE database. For the training set, 160 images of 63 speakers constituting 47 male and 16 female subjects are labelled for four different expression

Emotion	Surprise	Happy	Neutral	Sad
Surprise	78.2%	8.9%	8.6%	4.3%
Happy	6.2%	75.0%	12.5%	6.3%
Neutral	0.0%	4.0%	92.0%	4.0%
Sad	0.0%	5.0%	10.0%	85.0%

Table 5.1: Classification accuracy of SSIM-NN on EI generated from Multi-PIE images. Here the row is the true class and column is the assigned class.

categories viz. *Surprised*, *Happy*, *Neutral* and *Sad*. For the test set, 100 images of 62 subjects containing 34 male and 28 female subjects are chosen. The subjects in the test and training data sets are mutually exclusive. The iterative-discriminative AAM model Saragih and Goecke [2009] is trained on the Multi-PIE dataset and the shape vector are extracted for the test and training images. Table 5.1 displays the confusion matrix for SSIM-NN performance on the Multi-PIE database. Figure 5.3 shows some output of SSIM-NN on Multi-PIE data corpus. Here, Row 1 is the input and Row 2 is the corresponding most similar facial expression. For comparison, the Largest Margin Nearest Neighbour (LMNN) algorithm Weinberger and Saul [2009] is compared with SSIM-NN. LMNN learns a Mahalanobis distance metric over the labelled training set. Two different forms of input vectors are tested, the first one constitutes the *landmark points* chosen for building EI (this is referred to as LMNN-1) and the second one constitutes of the *distance vectors* between these EI points (LMNN-2). SSIM-NN is also compared with kNN with L2 norm distance and the single Gaussian Naive Bayes (NB) classifier. Table 5.2 displays the comparison of classification accuracy of these approaches. The SSIM-NN classifier performs better than the others, as it compares the structure and the EI retains the structural perceptive aspect of the local moments of the face, which result in facial expressions. The performance of any classification algorithm is highly affected by the type of descriptor. This can be stated a possible reason for average performance of LMNN.

As the SSIM distance metric calculation uses a set of convolutions, there is scope

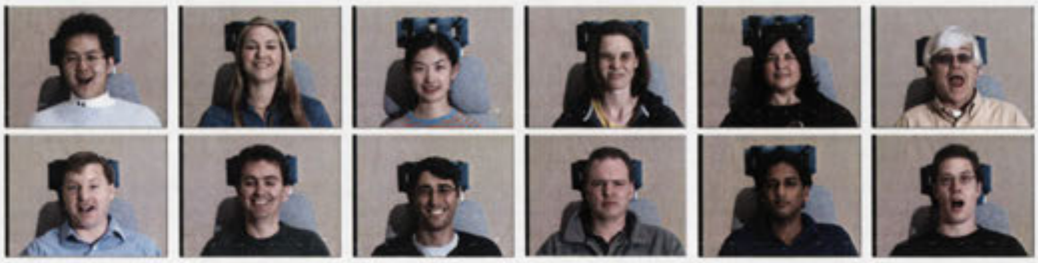


Figure 5.3: Some outputs of the SSIM-NN technique for k=1 on images from Multi-PIE data corpus. Row 1 are the input images and Row 2 are the corresponding most similar expression images.

Algorithm	SSIM-NN	LMNN-1	LMNN-2	KNN-L2	NB
Performance	82.0%	54.0%	56.0%	50.0%	52.0%

Table 5.2: Classification accuracy comparison of the SSIM-NN, LMNN-1, LMNN-2, KNN-L2 and NB classifiers.

of parallelisation via a GPU, which is becoming common in modern computers. SSIM based on the Jacket Jac [2009] Matlab plug-in is tested on a NVIDIA GT230M chipset. This gives a 10-12x computational performance gain compared to an Intel Core2Duo @ 2.53GHz. Without GPU, SSIM takes 3.46 seconds on average. With parallelisation it takes on average 317 milliseconds.

5.3 Facial Expression Based Album Creation

With the advancement in sensor technology, users are collecting millions of images. Due to this increase, searching, browsing and managing images in multi-media systems has become more complex. One solution to this problem is to divide images into albums for meaningful and effective browsing. A novel automated, expression driven image album creation for consumer image management systems based on SSIM is proposed. The system groups images with faces having similar expressions into albums. The method constitutes of four major steps, which are described in the following sub sections. Figure 5.4 depicts the flow of the system.

5.3.1 Semi-unsupervised Album Creation

For image album formation, the K-Means algorithm is applied over the expression data. The K-means clustering algorithm splits a set of observations into subsets by minimizing the intra-cluster variation. The number of image albums k serves as the initial number of clusters for K-Means clustering algorithm where the distance metric is SSIM. Therefore, the clustering becomes:

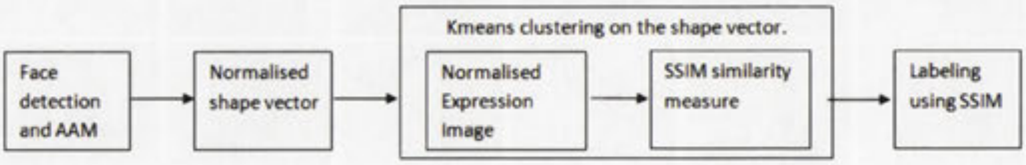
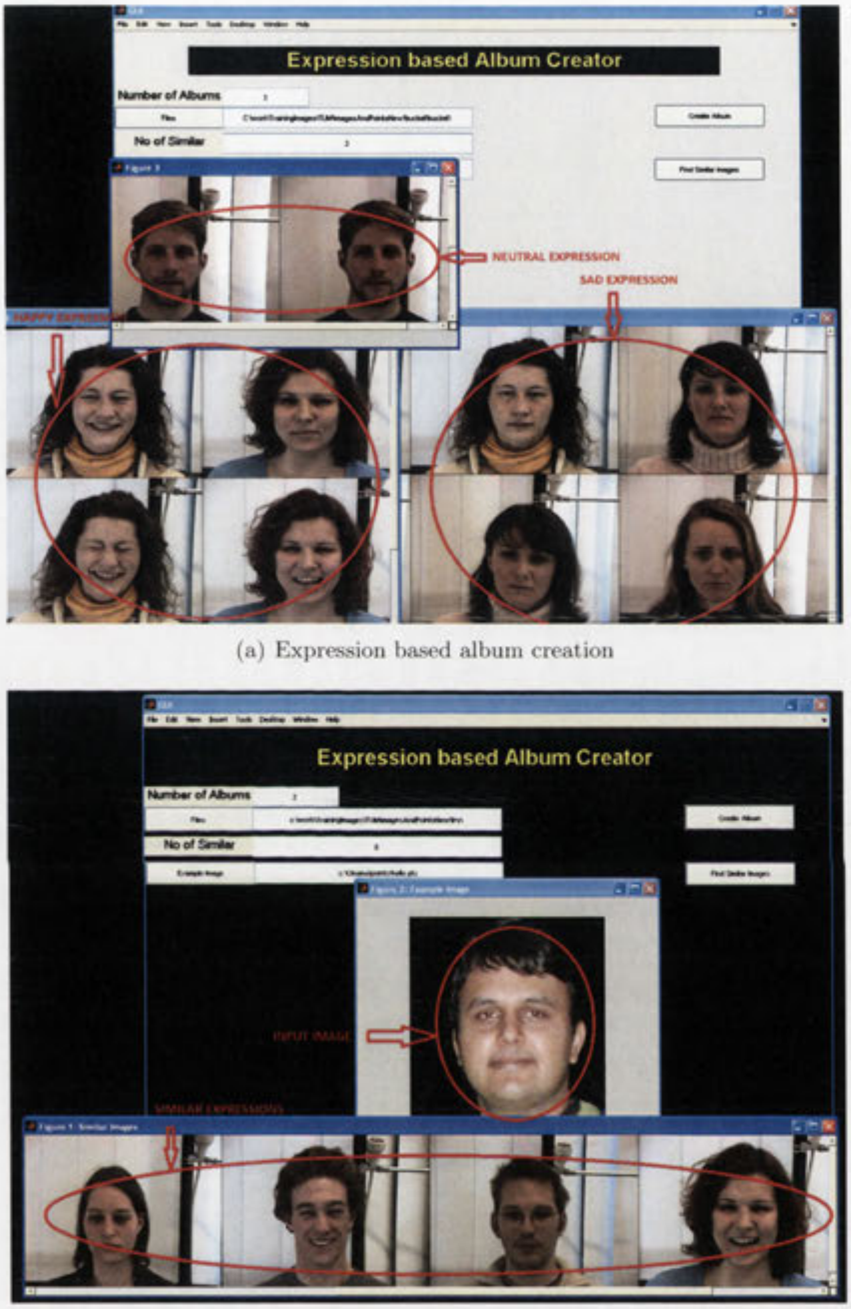


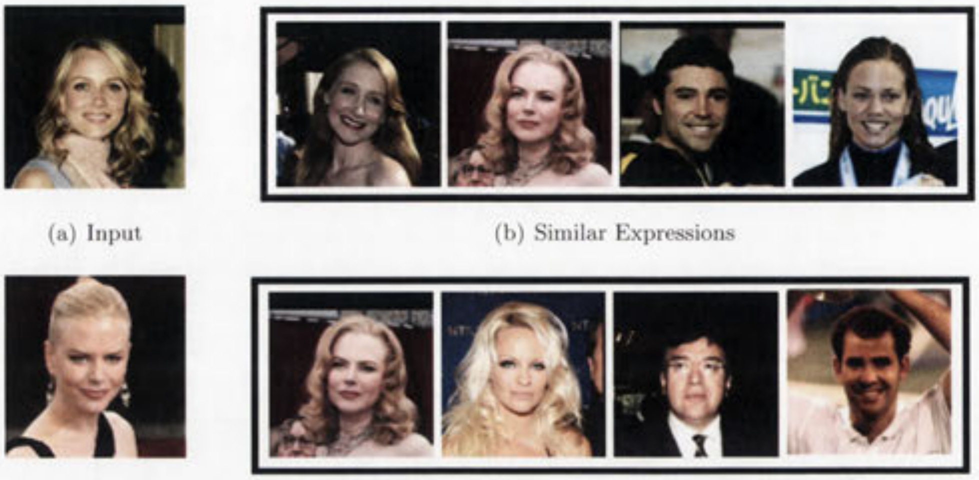
Figure 5.4: Block diagram of the facial expressions based album creation system Dhall et al. [2010].



(a) Expression based album creation

(b) Album by similar expression

Figure 5.5: The figures (a) & (b) are snapshots of Matlab based prototypes for expression based album creation and album by similar expression Dhall et al. [2010].



(a) Input

(b) Similar Expressions

(c) Input

(d) Similar Expressions

Figure 5.6: Album by similar expression example, (a) and (c) are the input images. (b) and (d) are the corresponding similar expressions.

$$\arg_{\mathbf{S}} \min \sum_{i=1}^{k} \sum_{x_j \in \mathbf{S}_i} \mathbf{SSIM}(\mathbf{x}_j, \mu_\mathbf{i}) \tag{5.5}$$

where $(x_1, x_2, .., x_n)$ are the Expression Images EI and μ_i in S_i is the mean EI. The clustering is done on the normalise landmark points of the shape vector, which are used to construct the distance vectors of EI. This is done to keep a low dimensionality during clustering. Though the distance comparison is calculated on EI. The mean EI representing each clusters are then compared using the SSIM distance metric with pre-stored labelled EI. The pre-stored EI are labelled into four expressions (*Happy*, *Neutral*, *Sad*, and *Surprise*). This leads to automatic labelling of the albums into the fundamental expression classes. Once a new image arrives it is added to the existing albums via comparing its closeness to the mean image representing the respective albums. Figure 5.5(a) shows the Matlab based prototype of the clustering technique.

5.4 Album by similar expression

A user may be interested in finding images, which have the expression similar to a specific facial expression. In this case, the user provides the system with one example image. The user also specifies the number of similar images, which decides the size of this album. The system extracts the EI for the example and the group of images. The example EI is then compared using SSIM with all other images in the group. The

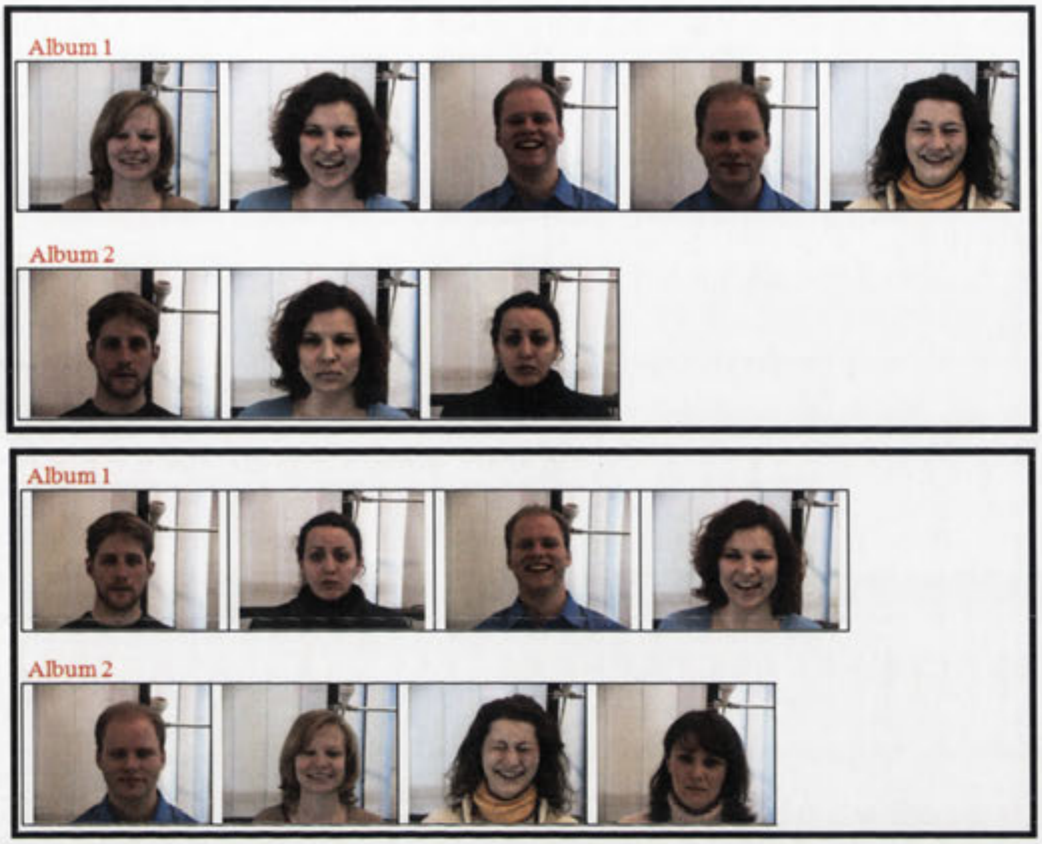


Figure 5.7: Sample result on images from the FEEDTUM Wallhoff [2006] database after executing the proposed system (upper box) and the Fuzzy clustering algorithm (lower box) Bezdek [1981].

similarity distances are then sorted and the user desired number of similar images is selected as an album with respect to the relevance. Figure 5.5(b) shows the snapshot of a Matlab based prototype of the album by example system. Figure 5.6 depicts two examples of this function.

5.4.1 Perceptual Experiment and Outputs

Since different users may have different perception about an expression, analysing the correct clustering performance is a non-trivial task. To validate the performance a test set of sixty images from the FEEDTUM Wallhoff [2006] and LFW Huang et al. [2007] databases is created. A total of fifteen users are asked to judge the album creation performance by finding the the images, which seem to have a different expression (as compared to their cluster) and do not belong to the album created. The average total error classification rate (based on images for which the users perceived the facial expression different from the common facial expression of the cluster, to which the image belonged) came out to be 13.7%.

The Figure 5.5(a) displays the experimental GUI of the system. In 5.5(a) the images are from the FEEDTUM database Wallhoff [2006]. The three sub windows in the figure are the albums created after SSIM based clustering. Please note that the system groups faces of similar facial expressions into one set.

Figure 5.5(b) depicts album by similar expression example. The user inputs an image, which contains an example expression. The user also specifies the number of similar images desired in the album. This input serves as the number of similar images to be presented. The larger ellipse shows the searched similar images and the upper one is the user input image.

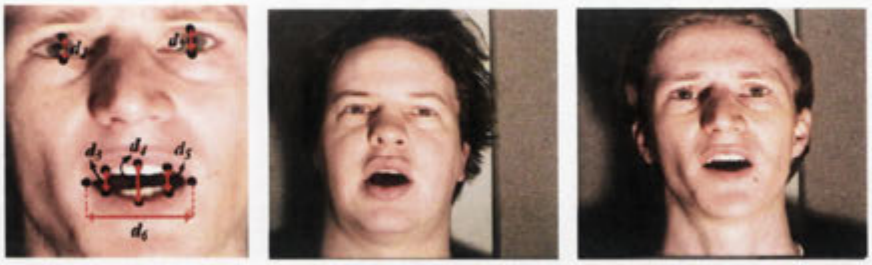
Figure 5.6, shows the results of a similar experiment on images from the LFW database Huang et al. [2007]. Figure 5.6(a) is the user example input with a happy expression. Figure 5.6(b) depicts the album of top matching expressions with decreasing relevance from left to right. Similarly, Figure 5.6(c) shows a face with smiling expression and Figure 5.6(d) are the similar expressions.

The proposed album creation method is also compared with the fuzzy clustering algorithm. Figure 5.7 shows the outputs of the systems.

5.5 Locate Corresponding Facial Expression Images

In this section, the application of SSIM-NN to solve the problem of finding corresponding facial expression images for the application of facial expression transfer is described. The issue of transferring facial expressions from one person's face to another one's has been an area of interest for the movie industry and the computer graphics community for quite some time. The goal is to enable meaningful facial expression transfer. 'Meaningful' here means taking the personal characteristics into account. For example, if the source subject exhibits large facial movements while the target subject normally shows little, it would look unrealistic if the source's large movements were transferred directly 1-1 to the target face. Various approaches, such as Blanz et al. [2003], Chang and Ezzat [2005], Noh and Neumann [2001], Song et al. [2006], Zhang et al. [2006], have been proposed for facial expression transfer/cloning. Generally, all these approaches require a set of corresponding facial expression images across both source and target subject to create a model that can be used for automatic expression transfer.

The aim is to find the set of closest corresponding expression images between the source and target subject, taking into the account certain personal characteristics and mannerisms. For example, it is possible for two people to exhibit a different amount of lip movement (for example Figure 5.8(b) and Figure 5.8(c), while saying the same word. The intent is to capture this difference while extracting the corresponding EI, so



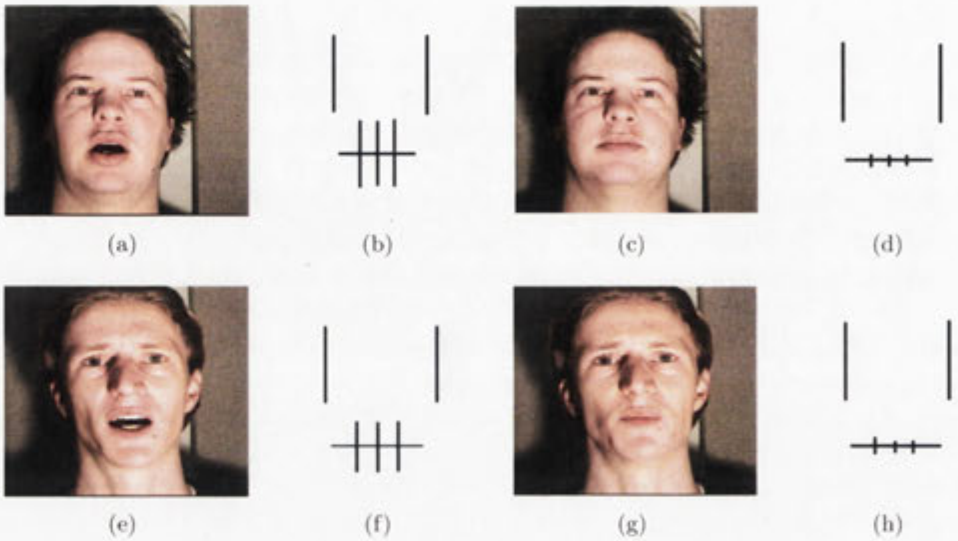
(a) Normalised distance (b) (c)

Figure 5.8: (a) depicts the extracted distance features. (b) and (c)are the manually selected source and target reference images.

that the model trained on this set of images exhibits a high degree of realism. This can be done by using methods, such as Sumner and Popović [2004], that allow establishing a correspondence between the meshes of any two arbitrary objects. Since this method is designed for general meshes, it needs manual intervention and is quite tedious Song et al. [2006]. The problem of finding the correspondence between the faces of two different subjects is simpler owing to the structural similarity between human faces and the constrained nature of the face models Song et al. [2006]. A simple extension of the SSIM-based approach in Section 5.2 efficiently finds the closest correspondence images with minimal manual intervention and can be used directly for collecting the training data for facial expression transfer systems such as Blanz et al. [2003], Chang and Ezzat [2005], Noh and Neumann [2001], Song et al. [2006], Zhang et al. [2006] is presented.

The landmark points of the shape vector obtained from AAM fitting are used to construct the *distance feature* (Figure 5.8(a). The distance features are extracted from the facial features, such as the mouth and eye regions, since they are the region of interest for the experiments presented in this section. However, this method could be extended to cover other facial features such as eyebrows and forehead etc., as the landmark position of these facial features is available in the AAM fitted shape vector.

Step 1 - Selecting reference images: Reference images are required for taking base distances for normalisation. A reference image, exhibiting the same expression, for the region of interest of the source and target subjects is manually selected. In the experiments presented below (Section 5.5.1), an images that exhibits open mouth expression with both tongue and teeth visible is selected as a reference image. Figures 5.8(b) and 5.8(c) show the reference images for the source and target subjects. Before extracting the *distance feature* from the reference images, both of these reference images are similarity normalised, assuming the eye corners as the reference points. The reference *distance feature* is represented as $\mathcal{D}^R = \{d_1^R, \ldots, d_n^R\}$ where $n = 6$ (see Figure 5.8(a)).


(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 5.9: [(a):(b)], [(c):(d)], [(e):(f)] and [(g):(h)] are the input images and the corresponding normalised EI images, drawn from the normalised distance vectors.

Step 2 - Extracting features from training sequences: As a preprocessing step, shapes of all the frames are aligned into a common coordinate frame with respect to the reference images via similarity normalisation, as done in the previous step. The *normalised distance features vector* is extracted from all the frames in the training sequences that will be used for finding the correspondence images

$$\bar{\mathcal{D}}_m = \left[\frac{d_1^m}{d_1^R};\dots;\frac{d_n^m}{d_n^R}\right]^T = \left[\bar{d}_1^m;\dots;\bar{d}_n^m\right]^T, \qquad m = 1\dots M \tag{5.6}$$

where $\bar{\mathcal{D}}$ is the *normalised distance features vector*, M is the total number of frames extracted from the training sequences and d_n^m is the n^{th} *distance feature* extracted from the m^{th} frame. **Step 3 - Finding candidate correspondence images:** From the training sequences, a set of candidate correspondence images of the target subject for every frame of the source subject using the *normalised distance features vector* is computed. Let $\bar{\mathcal{D}}_p^{\text{Source}}$ be the *normalised distance feature vectors* for the p^{th} source frame and $\bar{\mathcal{D}}_q^{\text{Target}}$ be the *normalised distance feature vectors* for the q^{th} target frame

$$\bar{\mathcal{D}}_p^{\text{Source}} = \left[\bar{s}_1^p;\dots;\bar{s}_n^p\right]^T \tag{5.7}$$

$$\bar{\mathcal{D}}_q^{\text{Target}} = \left[\bar{t}_1^q;\dots;\bar{t}_n^q\right]^T \tag{5.8}$$

These vectors are used to construct the normalised EI images described in the image (Figure 5.9). Next, the EIs of the source and the target frame are compared using SSIM-NN. The resulting most similar image is the best correspondence image for the

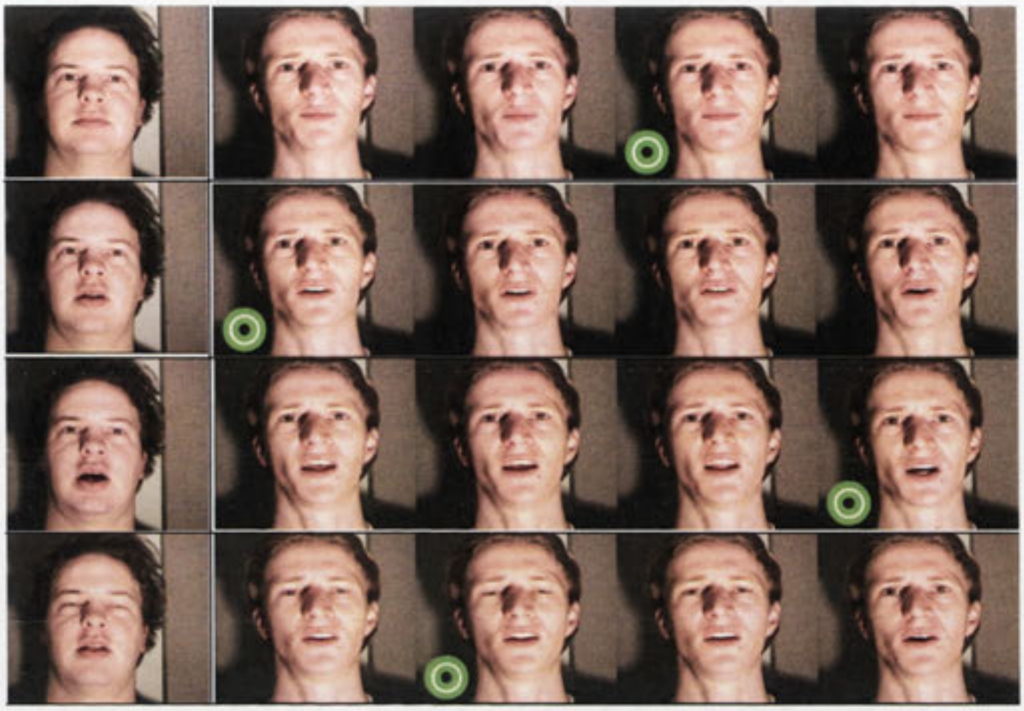


Figure 5.10: Example outputs of SSIM-NN for locating corresponding facial expression images on the AVOZES data corpus Goecke and Millar [2004]. Column one images are of the source subject and columns two - five are the four most similar images of the target subject as calculated by the SSIM-NN. The most relevant output which is manually selected out of the four results is marked with a green circle.

source frame in consideration. Hence, for every source frame $p = 1 \dots P$, where P is the total number of source frames in the training sequences, candidate correspondence images are selected target frames with the maximum similar result from the SSIM-NN. In the experiments, the value of k is set as k=4 for SSIM-NN.

Step 4 - Manually selecting the candidate set: In the previous step, the candidate correspondence images are selected considering only the shape of the face. The texture of the oral cavity is ignored. This can result in erroneous candidate corresponding images. This is particularly a difficult problem because different people have different amounts of oral cavity and teeth/tongue visible, while otherwise exhibiting similar expressions (Figure 5.8(b) and 5.8(c)). The candidate set for a source frame in consideration and the visually best correspondence image from the candidate correspondence images are selected manually.

5.5.1 Experiments and Comparison

Figure 5.10 shows the experimental results of the process described above. Column one are the source image frames and columns two-five are the four most similar expressions

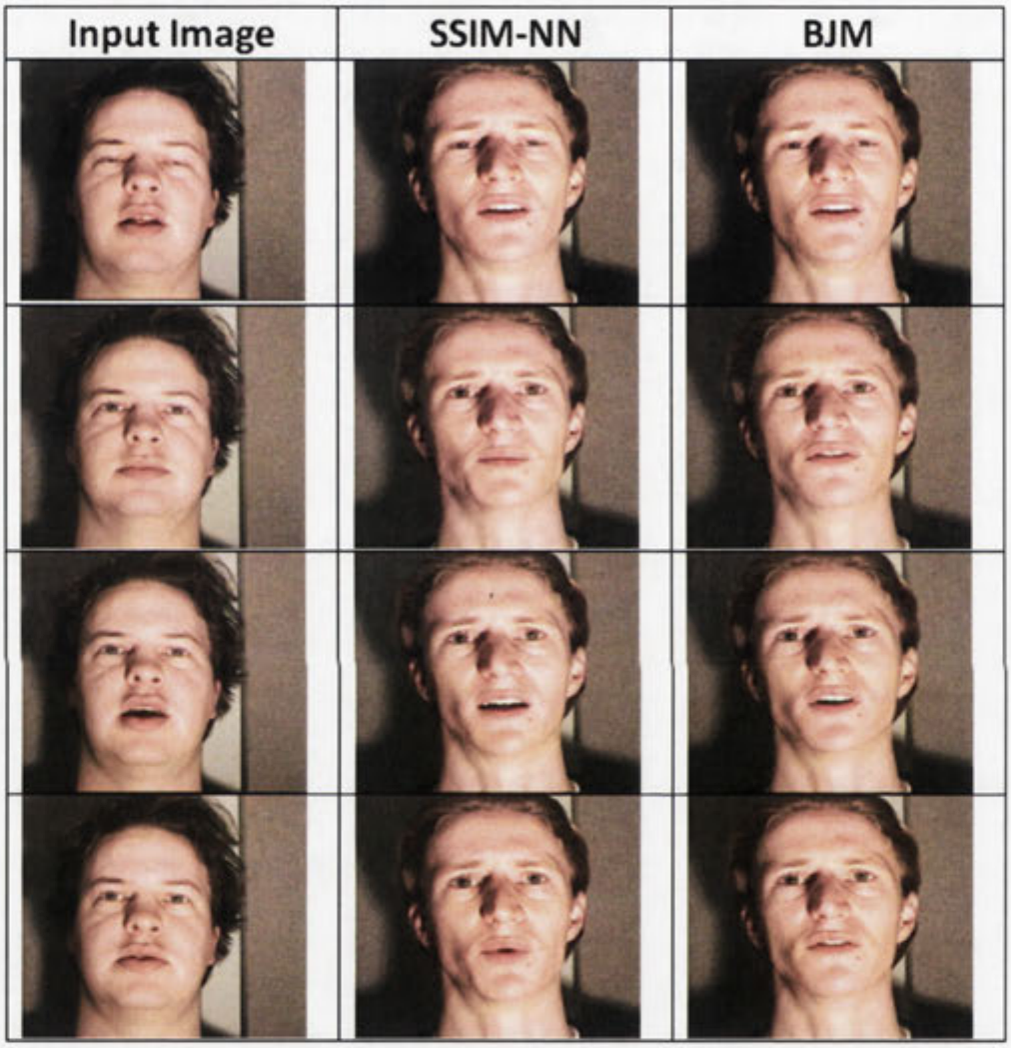


Figure 5.11: Column One are the input images and Column two and three are the corresponding expressions outputs from our algorithm and BJM Kemelmacher-Shlizerman et al. [2010]. The system performs similar to BJM and in some cases better. For example in Row 2,3 & 4 please observe the little difference in the lip shapes. Here k = 1 for SSIM-NN.

as calculated by SSIM-NN. The closest of the four is manually selected and has been marked with a green circle in the example. In row one, the source target has neutral expression and the found images have similar expressions. The most suitable is the third result located in the fourth column. In row three, the source exhibits large mouth opening and the corresponding images of the target speaker also have a large mouth opening. With manual selection out of the four results, we chose result number four located in column five. Note that as two people can say the same word differently, their maxima and minima of mouth movement will be different. In row four, the source subject closes eyes and the SSIM-NN returns similar expression and the closest match is found in its corresponding column three.

In Kemelmacher-Shlizerman et al. [2010], the authors presented a system for controlling a face of a subject using a gallery of images of the subject with an input face image. LBP Ojala et al. [2002] feature descriptor is extracted from faces. Further, χ^2 distance is computed between the input image and the images in the dictionary. This method Kemelmacher-Shlizerman et al. [2010] is referred to as BJM (Being John Malkovich) for further use. A system similar to BJM, for finding the corresponding expressions is implemented. Figure 5.11 compares the outputs from our system and BJM Kemelmacher-Shlizerman et al. [2010]. Please note the subtle in the shape of the lips as detected in the corresponding images in Row 2, 3 and 4.

5.6 Conclusions

This chapter discusses the usefulness of facial expression analysis to three different applications: (a) facial expressions based album creation; (b) creating album by facial expression search and (c) finding image candidates for facial performance transfer. The applicability of facial structure for finding similar facial expressions is explored. Lab based databases Wallhoff [2006] are used, so as to analyse the performance of the proposed method without the effect of error in fiducial points detection.

A novel method for finding similar facial expressions using SSIM is proposed. This work utilises the shape characteristics of the face and demonstrates the power of the NN classifier when an efficient discriminative image descriptor is used. The experiments are performed on the Multi-PIE and FEEDTUM databases. Furthermore, technique is described to find corresponding images, which is a central problem in applications such as the expression transfer, similar image search and clustering. Instead of commonly used machine learning techniques, the method capitalizes on the visual structure aspect of the expression.

It is interesting to note that structural information alone is useful for finding similar facial expressions. A major limitation of this work is dependence on accurate fiducial points detection. Accurate fiducial points detection on database such as AFEW (Chapter 3) can be error-prone, therefore similar facial expressions can be noisy on data from AFEW. Further for problems such as matching AUs, points alone can be insufficient. Texture information needs to be added to the current framework. Generally, in the problem of automatic album creation and for images collected from social events, there can be multiple subjects present in a scene. In the next chapter 6, a framework for analysis of expression of a group of people in an image is presented.

To summarise, in Chapter 3, the focus is on how to construct facial expressions database representing different challenging real-world conditions. From the learning in Chapter 3 that MoPS framework is the current state-of-the-art face and fiducial points detector on AFEW, a HPN method based on MoPS framework is proposed in Chapter 4. In the current chapter, the problem of structural similarity for finding similar facial expressions is evaluated. The problem solution is based on fiducial points. As discussed in the Section 1.1, along with lack of labelled data, presence of multiple subjects in a sample poses a significant challenge to FER in real-world conditions. In the next Chapter 6, a framework is proposed for handling the expression of multiple subjects in an image.

Chapter 6

Analysis of the Mood of a Group of People

Automatic facial expression analysis has seen much research in recent times. However, little attention has been given to the estimation of the overall expression theme conveyed by an image of a group of people. Analysing the theme expression conveyed by images of groups of people is an unexplored problem that has many real-world applications: image search, retrieval, representation and browsing; event summarisation and highlight creation; candid photo shot selection; expression apex detection in video; video thumbnail creation etc. In a recent 2012 Forbes article Caroll [2012] discusses the lack of ability based on use of context of the current image search engines. Information such as the mood of a group can be used to model the context. These problems where group mood information can be utilised are a motivation for exploring the various group mood models. One basic approach is to average the happiness intensities of all people in a group. However, perception of the mood of a group is defined by attributes such as where people stand, how much of their face is visible etc. These social attributes play an important role in defining the overall happiness[1] an image conveys. The contents of this chapter have been published in Dhall et al. [2012b] and Dhall and Goecke [2012].

6.1 Contribution of the Chapter

1. An automatic framework for mood analysis of a group of people in images based on the social context.

2. A weighted model is presented, which takes into consideration the global and local attributes, which affect the perception of the mood of a group.

[1] This chapter uses the terms 'mood', 'expression' and 'happiness' interchangeably for a group in an image.

3. A labelled 'in the wild' database containing images of groups of people is collected and compared with the existing databases. A semi-automatic process was followed to create the database.

The remainder of the chapter is organised as follows: Section 2.7 discusses the prior work in visual analysis of focussing on various aspects of the analysis of a group of people. Section 6.2 describes the problems and challenges involved in automatic Group Mood Analysis (GMA). The details of a 149-user survey investigating attributes affecting the perception of mood is discussed in Section 6.3. An 'in the wild' database collection method is detailed in Section 6.4. Section 6.5 discusses a basic GMA model based on averaging. Global context based social features are discussed in Section 6.6. Local context based on occlusion intensity is described in Section 6.7. The global and local context are combined and applied to the averaging approach in Section 6.8. The manual attributes are combined with data-driven attributes in a supervised hierarchal Bayesian framework in Section 6.9. Section 6.10 discusses the results of the proposed frameworks, including both quantitative and qualitative experiments.

6.2 Challenges

The following subsections discuss the challenges in creating an automatic system for mood analysis.

6.2.1 Attributes

Human perception of the mood of a group of people is very subjective. Kelly and Barsade [2001] argue that the mood of a group is composed by two broad categories of components: top-down and bottom-up. Top-down is the affective context, i.e. the effect induced by attributes such as group history, background, social event etc. Top-down has an effect on the group members. For example, a group of people laughing at a party displays happiness in a different way than a group of people in an office meeting room. From an image perspective, this means that the scene/background information can be used as affective context. Bottom-up component deals with the subject in the group. Attributes of individuals that affect the perception of group mood. The bottoms-up component defines the contribution of individuals to the overall group mood.

From now on, the top-down component is referred to as 'global context' and bottom-up component as 'local context'. There can be various attributes, which define these two components. For example, global context contains but is not limited to scene information, social event information, who is standing with whom, where are people standing in an image and with respect to the camera etc. Local context ,i.e. individual specific attributes cover individual's mood/emotion, face visibility, face size with respect to neighbours, age, gender, head pose and eye blink etc. To further understand these attributes, a perception user study is performed. The study and its results are detailed in Section 6.3.

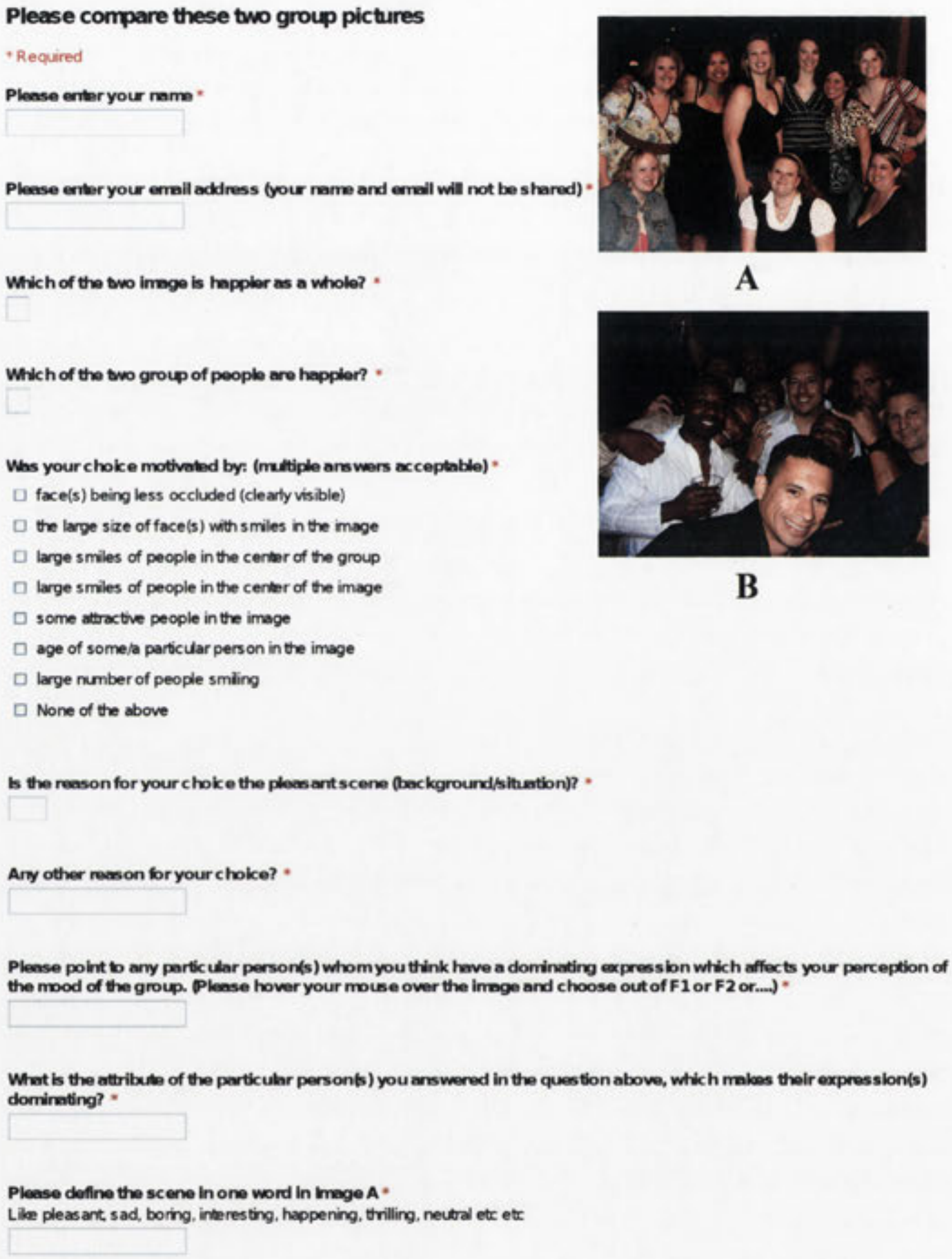
6.2.2 Data and Labelling

Data simulating 'in the wild' conditions is a major challenge for making emotion recognition methods work in real-world conditions. Generally, emotion analysis databases are lab-recorded and contain a single subject in an image or video. It is easy to ask people to pose in labs. However, anyone working in emotion analysis will attest to the difficulty of collecting spontaneous data in real-world conditions. For learning and testing a mood analysis system, labelled data containing groups of people in different social scenarios is required. Once the data is available the next task is labelling. According to Forgas [1992], moods are low-intensity, diffuse feeling states that usually do not have a clear antecedent. Mood can be positive/pleasant and negative/unpleasant. A problem defines the types of labelling required: discrete or continuous. The database proposed in this paper: HAPPEI is labelled for neutral to pleasant mood with discrete levels.

6.2.3 Visual Analysis

Inferring the mood of a group involves classic computer vision tasks. As a pre-processing step, the first challenge is face and facial point detection. Ideally, for a facial dynamics analysis system, one will want a subject independent facial landmark detector Chew et al. [2012]. Further Chew et al. [2012], argue that a subject dependent facial parts method, such as AAM, performs better than subject dependent constrained local models Saragih et al. [2009]. However, if a proper descriptor is used on top of previously aligned faces from a subject independent detector, the error in alignment can be compensated for. Moreover Zhu and Ramanan [2012], show the effectiveness of MoPS Zhu and Ramanan [2012] over CLM and AAM for facial landmark localisation when there is much head movement. Motivated by these arguments, the parts based model of Everingham et al. [2006][2] is used in the experiments (Section 6.10). The images in the proposed database HAPPEI, have been downloaded from Flickr. This creates the challenge of real-world varied illumination scenarios and difference face resolution. To overcome this, LPQ and PHOG are used in this work. LPQ is robust to varied illumination and PHOG is robust to scale Dhall et al. [2011a].

[2] Zhu and Ramanan [2012] report that their method works better than Everingham et al. [2006], however due to its computational complexity Everingham et al. [2006] is used here.



Please compare these two group pictures

* Required

Please enter your name *

Please enter your email address (your name and email will not be shared) *

Which of the two image is happier as a whole? *

Which of the two group of people are happier? *

Was your choice motivated by: (multiple answers acceptable) *

- □ face(s) being less occluded (clearly visible)
- □ the large size of face(s) with smiles in the image
- □ large smiles of people in the center of the group
- □ large smiles of people in the center of the image
- □ some attractive people in the image
- □ age of some/a particular person in the image
- □ large number of people smiling
- □ None of the above

Is the reason for your choice the pleasant scene (background/situation)? *

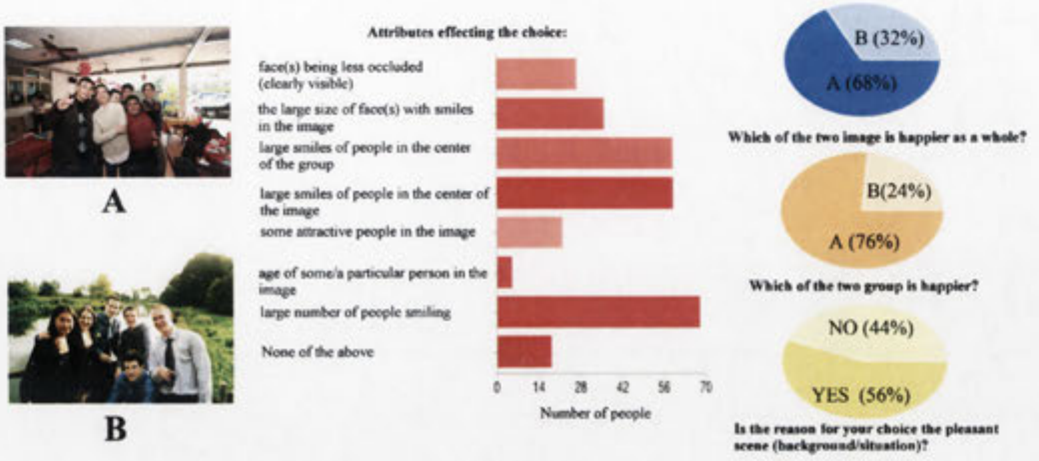
Any other reason for your choice? *

Please point to any particular person(s) whom you think have a dominating expression which affects your perception of the mood of the group. (Please hover your mouse over the image and choose out of F1 or F2 or....) *

What is the attribute of the particular person(s) you answered in the question above, which makes their expression(s) dominating? *

Please define the scene in one word in image A *
Like pleasant, sad, boring, interesting, happening, thrilling, neutral etc etc

Figure 6.1: Snapshot of the survey used for understanding the attributes, which affect the perception of the mood of a group.



(a) Case 1

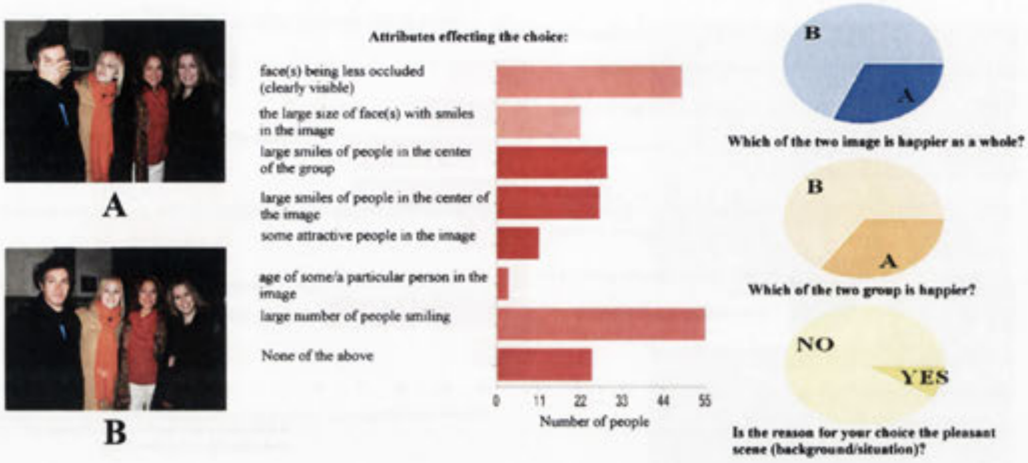
Figure 6.2: Results of the analysis of a comparative case in the survey (Section 6.3).

6.3 Survey

In order to understand the attributes (Section 6.2.1) which affect the perception of a mood of a group, a user study is conducted. Two sets of surveys are developed. In the first part (Figure 6.1), subjects are asked to compare two images for their apparent mood and rate the one with a higher positive mood. Further, they are asked various questions about the attibutes/reason, which made them choose a specific image/group out of the two images/groups. A total of 149 subjects participated in this survey (94 males and 55 female subjects). There are a total of three cases in the first survey, Figure 6.1 shows one such case of questions. Figures 6.2 and 6.3 describe the analysis of two cases in the survey. On the left of the figures, the two images to be compared are displayed.

The images in the survey are chosen on basis of two criteria: a) to validate a hypothesis - for example: from the earlier experiments in Dhall et al. [2012b], it is noticed that adding the effect of occlusion attribute to the model decreased the error (user perception vs model output). Therefore, in one case, two images shot in succession are chosen, in which one of the subjects covered his face in the first capture (Figure 6.3). It is interesting to note that a larger number of survey participants (69.0%) chose image B in Figure 6.3 as having a higher mood score (more positive mood on the scale of neutral towards thrilled) and out of these 69.0%, 51.1% chose 'faces being less occluded' as one of the reason. The other dominating attribute for their decision was the larger number of people smiling (54.6%). Both the attributes are correlated; it is easy to infer the expression of a person when the face is clearly visible.

In the other case (Figure 6.2), two random images were chosen. 76.0% participant



(a) Case 2

Figure 6.3: Results of the analysis of a comparative case in the survey (Section 6.3).

ranked Image A higher than B. 46.0% of all participants chose 'larger number of people smiling' as the most dominant attribute. There were 49.1% of the participants who selected Image A, 'larger number of people smiling' and 37.1% of the participants who selected Image B, 'larger number of people smiling'. 46.4% of participants who chose A, selected 'large smiles of people in the center of the group'. Also, 56.0% of the people chose the presence of a pleasant background as the reason for their selection. In another question: 'Any other reason for your choice?', participants pointed to the party like scenario in Image A (Figure 6.2), which made them to consider group in Image A being more happy. Other considerable responses were: body pose, people closer to each other in the group and spontaneous expression (i.e. when subjects are not explicitly posing in front of the camera) etc. For the question: 'Please define the scene in one word in Image A' and 'Please define the scene in one word in Image B', the majority of the participants defined the scene in context of mood such as 'relaxed', 'pleasant', 'interesting', 'happening', 'enjoyable'. Based on the user survey, in the following sections the attributes which are discussed are: relative location of members of a group, relative face size to estimate if a person is in the front or back, face clarity/occlusion and mood of a member.

6.4 Database Creation and Labelling

Popular facial expressions databases such as the CK+ Lucey et al. [2010], FEEDTUM Wallhoff [2006], GEMEP Bänziger and Scherer [2010], MultiPIE Gross et al. [2008a], VAM Grimm et al. [2008] are all 'individual' centric databases, i.e. contain a single subject only in any particular image or video. For the problem in this chapter, there

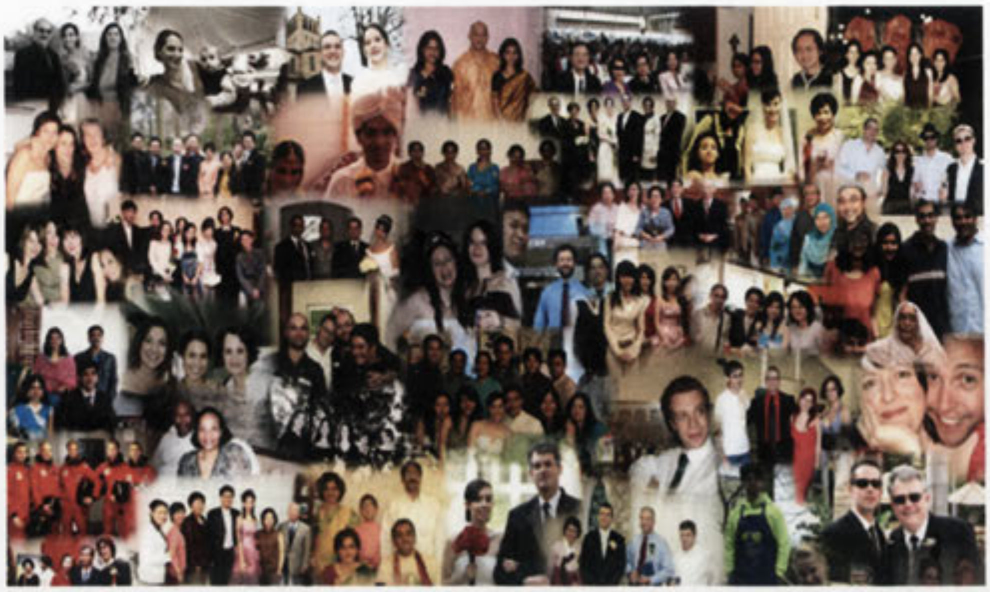


Figure 6.4: The image shows collage of various images in the HAPPEI database Dhall et al. [2012b].

are various databases, which are partially relevant Whitehill et al. [2009], Dhall et al. [2012a]. In an interesting work Whitehill et al. [2009] proposed the GENKI database. It has been collected from Google images and is an uncontrolled database containing single subject per image smiling and non-smiling. However, it does not fullfill the requirement here as the intensity level of happiness is not labelled at both image and face level. As part of this PhD project an 'in the wild' database AFEW Dhall et al. [2012a] (Chapter 3), which is a dynamic facial expressions database collected from movies is proposed. It contains both single and multiple subject videos. However, there is no intensity labels present and multiple subject video clips are few in number. Therefore, a new labelled database for learning and testing a group mood analysis system is required.

Databases such as GENKI and AFEW have been compiled using semi-automatic approaches. Whitehill et al. [2009] used Google images based on a keyword search for finding relevant images. Dhall et al. [2012a] used a recommender system based approach where a system suggested video clips to the labellers based on emotion related keywords in closed caption subtitles (Section 3.2). This makes the process of database creation and labelling easier and less time consuming. Inspired by Dhall et al. [2012a], Whitehill et al. [2009], a semi-automatic approach is followed. Web based photo sharing websites such as Flickr and Facebook contain billions of images. From a research perspective, not only are these a huge repository of images but also come with rich associated labels, which contain very useful information describing the scene in them. An 'in the wild' database from Flickr was collected. The database contains 4886 images. An automatic

Matlab based program is used to search and download images which had keywords associated with group of people and events, was developed. A total of 40 keywords were used (*eg.* 'party+people', 'group+photo', 'graduation+ceremony', 'marriage', 'bar', 'reunion', 'function', 'convocaction' etc). After downloading the images, VJ object detector trained on different data (frontal and pose models in OpenCV) was executed on the images and only the images which contained more than one subjects were kept. Images with false detections were manually removed. This labelled image collection is called HAPpy PEople Images (HAPPEI). Figure 6.4 shows a collage of images in the database.

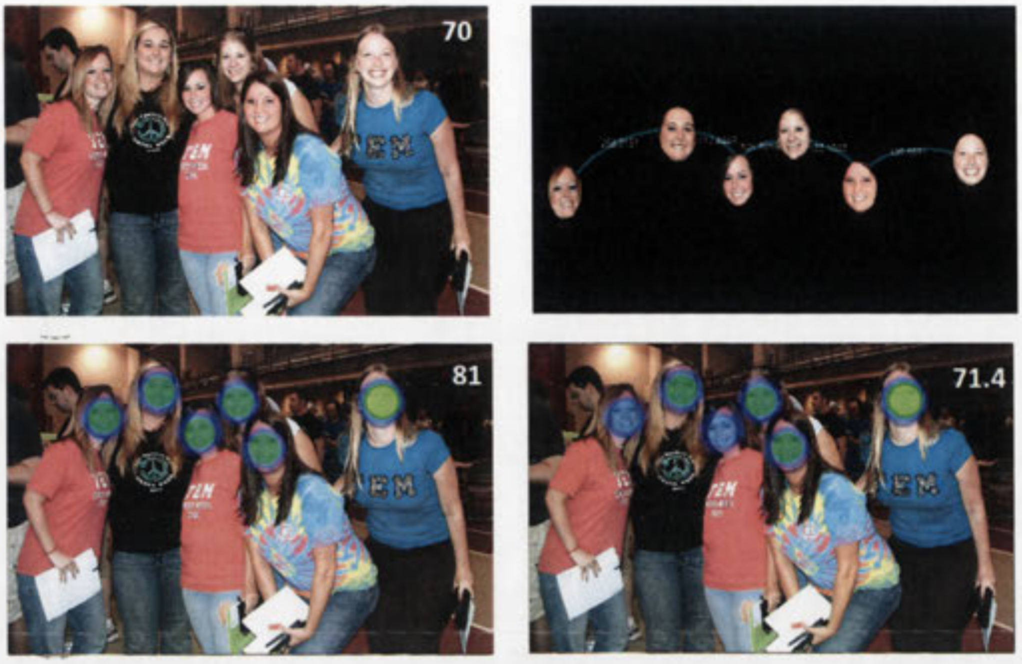
The labellers annotated all the images with the group level mood intensity ('neutral' to 'thrilled'). Moreover in 4886 images, 8500 faces were manually annotated for face level happiness intensity, occlusion intensity and pose by 4 human labelers, who annotated different images. For representing the mood, happiness intensity corresponding to six stages of happiness: *Neutral*, *Small Smile*, *Large Smile*, *Small Laugh*, *Large Laugh* and *Thrilled*. The LabelMe Russell et al. [2008] based Bonn annotation tool Korc and Schneider [2007] was used for labeling. It is interesting to note that ideally one would like to infer the mood of the group by the means of self rating along with the perception of the mood of the group. In this work, no self rating is conducted as the data is collected from the internet. In this database the labels are based on the perception of the labellers. One can see this work as a stepping stone to group mood analysis. The aim of the models (Section 6.5, 6.8 and 6.9.1) proposed in this chapter is to infer as close as possible to that of the human observers.

6.5 Group Expression Model

Given an image **I** containing a group of people $\mathcal{G}$ of size **s** and their happiness intensity level $\mathcal{I}_{\mathcal{H}}$, a simple *Group Expression Model (GEM)* can be formulated as an average of the happiness intensities of all faces in the group

$$\text{GEM} = \frac{\sum_i \mathcal{I}_{\mathcal{H}i}}{s} \quad . \tag{6.1}$$

In this simple formulation, both global information, e.g. the relative position of the people in the image, and local information, e.g. the level of occlusion of a face, are being ignored. In order to add the bottom-up and top-down components (Section 6.2.1), it is proposed here to add these social context features as weights to the process of determining the happiness intensity of a group image. The experiments (Section 6.10) on the HAPPEI confirm the positive effect due to the addition of the social feature weights to GEM. In the next section, the methods for computing the global context is discussed.



(a)

Figure 6.5: *Top Left:* Image with mood intensity score=70. *Top Right:* Min-span tree depicting connection between faces. *Bottom Left:* Happiness intensity heat map generated using the GEM model, here the mood intensity score=81. *Bottom Right:* Happiness intensity heat map with social context, the contribution of the **faces with respect to their neighbours (F2 and F4)** towards the overall intensity of the group is weighted, here the mood intensity score=71.4. It can be observed that adding the global context feature reduces the error.

6.6 Global Context

For analysing the effect of the group on a subject the following technique is applied. The tip of the nose $\mathbf{p_i}$ is considered as the position of a face f_i in the image. To map the global structure of the group, a fully connected graph $G = (V, E)$ is constructed. Here, $V_i \in \mathcal{G}$ represents a face in the group and each edge represents the link between two faces $(V_i, V_m) \in E$. The weight $w(V_i, V_m)$ is the Euclidean distance between $\mathbf{p_i}$ and $\mathbf{p_m}$. Prim's minimal spanning tree algorithm Prim [1957] is computed on $\mathcal{G}$, which provides the information about the relative position of people in the group with respect to their neighbours. In Figure 6.5, the min-span tree of the group graph is displayed.

Once, the location and minimally connected neighbours of a face are known, the relative size of a face f_i with respect to its neighbours is calculated. The size of a face is taken as the distance between the location of the eyes (intraocular distance),

$d_i = ||\mathbf{l} - \mathbf{r}||$. The relative face size θ_{i} of f_i in region r is then given by

$$\theta_{\mathrm{i}} = \frac{d_i}{\sum_i d_i/n} \tag{6.2}$$

where the term $\sum_i d_i/n$ is the mean face size in a region r around face f_i, with r containing a total of n faces including f_i. Generally speaking, the faces which have a larger size in a group photo are of the people who are standing closer to the camera. Here, it is assumed that the expression intensity of the faces closer to the camera contributes more to the overall group mood intensity as compared to the faces of people standing in the back. Eichner and Ferrari Eichner and Ferrari [2010] made a similar assumption to find if a person is standing in the foreground or at the back in a multiple people pose detection scenario.

Based on the centre locations $\mathbf{p_i}$ of all faces in a group $\mathcal{G}$, the centroid $\mathbf{c_g}$ of $\mathcal{G}$ is computed. The relative distance δ_{i} of each face f_i is described as

$$\delta_{\mathrm{i}} = ||\mathbf{p}_i - \mathbf{c}_g|| \quad . \tag{6.3}$$

δ_i is further normalised based on the mean relative distance. Faces closer to the centroid are given a higher weighting than the faces further away. Using Equations 6.2 and 6.3, a global weight is assigned to each face in the group

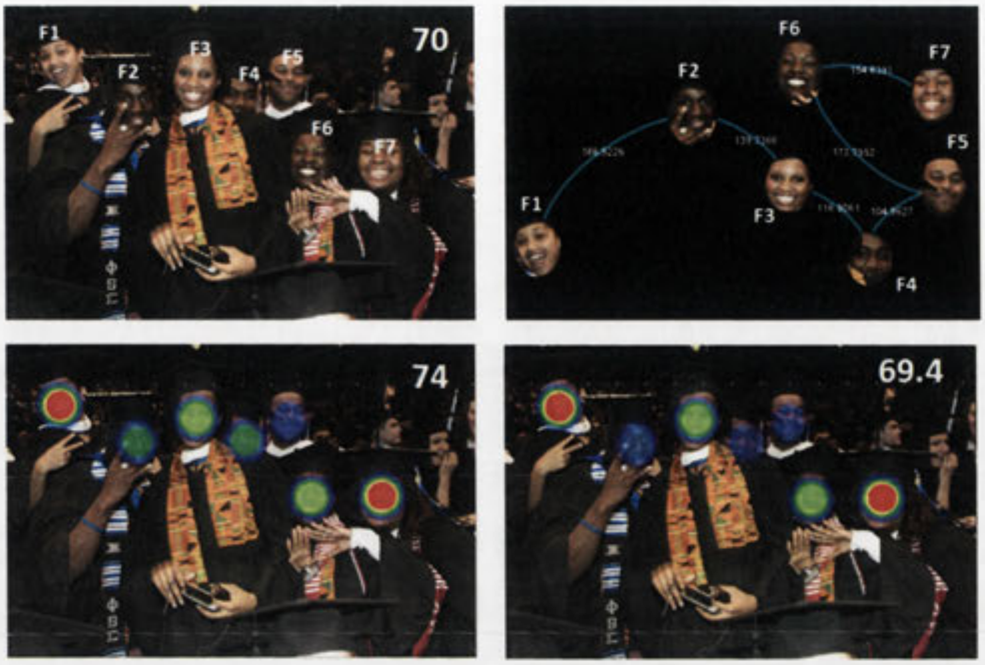
$$\psi_{\mathrm{i}} = ||1 - \alpha\delta_i|| * \frac{\theta_i}{2^{\beta-1}} \tag{6.4}$$

where α and β are the parameters, which control the effect of these weight factors on the global weight. Figure 6.6 demonstrates the effect of the global context on the overall output of GEM.

6.7 Local context

In the previous section, global context features, which compute weights on the basis of two factors: (1) where are people standing in a group and (2) how far are they away from the camera, are defined. The local context is described in terms of an individual person's level of face visibility and happiness intensity.

Occlusion Intensity Estimate: Occlusion in faces, whether self-induced (e.g. sunglasses) or due to interaction between people in groups (e.g. one person standing partially in front of another and occluding the face), is a common problem. Lind and Tang Lin and Tang [2007] introduced an automatic occlusion detection and rectification method for faces via GraphCut-based detection and confident sampling. They also proposed a face quality model based on global correlation and local patterns to derive occlusion detection and rectification.



(a)

Figure 6.6: *Top Left:* Image with happiness intensity score=70. *Top Right:* Min-span tree depicting connection between faces. *Bottom Left:* Happiness intensity heat map. *Bottom Right:* Happiness intensity heat map with social context, the contribution of the **occluded faces (F2 and F4)** towards the overall intensity of the group is penalised.

The presence of occlusion on a face reduces its visibility and, therefore, hampers the clear estimation of facial expressions. It also reduces the face's contribution to the overall expression intensity of a group portrayed in an image. Based on this local face level phenomenon, the happiness intensity level $\mathcal{I}_{\mathcal{H}}$ of a face $\mathbf{f_i}$ in a group is penalised if (at least partially) occluded. Thus, along with an automatic method for occlusion detection, an estimate of the level of occlusion is required. Unlike Lin and Tang [2007], it is proposed to learn a mapping model $\mathcal{F} : \mathbf{X} \rightarrow \mathbf{Y}$, where $\mathbf{X}$ are the descriptors calculated on the faces and $\mathbf{Y}$ is the amount of occlusion.

The mapping function $\mathcal{F}$ is learnt using the Kernel Partial Least Squares (KPLS) Rosipal [2011] regression framework. The PLS set of methods have recently become very popular in computer vision Guo and Mu [2011], Schwartz et al. [2009, 2010]. In Schwartz et al. [2010] and Schwartz et al. [2009], PLS is used for dimensionality reduction as a prior step before classification. Guo and Mu [2011] use KPLS based regression for simultaneous dimensionality reduction and age estimation and show that KPLS works well for face analysis when the feature vector is high dimension. For the occlusion intensity estimation problem, the training set $\mathbf{X}$ is a set of input samples x_i of dimension N. $\mathbf{Y}$ is the corresponding set of vectors y_i of dimension M. Then, the

PLS framework defines a decomposition of matrices $\mathbf{X}$ and $\mathbf{Y}$ as

$$\begin{aligned} \mathbf{X} &= \mathbf{TP^T} + \mathbf{E} & (6.5)\\ \mathbf{Y} &= \mathbf{UQ^T} + \mathbf{E} & (6.6) \end{aligned}$$

where $\mathbf{T}$ and $\mathbf{U}$ are the $n \times p$ score matrices of the p extracted latent projections. The $N \times p$ matrix $\mathbf{P}$ and $M \times p$ matrix $\mathbf{Q}$ denote the corresponding loading matrices and the $n \times N$ matrix $\mathbf{E}$ and $n \times M$ matrix $\mathbf{F}$ denote the residual matrices, that account for the error made by the projection. The classical NIPALS method Blalock [1975] is used to solve the optimisation criteria

$$\begin{aligned} [cov(\mathbf{t},\mathbf{u})]^2 &= [cov(\mathbf{Xw},\mathbf{Yc})]^2 \\ &= max_{|r|=|s|=1}[cov(\mathbf{Xr},\mathbf{Ys})]^2 \qquad (6.7) \end{aligned}$$

where $cov(\mathbf{t};\mathbf{u}) = \mathbf{t^T u}/n$ represents the sample covariance between the score vectors $\mathbf{t}$ and $\mathbf{u}$. The score vectors $\{\mathbf{t}_i\}_{i=1}^{p}$ are good predictors of $\mathbf{Y}$ and the inner relation between the score vectors $\mathbf{t}$ and $\mathbf{u}$ is given by $\mathbf{U} = \mathbf{TA} + \mathbf{H}$, where $\mathbf{A}$ is a $p \times p$ diagonal matrix and $\mathbf{H}$ is the residual matrix.

To perform classification, the regression matrix $\mathbf{B}$ is calculated as:

$$\mathbf{B} = \mathbf{X}^T\mathbf{U}(\mathbf{T}^T\mathbf{X}\mathbf{X}^T\mathbf{U})^{-1}\mathbf{T}^T\mathbf{Y} \quad . \qquad (6.8)$$

For a given test sample matrix X_{test}, the estimated labels matrix $\hat{Y}$ is given by:

$$\hat{\mathbf{Y}} = \mathbf{X}_{test}\mathbf{B} \qquad (6.9)$$

For a detailed description, readers may refer to Rosipal [2011]. Now, for a non-linear mapping, the kernel trick can be applied to the PLS method. $\mathbf{X}$ is substituted with $\mathbf{\Phi} = [\Phi(x_1), ..., \Phi(x_n)]^T$, which maps input data to a higher-dimensional space. The kernel matrix is then defined by the Gram matrix $\mathbf{K} = \mathbf{\Phi\Phi}^T$, in which the kernel function defines each element $\mathbf{K}_{i,j} = k(x_i, x_j)$. Therefore, Equations 6.8 and 6.9 can be rewritten as

$$\begin{aligned} \hat{\mathbf{Y}} &= \mathbf{K}_{test}\mathbf{R} & (6.10)\\ \mathbf{R} &= \mathbf{U}(\mathbf{T}^T\mathbf{K}^T\mathbf{U})^{-1}\mathbf{T}^T\mathbf{Y} & (6.11) \end{aligned}$$

where $\mathbf{K}_{test} = \mathbf{\Phi_{test}}\mathbf{\Phi}^T$ is the kernel matrix for test samples.

The input sample vector x_i is a normalised combination of Hue, Saturation and the PHOG Bosch et al. [2007] for each face. In the training set, $\mathbf{X}$ contains both occluded and non-occluded faces, $\mathbf{Y}$ contains the labels identifying the amount of occlusion (where 0 signifies no occlusion). The labels were manually created during the database creation process (Section 6.4). The output label y_i is used to compute the local weight λ_i, which will penalise $\mathcal{I}_{\mathcal{H}}$ for a face f_i in the presence of occlusion. It is defined as

$$\lambda_{\mathrm{i}} = ||1 - \gamma y_i|| \tag{6.12}$$

where γ is the parameter, which controls the effect of the local weight.

Happiness Intensity Computation: A regression based mapping function $\mathcal{F}$ is learnt using KPLS for regressing the happiness intensity of a subject's face. The input feature vector is the PHOG descriptor computed over aligned faces. As discussed earlier, the advantage of learning via KPLS is that it performs dimensionality reduction and prediction in one step. Moreover, KPLS based classification has been successfully applied to facial action units Gehrig and Ekenel [2011].

6.8 Weighted Group Expression Model

The global and local contexts defined in Eq. 6.4 and Eq. 6.12 are used to formulate the relative weight for each face f_i as

$$\pi_i = \lambda_i \psi_i \tag{6.13}$$

This relative weight is applied to the $\mathcal{I}_{\mathcal{H}}$ of each face in the group $\mathcal{G}$ and based on Eq. 6.1 and Eq. 6.13, the new weighted GEM is defined as

$$\mathrm{GEM}_w = \frac{\sum_i \mathcal{I}_{\mathcal{H}i} \pi_i}{s} \tag{6.14}$$

This formulation takes into consideration the structure of the group and the local context of the faces in it. The contribution of each face f_i's $\mathcal{I}_{\mathcal{H}i}$ towards the overall perception of the mood of the group is weighted relatively, here $\mathcal{I}_{\mathcal{H}i}$ is the happiness intensity of f_i.

6.9 Social Context as Attributes

The social features described above can also be viewed as manually defined attributes. From the survey (Section 6.3), it is evident that along with the social context features there are many other properties such as age, attractiveness, gender etc., which affect the perception of the mood of a group. The assumptions in GEM_w do not hold true for some scenarios, for example: when a baby is in the lap of the mother. In this case the system will give higher weight to mother and lower to the baby assuming that the mother is in the front and the baby is in the background. In order to implicitly add the effect of other attributes, a feature augmentation approach is presented next.

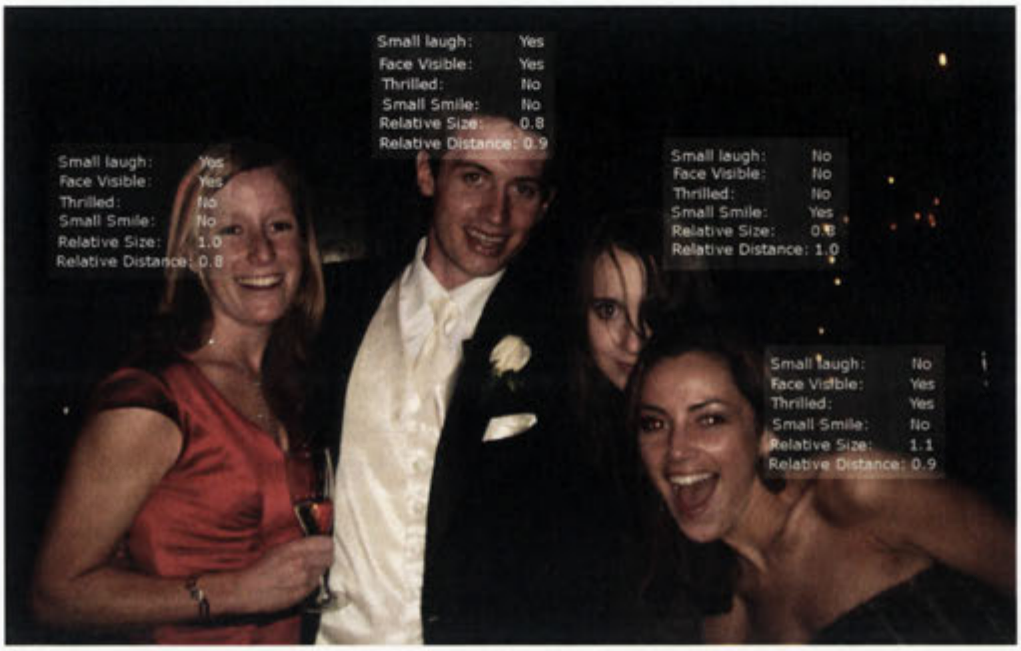


Figure 6.7: The picture shows some manually defined attributes for subjects in a group.

Lately, attributes have been very popular in the computer vision community (e.g. Parikh and Grauman [2011]). Attributes are defined as high-level semantically meaningful representations. They have been, for example, successfully applied to object recognition Parikh and Grauman [2011], scene analysis Li and Perona [2005] and face analysis Jain et al. [2007].

Based on the regressed happiness intensities, the attributes are defined as *Neutral*, *Small Smile*, *Large Smile*, *Small Laugh*, *Large Laugh*, *Thrilled*, and for occlusion as *Face Visible*, *Partial Occlusion* and *High Occlusion*. These attributes are computed for each face in the group. Attributes based on global context are *Relative Distance* and *Relative Size*. Figure 6.7 describes the manual attributes for faces in a group.

Defining attributes manually is a subjective task, which can result in many important discriminative attributes being ignored. Inspired by Tsai et al. [2011], low-level feature based attributes are computed. They propose the use of manually defined attributes along with data-driven attributes. Their experiments show a leap in performance for human action recognition based on combination of manual and data-driven attributes. Weighted bag of visual words based on extracting low-level features are computed.

Furthermore, a topic models is learnt using Latent Dirichlet allocation (LDA) Blei et al. [2001]. The manually defined and weighted data-driven attributes are combined to form a single feature.

6.9.1 Augmented Group Expression Model

Topic models, though originally developed for document analysis domain, have found a lot of attention in computer vision problems. One very popular topic modelling technique is Latent Dirichlet Allocation (LDA) Blei et al. [2001], a hierarchical Bayesian model, where topic proportions for a document are drawn from a Dirichlet distribution and words in the document are repeatedly sampled from a topic, which itself is drawn from those topic proportions.

Jain et al. [2007] introduced people-LDA, where topics were modelled around faces in images along with titles from news. Jain et al. [2007]'s work is considered different from the method proposed in this section. The single biggest difference is that Jain et al. [2007] propose to create topics around single people rather around group of people. The proposed group work creates topics around intensities of happiness for a group of people. For learning the topic model, a dictionary is learnt first.

Weighted Soft Assignment: K-means is applied to the image features for defining the visual words. For creating a histogram, each word of a document is assigned to one or more visual words. If the assignment is limited to one word, it is called hard assignment and if multiple words are considered, it is called soft assignment. The cons of hard assignment are that if a patch (face in a group $\mathcal{G}$) in an image is similar to more than one visual word, the multiple association information is lost. Therefore, Jiang et al. [2007] defined a soft-weighting assignment to weight the significance of each visual word towards a patch. For a visual vocabulary of K visual words, a K-dimensional vector $T = [t_1...t_K]$ with each component t_k representing the weight of a visual word k in an group $\mathcal{G}$ is defined as

$$\mathbf{t_k} = \sum_{i}^{N} \sum_{j}^{M} \frac{1}{2^{i-1}} sim(j,k), \tag{6.15}$$

where M_i represents the number of face f_j whose i^{th} nearest neighbour is the visual word k. The measure $sim(j,k)$ represents the similarity between face f_j and the visual word k. It is worth noting that the contribution of each word is dependent to its similarity to a visual word weighted by the factor $\frac{1}{2^{i-1}}$.

Relatively weighted soft-assignment: Along with the contribution of each visual word to a group G, it is interesting to add the global attributes as weights here. It is intuitive to note that the weights effect the frequency component of words, (words here represent faces in a group). Therefore, it is similar to applying weights in GEM_w and looking at the contribution based on neighbourhood analysis of a particular subject under consideration. As the final goal is to understand the contribution of each face f_i towards the happiness intensity of its group $\mathcal{G}$, the relative weight formulated in Eq. 6.13 is used to define a 'relatively weighted' soft-assignment. Eq. 6.15 can then be modified as

$$\mathbf{t_k} = \sum_{i}^{N} \sum_{j}^{M} \frac{\psi_j}{2^{i-1}} sim(j,k) \quad . \tag{6.16}$$

Now, along with weights for each nearest visual word for a patch, another weight term is being induced, which represents the contribution of the patch to the group. These data-driven visual words are appended with the manual attributes. Note that the histogram computed here is influenced by the global attributes of the faces in the group.

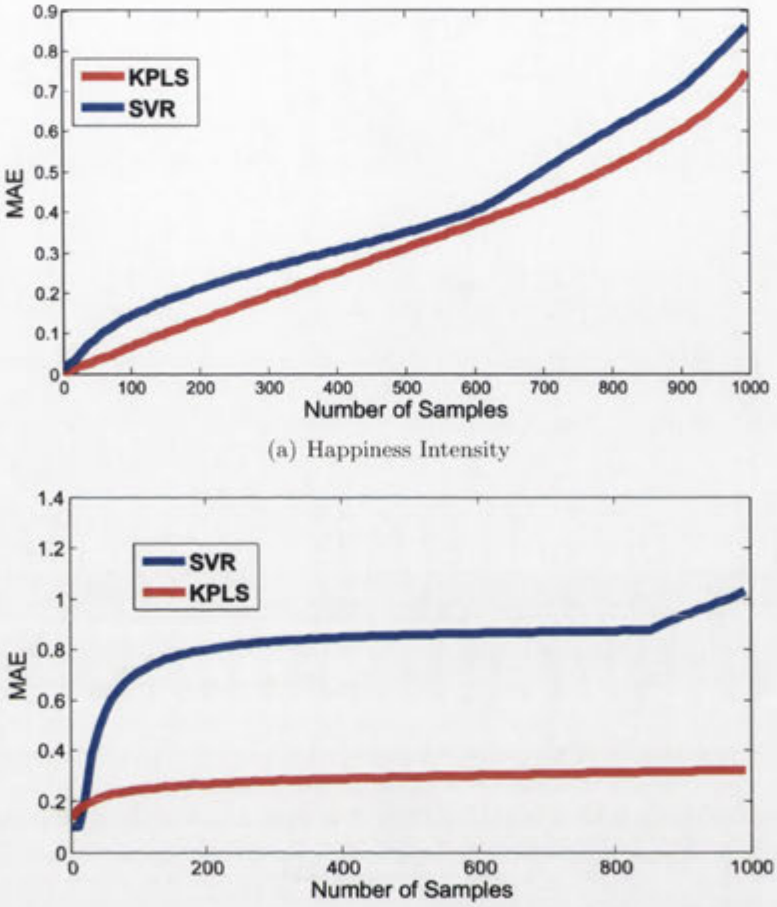
The default LDA formulation is an unsupervised Bayesian method. In their recent work, Blei and McAuliffe [2007] proposed the Supervised LDA (SLDA) by adding a response variable for each document. It is shown to perform better for regression and classification task. Since, in HAPPEI the human annotated labels for the happiness intensities at the image level are present, using a supervised topic model is a natural choice. The document corpus is the set of groups $\mathcal{G}$. The word here represents each face in $\mathcal{G}$. Max Entropy Discriminant LDA (MedLDA) Zhu et al. [2009] is computed for topic model creation and test label inference. The LDA formulation for groups is called GEM_{LDA}. In the results, section the average model GEM is compared with the weighted average model GEM_w and the feature augmented topic model GEM_{LDA}.

6.10 Experiments

6.10.1 Face Processing Pipeline

Given an image, the VJ object detector Viola and Jones [2001] models trained on frontal and profile faces are applied on the images. For extracting the fiducial points, Everingham et al. [2006]'s part based point detector is applied. This gives nine points, which describe the location of the left and the right corners of both eyes, the centre point of the nose, left and right corners of the nostrils, and the left and right corners of the mouth. For aligning the faces, an affine transform is applied.

As the images have been collected from Flickr and contain different scenarios and complex backgrounds, classic face detectors, such as the VJ object detector, give a fairly high false positive rate (13.6%). To minimise this error, a non-linear binary SVM Chang and Lin [2001] is trained. The training set contains samples containing faces and non-faces. For face samples, all the true positives from the output of VJ detector executed on 1300 images from the HAPPEI database are selected. For non-faces, the samples are manually selected from the same VJ output on the HAPPEI database. To create a large number of false positives from real world data, an image set containing monuments, mountains and water scenes (but no persons facing the camera) is constructed. To learn the parameters for SVM, five-fold cross validation is



(a) Happiness Intensity

(b) Occlusion Intensity

Figure 6.8: The two graphs describe the MAE. a) Happiness intensity comparison. b) Occlusion intensity comparison.

performed.

6.10.2 Implementation Details

Given a test image I containing group $\mathcal{G}$, the faces in the group are detected and aligned. The faces are cropped to 70×70 pixel size. For happiness intensity detection, PHOG features are extracted from the face. Here, pyramid level $L = 3$, angle range $= [0 - 360]$ and bin count $= 16$. The number of latent variables are chosen as 18 after empirical validation. PHOG is scale invariant. The choice of using PHOG is motivated by Dhall et al. [2011b], where PHOG performed well for facial expression analysis.

The parameters for MedLDA are $\alpha = 0.1$, $k = 25$, for SVM $fold = 5$. 1500 documents are used for training and 500 for testing. The range of label is the group mood intensity range [0-100] with a step size of 10. For learning the dictionary, k the

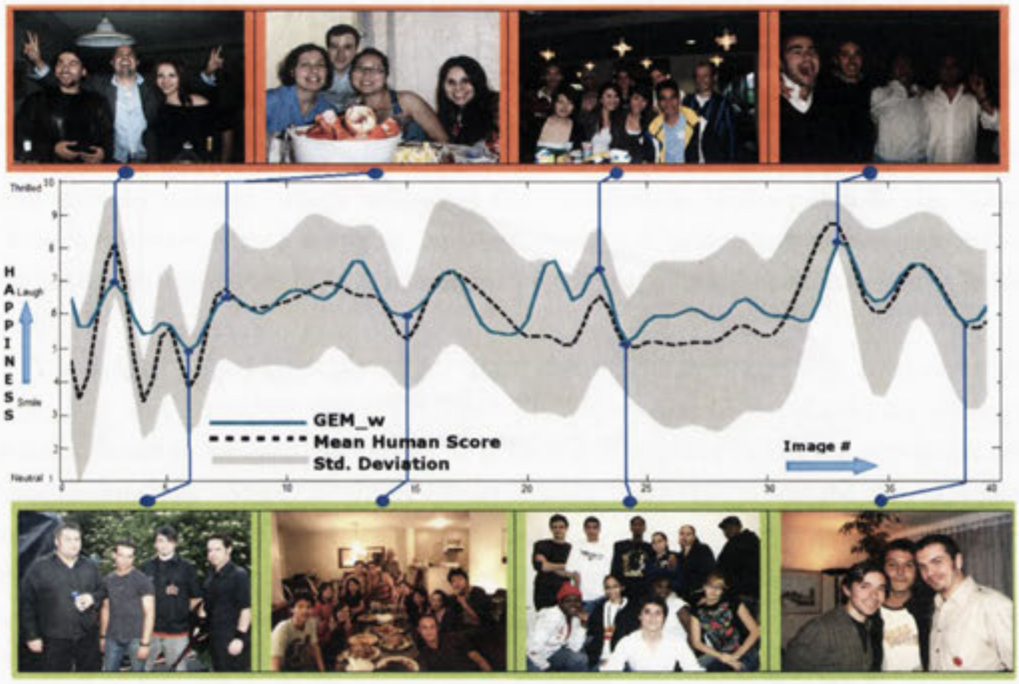


Figure 6.9: The graph describes the comparison of the group mood intensity as calculated by the proposed methods with the results from the user study. The top row shows images with high intensity score and the bottom row shows images which are close to neutral. Please note that the images are from different events.

number of words is empirically set to 60. In Eq. 6.4 and Eq. 6.12, the parameters are set as follows: $\alpha = 0.3$, $\beta = 1.1$ and $\gamma = 0.1$. For a fair comparison between the three models (GEM, GEM_w and GEM_{LDA}) both quantitative and qualitative experiments are performed. 2000 faces are used for training and 1000 for testing of the happiness and occlusion intensity regression models.

6.10.3 Human Label Comparison

Mean Average Error (MAE) is used for comparing GEM, GEM_w and GEM_{LDA}. The performance of the occlusion intensity and happiness intensity estimators, which are based on KPLS are compared with Support Vector Regression (SVR) Chang and Lin [2001] based occlusion intensity and happiness intensity estimators. Figure 6.8 displays

Method	GEM	GEM_w	GEM_{LDA}
MAE	0.455	0.434	0.379

Table 6.1: The table compares the Mean Average Error for the three group expression models: GEM, GEM_w and GEM_{LDA}.

the comparison based on the MAE scores. The MAE for occlusion intensity are 0.79 for KPLS and 1.03 for SVR. The MAE for happiness intensity estimation for KPLS is 0.798 and for SVR 0.965. Table 6.1 shows the MAE comparison of GEM, GEM_w and GEM_{LDA}. As hypothesised, the effect of adding social features is evident in the lower MAE in GEM_w and GEM_{LDA}.

6.10.4 User Study

A two-part user survey is performed. A total of 15 subjects were asked to a) give happiness intensities to 40 images and b) rate the output of the three methods for their output of the top-5 happiest images from an event. Here, the users were asked to rate a score from the range 0 (not good at all) to 5 (good) for the three methods for three social events each. They did not know, which output belonged to which method. For part a), Figure 6.9 shows the output. Note that the happiness scores computed by the GEM_w are close to the mean human score and are well within the range of the standard deviation of the human labellers' scores. For part b), ANOVA tests were performed with the hypothesis that adding social context to group mood analysis leads to an estimate closer to human perception. For GEM and GEM_w, $p < 0.0006$, which is statistically significant in the one-way ANOVA. For GEM and GEM_{LDA}, $p < 0.0002$, which is also statistically significant.

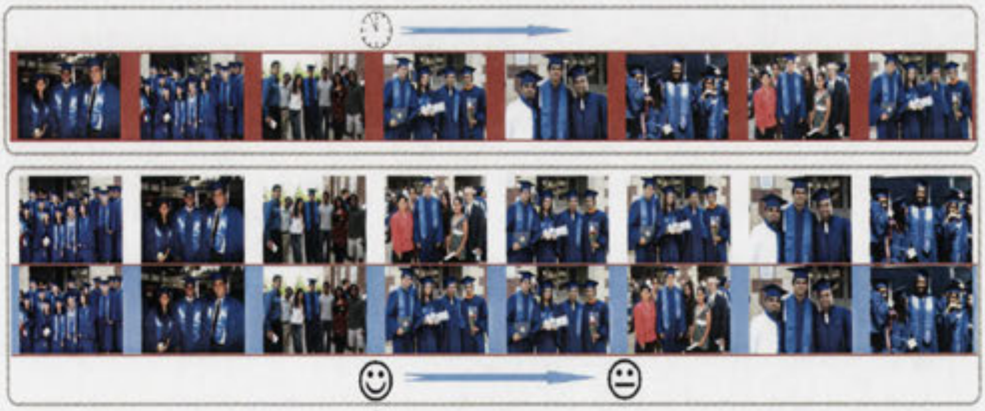


Figure 6.10: This figure shows a graduation ceremony. The top row (red background) are the images from a graduation ceremony organised by timestamps. The second row (white background) are the images ranked by their decrease in intensity of happiness (from left to right) by human annotators. The third row (blue) are the images ranked by their decrease in intensity of happiness (from left to right) by $\mathbf{GEM}_w$. A higher resolution figure can be found in the Appendix C.

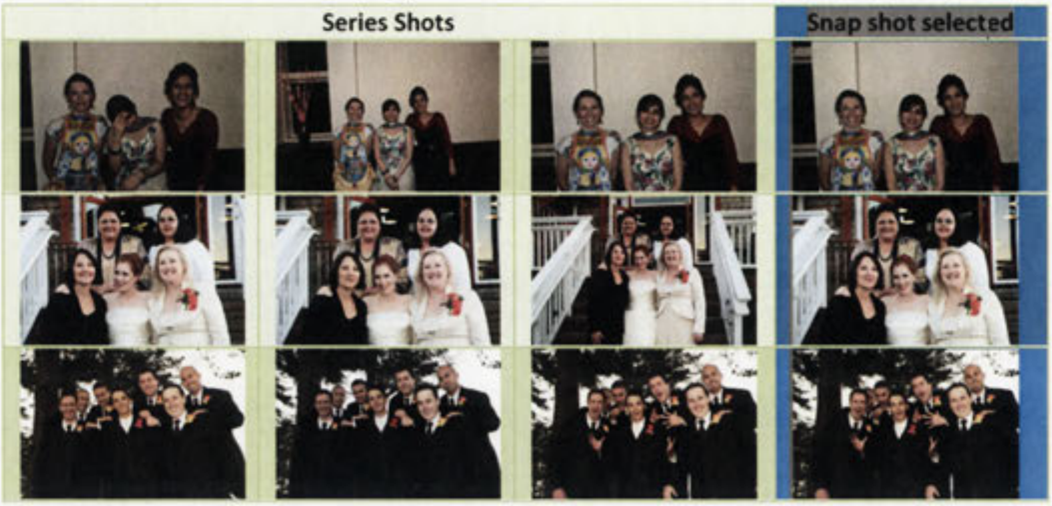


Figure 6.11: Candid Group Shot Selection: Here, each row represents a series of photographs of the same people. The fourth column is the selected shot based on the highest score from $\mathbf{GEM}_w$ (Eq. 6.14). Please refer to the Appendix C for more experiments.

6.10.5 Image Ranking from an Event

For comparison of the proposed framework, volunteers were asked to rank a set of images containing a group of people from an event. Now the task is as follows: Given a social event with different or the same people present in the same or different photographs, the happiest moment of the event is to be found. Therefore, all the images are ranked on the basis of their decreasing amount of happiness intensity. Figure 6.10 is a snapshot for an event ranking experiment. In the first row, the images are arranged based on their timestamp, i.e. when they were shot. The second row shows the ranking by human labellers. The highest happiness intensity image is on the left and decreases from left to right. Now, the output of the $\mathbf{GEM}_w$ is in row 3, where the proposed method ranked the images in order of their decreasing happiness intensity.

6.10.6 Candid Group Shot Selection

There are situations in social gatherings when multiple photographs are taken for the same subjects in a similar scene within a short span of time. Due to the dynamic nature of groups of people, it is a challenging task to get the most favourable expression together in a group of people. Here, the group mood analysis method is applied to shot selection after a number of pictures have been taken. In Figure 6.11, the rows are the shots taken at short intervals. The $\mathbf{GEM}_w$ ranks the images containing the same subjects and the best image (highest happiness quotient) is displayed in the fourth column. More experiments and visual outputs can be found in the Appendix C.

6.11 Conclusion

Social gathering events generate many group shots. In this chapter a framework for estimating the mood of a group of an image, focussing on positive mood, is proposed. To the best of my knowledge, this is the first work for analysing mood of a group based on the structure of a group and local attributes such as occlusion. An 'in the wild' database called HAPpy PEople Images (HAPPEI) is collected from Flickr based on keyword search. It is labelled at both image and face level. From the perspective of social context, the global structure of the group is explored. Relative weights are assigned to the happiness intensities of individual faces in a group, so as to estimate their contribution to the group mood. The experiments show that assigning relative weights to intensities helps in better predication of the group mood. Further, feature augmented topic model based group mood analysis model performs better than the average and weighted group expressions model.

In the future, social context factors such as age, gender *etc.* will be explored and their effect on mood analysis of a group will be evaluated. As this work focuses on defining the framework and social constraint, standard detector algorithms and descriptors have been used. Further extensions of the work can benefit immensely from robust face detection and alignment and the use of new, faster and more discriminative descriptors. In its current form, the proposed framework has a limitation w.r.t. extreme poses. In future, pose handling will be added and a system will be developed, which selects group photographs that can benefit immensely from pose information. Further, computer vision problems such as early event detection and abnormal event detection for multiple subjects can be dealt with this framework.

To summarise, in Chapter 3, the focus is on how to construct facial expressions database representing different challenging real-world conditions. From the learning in Chapter 3 that MoPS framework is the current state-of-the-art face and fiducial points detector on AFEW, a HPN method based on MoPS framework is proposed in Chapter 4. In Chapter 5, the problem of structural similarity for finding similar facial expressions is evaluated. The problem solution is based on fiducial points. As discussed in the Section 1.1, along with lack of labelled data, presence of multiple subjects in a sample poses a significant challenge to FER in real-world conditions. In the current Chapter 6, a framework is proposed for handling the expression of multiple subjects in an image based on social contextual features. The proposed group models are evaluated on the application of 'candid group shot selection' and 'event summarisation'.

Chapter 7

Conclusion and Future Work

This thesis makes contributions towards the understanding of automatic facial expression analysis in real-world conditions and its applications to various problems.

7.1 Summary of Contributions

A summary of contributions of the work presented in this thesis is as follows -

1. Collecting an 'in the wild' facial expression database

Standard facial expression databases such as the CK+ Lucey et al. [2010], Multi-PIE Gross et al. [2008a], MMI Valstar et al. [2005] etc. have been recorded in lab-controlled settings where subjects were asked to pose for specific facial expression. For moving facial expression analysis methods from labs to the real-world, a database representing real-world scenarios (AFEW Dhall et al. [2012a]) is presented. Capturing databases is a time consuming process and widely used databases such as the CK+ Lucey et al. [2010], Multi-PIE Gross et al. [2008a], MMI Valstar et al. [2005] etc. have been manually constructed. To overcome the time consuming data collection process and create a non-posed database representing real-world conditions, movie data is used.

A recommender system based on closed caption subtitles parsing is proposed. The system scans subtitles for presence of emotion related keywords. The labeller is presented with a short video clip and a potential expression label. This helps the labeller in easily selecting or discarding data which is relevant to the proposed database AFEW. The labeller can further change the proposed label (if incorrect) and use facial expression, audio, body expression, scene context and subtitle for proposing a new label. This way context is introduced in the labelling process. AFEW forms the platform for the EmotiW challenge Dhall et al. [2013]. Recent works (Ebrahimi et al. [2013], Sikka et al. [2013b] etc.) proposed by participants in the EmotiW challenge, show that for 'in the wild' emotion recognition, scene information, body expressions and audio play vital role.

The labeller also records information about the subject(s) in the video clips using knowledge from the internet movie database (http://imdb.com). A XML based schema has been carefully designed to store this meta-data.

Further, a novel static database (SFEW Dhall et al. [2011c]) is proposed by manually selecting frames from the AFEW database. Strict experimental protocol is defined for both AFEW and SFEW. The performance of state-of-the-art FER methods is compared for AFEW with CK+ and for SFEW with JAFFE and Multi-PIE databases. The comparison clearly shows the current status of the FER methods and their limitations when dealing with real-world data.

Both AFEW and SFEW have been downloaded more than 120 times since their release and have defined a platform for the new research problem of facial expression analysis 'in the wild'.

2. Head Pose Normalisation

In the EmotiW challenge (Section 3.8), the MoPS framework was used for face and fiducial points detection. MoPS performs better in face detection specially for non-frontal poses as compared to the standard VJ face detector. However, even with MoPS fiducial points detection is not fully accurate on AFEW, although it was found to be better than state-of-the-art AAM and CLM fitting algorithms. Based on this motivation a HPN method is proposed based on the MoPS framework.

Traditional HPN methods Asthana et al. [2009a, 2011], Rudovic and Pantic [2011], Rudovic et al. [2010a,b] are based on fiducial points as input. However, given that fiducial points detection is an open problem for real-world images, traditional HPN methods do not perform well for images containing 'in the wild' scenario. A parts-detector based approach is proposed (Section 4.5). The confidence score of part-detectors is normalised using Twin-GPR algorithm. This is integrated within the MoPS framework. Post confidence maps normalisation, frontal shape model is applied. This applies a strict facial shape constraint on the virtual confidence map. Therefore, the proposed HPN method is shape constrained and has no explicit dependency on fiducial points.

Moreover, by following the mixture concept for faces of MoPS, HPN is performed for all non-frontal poses on an image and the highest scoring normalisation is chosen as the final result. This removes the dependency on head pose information. Similar to fiducial points detection, head pose inference is an open problem for real-world images. Therefore, the proposed HPN methods (Section 4.5) are shape-constrained, fiducial points free and head pose invariant.

The experiments are conducted on the Multi-PIE and the SFEW databases. The results show that the proposed HPN methods perform better than the traditional HPN methods Asthana et al. [2009a, 2011], Rudovic and Pantic [2011], Rudovic et al.

[2010a,b].

3. Similar facial expression analysis

A fully automatic method for finding similar facial expressions is proposed. A novel geometric descriptor: Expression Image (EI) is proposed. EI is computed using normalised fiducial points. EI of two faces are compared using Structural Similarity Index Metric (SSIM) Wang et al. [2004b]. Experiments are performed on the Multi-PIE and AVOZES databases. Results show that SSIM performs better than L2 distance metric.

Further, the proposed method is applied to the problem of 'facial expression based album creation' and 'album by similar expression'. Results are computed on the FEED-TUM Wallhoff [2006] and the LFW Huang et al. [2007] databases. The proposed method is applied for finding candidate sets for learning facial performance transfer method.

4. Group Mood Analysis

Images collected in social events, generally contain multiple subjects. Current FER methods have been trained and tested on samples containing a single subject. Therefore, an image based database containing multiple subjects per image is collected using Flickr search. The proposed database is called HAPPEI Dhall et al. [2012b]. A group expression model is proposed based on an averaging model. The mood intensity is inferred for all the subjects of an image and their contribution is evaluated. The contribution of each subject is weighted by its social attributes. The social attributes comprise of global and local features.

A basic averaging model is first proposed. Further, a weighted group model is proposed which overcomes the limitations of the averaging based model. Bag-of-words based feature augmentation is performed on the social attributes and a topic model is learnt. This topic model implicitly learns the relationship between words for them falling under a topic. The three proposed models are compared on the HAPPEI database. The group models are applied to the problem of 'candid photo shot selection' and 'group event synopsis'. To the best of our knowledge this is the first work where facial expression of a group of people in an image have been analysed for inferring the mood at group level.

7.2 Future Work

Based on the contributions of this thesis, various directions can be taken for the future work.

1. Semi-automatic database collection and modelling

In Chapter 3, a semi-automatic method based on closed caption parsing is proposed for collecting temporal facial expression analysis database representing real-world conditions. The labeller is recommended a possible label based on the keyword search. The labeller can select or ignore this label. Now the ground truth created with the temporal database can be used as coarse labels for creating an image based database in a semi-automatic method. Currently, the SFEW database is created by manually selecting the frames from the video. Two possible extensions for data collection are:

Based on fiducial points location, a relative ranking can be performed, by computing distances (as in Figure 5.8) within the frames of a video clip in AFEW. The frames with the largest difference from a neutral can be proposed as frames which show the expression label same as that of the video clip.

Recently, Multiple Instance Learning (MIL) has been applied to the problem of pain classification Sikka et al. [2013a]. The pain database video clips do not have the information of when the pain is displayed by the subjects during the inference. Correlating the problem of semi-automatic image based database creation from labelled video clip with MIL based pain classification, it is intuitive that pain localisation is similar to localising frames in a video clip which have the same label as the video clip. Therefore, MIL based classification can be performed on AFEW video clips and instances with high probability (frames) can be assigned the ground truth (facial expression) label as that of the video clip.

2. Context based facial expression recognition

The label of a video in AFEW is influenced not only by the facial expression but by the audio, body expression, subtitle and scene information. Therefore, inference model should take into consideration the body expression. Furthermore, given that fiducial points detection is an open problem, fusing it with body parts detection can give estimate on the rough location of the face. Cues can be take from the work of Kleinsmith and Bianchi-Berthouze [2013] for integrating body expressions at either feature or classifier end. Scene analysis is a hot topic in computer vision. A scene's mood in terms of positiveness or negativeness can be inferred and this can be used as a feature.

3. Extending spatio-temporal descriptors

Spatio-temporal descriptors such as LBP-TOP are computed on three orthogonal planes. Generally, it is assumed that the XY frame will be close to the apex of an expression. This assumption is OK for databases such as CK+, MMI *etc.*. However, for database such as AFEW, it is non-trivial to localise the apex of an expression. In order to incorporate this information in descriptors like LBP-TOP, key-frame selection can be performed. In Dhall et al. [2011c], key-frames were extracted based on clustering of normalised fiducial points. The cluster centres were chosen as the representative frames. LBP-TOP can be extended by adding one or multiple XY frames which are cluster centers. This suggest that the XY frame(s) will be representing various stages of an expression.

Further, in Dhall et al. [2011c], it is shown that PHOG Bosch et al. [2007] performs better than LBP Ojala et al. [2002], for subject independent FER. Therefore, similar to LBP-TOP, HOG-TOP and PHOG-TOP can be computed with the key-frame extension discussed above.

4. Pose normalised texture feature

In Chapter 4, the part-detectors are applied for generating confidence maps in the MoPS framework. The confidence maps are normalised by regressing from non-frontal to frontal head pose. The normalised confidence maps can be treated as texture. The fiducial parts localisation can be used to crop the region of interest in the normalised confidence maps. The normalised confidence are frontal and hence the texture is pose normalised. A BOW based framework can be used to learn a dictionary from various normalised confidence maps. This can be used as a powerful texture descriptor for FER.

5. Body expression and scene analysis for group mood inference

In Chapter 6, for inferring the mood of a group the global social features are based on the relative location of a person. The aim is to find salient or important faces which can be the leader in the group. An interesting direction is to compute image saliency and weight the confidence of subjects who fall in the highly salient area. The hypothesis here is that, the human visual system fixates on salient objects and intuitively, one does not need to look at all the faces to infer the mood of a group.

Further, human body pose can be merged with the face analysis of a group of people. Based on the latest work by Kleinsmith and Bianchi-Berthouze [2013], body pose can convey affect information. In the images downloaded from the internet, there can be challenges such as face blur and occlusion due to neighbours in a group. This can make the inference of the mood of a person non-trivial. Body pose information can be fused with face information for robust inference. Attributes such as clothe colors and background scene details can also give important information about the social event and hence, aid in inferring the mood of a group. The mood value of the group can be fused with other attributes such as the one mentioned in the Kansei image retrieval systems Berthouze and Berthouze [2001].

6. Continuous facial expression analysis in the wild

In this thesis, the categorical approach to facial expression analysis is proposed. The main motive is to simplify labelling and collect data representing real-world conditions. The next natural step is to first do continuous labelling of AFEW data. Semi-automatic approach based on scanning window and MIL can be used for labelling.

Appendix A

Facial Expressions In The Wild Database

The seventy-five movies used in the AFEW 3.0 database are:

1. 21
2. About a boy
3. American History X
4. And Soon Came The Darkness
5. Black Swan
6. Bridesmaids
7. Change Up
8. Chernobyl Diaries
9. Crying Game
10. Curious Case Of Benjamin Button
11. December Boys
12. Deep Blue Sea
13. Descendants
14. Did You Hear About the Morgans?
15. Dumb and Dumberer: When Harry Met Lloyd
16. Four Weddings and a Funeral

17. Friends with Benefits
18. Frost/Nixon
19. Ghoshtship
20. Girl With A Pearl Earring
21. Hall Pass
22. Halloween
23. Halloween Resurrection
24. Harry Potter and the Philosopher's Stone
25. Harry Potter and the Chamber of Secrets
26. Harry Potter and the Deathly Hallows Part 1
27. Harry Potter and the Deathly Hallows Part 2
28. Harry Potter and the Goblet of Fire
29. Harry Potter and the Half Blood Prince
30. Harry Potter and the Order Of Phoenix
31. Harry Potter and the Prisoners Of Azkaban
32. I Am Sam
33. It's Complicated
34. I Think I Love My Wife
35. Jennifer's Body
36. Juno
37. Little Manhattan
38. Margot At The Wedding
39. Messengers
40. Miss March
41. Nany Diaries
42. Notting Hill

43. Oceans Eleven
44. Oceans Twelve
45. Oceans Thirteen
46. One Flew Over the Cuckoo's Nest
47. Orange and Sunshine
48. Pretty in Pink
49. Pretty Woman
50. Pursuit of Happiness
51. Remember Me
52. Revolutionary Road
53. Runaway Bride
54. Saw 3D
55. Serendipity
56. Solitary Man
57. Something Borrowed
58. Terms of Endearment
59. There Is Something About Mary
60. The American
61. The Aviator
62. The Devil Wears Prada
63. The Hangover
64. The Haunting of Molly Hartley
65. The Informant!
66. The King's Speech
67. The Pink Panther 2
68. The Social Network

69. The Terminal
70. The Town
71. Valentine Day
72. Unstoppable
73. Wrong Turn 3
74. You've Got Mail

(a) Val Video

	An	**Di**	**Fe**	**Ha**	**Ne**	**Sa**	**Su**
An	26	0	2	6	8	11	6
Di	15	10	4	6	7	7	1
Fe	18	3	8	5	6	5	9
Ha	20	1	5	27	3	5	1
Ne	8	5	7	7	19	2	7
Sa	15	3	4	6	13	13	10
Su	11	5	4	8	11	8	5

(b) Val Audio

	An	**Di**	**Fe**	**Ha**	**Ne**	**Sa**	**Su**
An	25	10	7	6	1	4	6
Di	13	6	4	9	7	5	6
Fe	12	8	14	6	4	8	2
Ha	20	3	8	13	8	4	6
Ne	8	10	5	16	7	6	3
Sa	12	15	12	6	2	9	8
Su	14	7	7	7	8	4	5

(c) Val Audio-Video

	An	**Di**	**Fe**	**Ha**	**Ne**	**Sa**	**Su**
An	26	1	2	7	17	3	3
Di	4	0	0	14	30	1	1
Fe	11	2	3	14	17	4	3
Ha	11	0	2	16	30	2	1
Ne	7	1	0	12	35	0	0
Sa	7	0	2	17	28	5	5
Su	2	0	3	7	33	4	3

Table A.1: Confusion matrix for $\mathbf{Val}_{\textbf{audio}}$, $\mathbf{Val}_{\textbf{video}}$, $\mathbf{Val}_{\textbf{audio-video}}$. For details see Section 3.8.2.

(a) Test Video

	An	**Di**	**Fe**	**Ha**	**Ne**	**Sa**	**Su**
An	27	3	3	4	6	4	7
Di	14	6	4	7	6	4	8
Fe	9	4	0	4	9	2	5
Ha	9	5	1	24	1	4	6
Ne	11	13	1	5	9	6	3
Sa	8	3	3	11	10	3	5
Su	7	5	6	5	7	3	2

(b) Test Audio

	An	**Di**	**Fe**	**Ha**	**Ne**	**Sa**	**Su**
An	24	4	6	9	2	3	6
Di	14	10	2	9	7	4	3
Fe	8	4	9	2	4	2	4
Ha	17	4	4	8	5	7	5
Ne	6	8	6	7	13	6	2
Sa	12	6	6	7	3	4	5
Su	6	5	6	9	2	5	2

(c) Test Audio-Video

	An	**Di**	**Fe**	**Ha**	**Ne**	**Sa**	**Su**
An	36	0	1	2	14	0	1
Di	13	0	1	15	18	1	1
Fe	8	1	2	4	16	0	2
Ha	12	1	2	8	22	1	4
Ne	5	0	0	3	39	1	0
Sa	16	1	1	8	13	0	4
Su	10	1	2	10	9	2	1

Table A.2: Confusion matrix for $\textbf{Test}_{\textbf{audio}}$, $\textbf{Test}_{\textbf{video}}$, $\textbf{Test}_{\textbf{audio-video}}$. For details see Section 3.8.2.

Appendix B

Continuous Head Pose Normalisation

While continuous head pose normalisation is not the goal of Chapter 4, a proof of concept that it is possible to extend the current HPN method (Section 4.5) for continuous head pose normalisation is demonstrated here. For dealing with faces in videos, continuous HPN is required. Zhu and Ramanan [2012] argue that the appearance of a part does not changes with a subtle pose change, therefore a detector for part i in pose angle p can be shared for the same part i for a pose angle $p + \delta$. Further experiments in Zhu and Ramanan [2012] showed that sharing based models and independent model have comparable performance. However, sharing based models are faster upto ten times as compared to the independent models Zhu and Ramanan [2012]. The confidence maps based methods (CM-HPN$_{PS}$ and CM-HPN$_{PI}$) can be extended from discrete to continuous by sharing part-specific regression models $\mathcal{R}$, which are shared among neighboring pose angles.

Algorithm 2: Continuous head pose normalisation

```
Require: Video $V$.
    $p, L^*_f = Algorithm3(V_1, p)$ {For the first frame}
    for $i \leftarrow 2$ to $lenght(V)$ do
        $Score_p, L^*_f = Algorithm1(V_i, p)$
        if $L^*_f == null$ then
            $p, L^*_f = Algorithm3(V_i, p)$ {Update head pose $p$}
        end if
    end for
```

For a video V with n frames, the first frame's head pose p and frontal reconstructed points are computed using Algorithm 3 (Same Algorithm 2 in main paper). For the consecutive frames, the part-wise regression models specific to the pose computed for first frame are used for normalisation (Algorithm 1). If the current frame's head pose

$p + \delta$ has a large difference from the current p, it may lead to a poor facial landmark reconstruction (based on low $Score^*_f$, which if less than a threshold will discard the reconstruction). Based on this the current frame's head pose p and normalised frontal points are re-computed using 3 and the new value of head pose p is used for upcoming frames. Algorithm 2 defines the process.

```
Algorithm 3: Pose Invariant Confidence Map Regression
  Input: Frame $I$
  Output: Pose $p$ and frontal parts configurations $L^*_f$
for pose $p \in P$ do
    $Score_p, L^p_f = Algorithm1(I, p)$
end
$L^*_f = L^p_f$ for the largest $Score_p$ return $p$ with the largest $Score_p$
```

Appendix C

Analysis of the Mood of a Group of People

C.1 Candid Group Shot Selection

As discussed in Section 5 of the paper, we conducted an experiment on images from our HAPPIE database (Section 4) for shot selection, where the images of a group of people taken at short time intervals were rated based on $\mathbf{GEM}_w$ for selecting the best shot image. Figure C.1 describes some experiment images for shot selection. Each row is a set and column three is the selected shot.

C.2 Image Ranking from an Event

Based on the experiments presented in Section 6.10.6, Figure C.2 shows a set of images from a party. Here, the top row of each event (red background) contains the images ordered by timeline. The second line (white background) for each event is the human labeller preception ranking of the images in decreasing order of their happiness. The third line (green background) for each event is the images arranged on the based on decreasing happiness intensity as computed by the weighted Group Expression Model ($\mathbf{GEM}_w$, discussed in Section 6.10.6). Figure C.3 describes a reunion event.

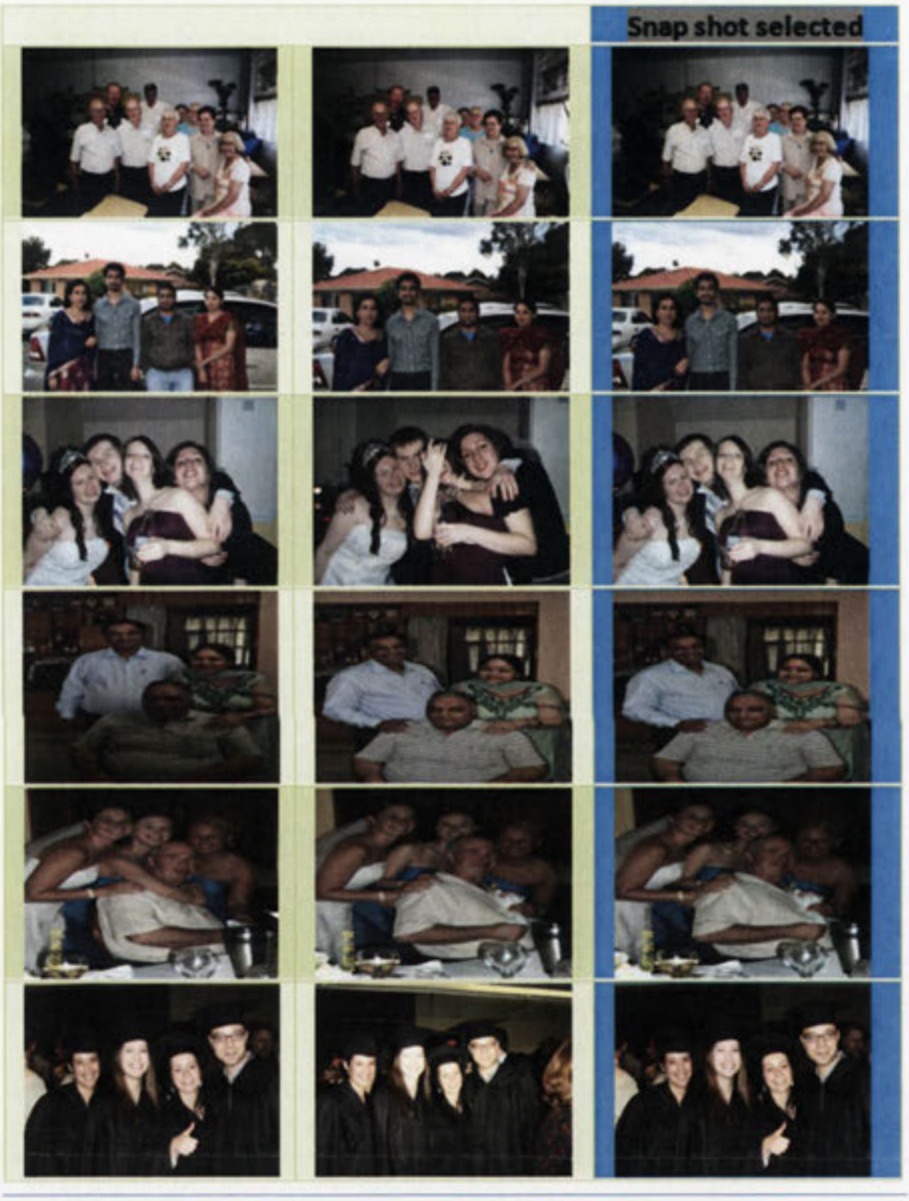


Figure C.1: Candid Group Shot Selection: Here, each row represents a series of photographs of same people. The third column is the selected shot based on the highest score from $\mathbf{GEM}_w$.

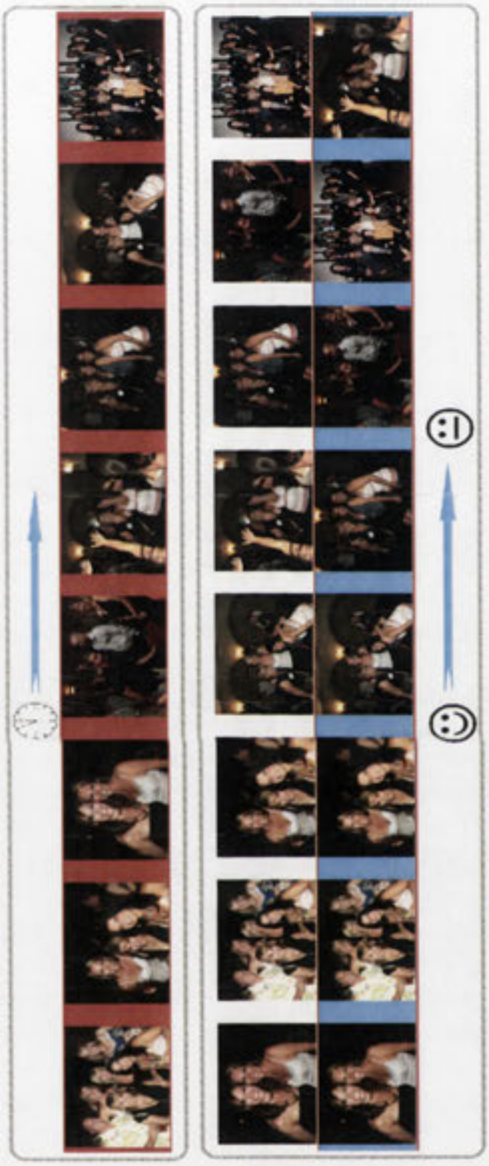


Figure C.2: This figure describes a party event. The top row (red background) are the images from a graduation ceremony organised by timestamps. The second row (white background) shows the images ranked on their decrease in intensity of happiness (from left to right) by human annotators. The third row (blue) shows the images ranked on their decrease in intensity of happiness (from left to right) by our method GEM_w.

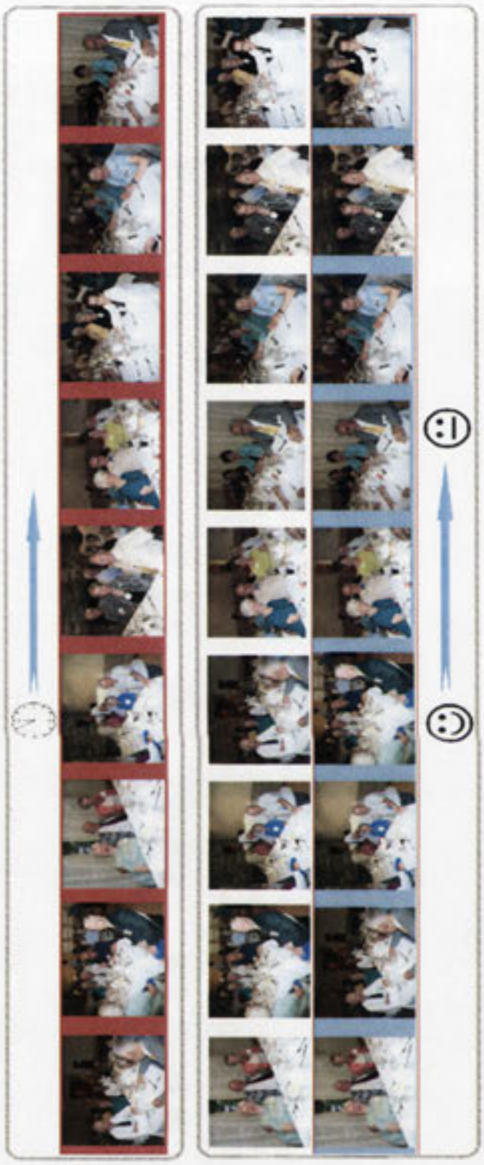


Figure C.3: This Figure describes a reunion party. The top row (red background) are the images from a graduation ceremony organised by timestamps. The second row (white background) shows the images ranked on their decrease in intensity of happiness (from left to right) by human annotators. The third row (blue) shows the images ranked on their decrease in intensity of happiness (from left to right) by our method GEM_w.

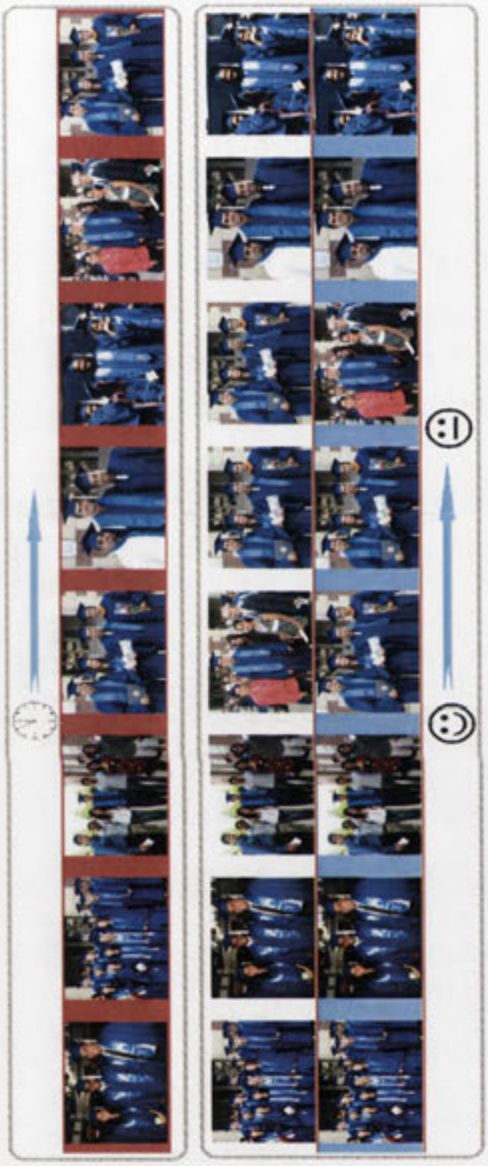


Figure C.4: High resolution version of Figure 3 in the main paper. This Figure describes a graduation ceremony. The top row (red background) are the images from a graduation ceremony organised by timestamps. The second row (white background) are the images ranked for their decrease in intensity of happiness (from left to right) by human annotators. The third row (blue) are the images ranked for their decrease in intensity of happiness (from left to right) by our method GEM_w.

Bibliography

Jacket : a gpu engine for matlab, 2009. `http://www.accelereyes.com`. 70

Timur Almaev and Michael Valstar. Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8, 2013. xvi, 15, 18, 22, 45

Timur R Almaev, Anıl Yüce, Alexandru Ghitulescu, and Michel F Valstar. Distribution-based iterative pairwise classification of emotions in the wild using lgbp-top. In *Proceedings of the ACM on International Conference on Multimodal Interaction (ICMI)*, pages 535–542, 2013. 44

Zara Ambadar, J. Schooler, and Jeffrey Cohn. Deciphering the enigmatic face: The importance of facial dynamics to interpreting subtle facial expressions. *Psychological Science*, pages 403–410, 2005. 20

Akshay Asthana. *The Utility of Synthetic Images for Face Modelling and Its Application*. PhD thesis, Australian National University, Canberra, Australia, April 2013. 11, 12

Akshay Asthana, Roland Goecke, Novi Quadrianto, and Tom Gedeon. Learning based automatic face annotation for arbitrary poses and expressions from frontal images only. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1635–1642, 2009a. 13, 25, 26, 48, 49, 50, 52, 53, 56, 58, 59, 60, 104

Akshay Asthana, Jason Saragih, Michael Wagner, and Roland Goecke. Evaluating AAM Fitting Methods for Facial Expression Recognition. In *Proceedings of the IEEE International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 598–605, 2009b. 15, 19

Akshay Asthana, Tim K. Marks, Michael J. Jones, Kinh H. Tieu, and M. V. Rohith. Fully automatic pose-invariant face recognition via 3d pose normalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 937–944, 2011. 25, 48, 49, 56, 104

Akshay Asthana, Miles de la Hunty, Abhinav Dhall, and Roland Goecke. Facial performance transfer via deformable models and parametric correspondence. *IEEE Transaction on Visualization and Computer Graphics*, 18(9):1511–1519, 2012. 14, 63

Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng, and Maja Pantic. Robust discriminative response map fitting with constrained local models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3444–3451, 2013a. 43

Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng, and Maja Pantic. Robust discriminative response map fitting with constrained local models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3444–3451, 2013b. 14

Simon Baker and Iain Matthews. Equivalence and Efficiency of Image Alignment Algorithms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1090–1097, 2001. 13

Simon Baker, Ralph Gross, and Iain Matthews. Lucas-Kanade 20 years on: A unifying framework: Part 3. Technical report, RI, Carnegie Mellon University, USA, 2003. 13

T. Bänziger and K.R. Scherer. Introducing the Geneva Multimodal Emotion Portrayal (GEMEP) Corpus. In K.R. Scherer, T. Bänziger, and E. Roesch, editors, *Blueprint for affective computing: A sourcebook*. Oxford, England: Oxford University Press, 2010. 24, 86

Marian Stewart Bartlett, Gwen Littlewort, Claudia Lainscsek, Ian Fasel, and Javier Movellan. Machine learning methods for fully automatic recognition of facial expressions and facial actions. In *Proceeding of the IEEE International Conference on Systems Man and Cybernetics (SMC)*, pages 592–597, 2004. 21

Marian Stewart Bartlett, Gwen Littlewort, Mark G. Frank, Claudia Lainscsek, Ian R. Fasel, and Javier R. Movellan. Automatic recognition of facial actions in spontaneous expressions. *Journal of Multimedia*, 1(6):22–35, 2006. 3, 23, 24

John N Bassili. Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face. *Journal Personality Socical Psychology*, 1979. 20

Nadia Berthouze and Lun Berthouze. Exploring kansei in multimedia information. *Kansei Engineering International*, 2(2):1–10, 2001. 107

Vinay Bettadapura, Grant Schindler, Thomas Plötz, and Irfan Essa. Augmenting bag-of-words: Data-driven discovery of temporal and structural information for activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2619–2626, 2013. 19

James C Bezdek. *Pattern recognition with fuzzy objective function algorithms*. Kluwer Academic Publishers, 1981. 73

Irving Biederman and Peter Kalocsais. Neurocomputational bases of object and face recognition. In *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, pages 1203–1219, 1997. 53

Michael J. Black, David J. Fleet, and Yaser Yacoob. A framework for modeling appearance change in image sequences. In *Proceedings of the IEEE Conference on Computer Vision (ICCV)*, pages 660–667, 1998. 20

Hubert M Blalock. *Quantitative Sociology: International perspectives on mathematical and statistical model building*, chapter Path models with latent variables: The NIPALS approach. Academic Press, 1975. 92

Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 187–194, 1999. 25

Volker Blanz, Curzio Basso, Thomas Vetter, and Tomaso Poggio. Reanimating faces in images and video. In *EUROGRAPHICS*, volume 22, pages 641–650, 2003. 74, 75

David M. Blei and Jon D. McAuliffe. Supervised Topic Models. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, 2007. 96

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, pages 601–608, 2001. xv, 94, 95

Liefeng Bo and Cristian Sminchisescu. Twin gaussian processes for structured prediction. *International Journal of Computer Vision*, 87(1-2):28–52, 2010. 51, 52

Aaron F. Bobick and James W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267, 2001. 21

Oren Boiman, Eli Shechtman, and Michal Irani. In defense of Nearest-Neighbor based image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008. 68

Anna Bosch, Andrew Zisserman, and Xavier Munoz. Representing Shape with a Spatial Pyramid Kernel. In *Proceedings of the ACM international conference on Image and video retrieval (CIVR)*, pages 401–408, 2007. xv, 18, 20, 39, 40, 92, 107

M. Caroll. How tumblr and pinterest are fueling the image intelligence problem. *Forbes*, Jaunary 2012. 81

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines, 2001. `http://www.csie.ntu.edu.tw/~cjlin/libsvm`. xvi, 96, 98

Yao-Jen Chang and Tony Ezzat. Transferable videorealistic speech animation. In *Proceedings of the 2005 ACM SIGGRAPH/Eurographics symposium on Computer Animation (SCA)*, pages 143–151, 2005. 74, 75

Sien W. Chew, Patrick Lucey, Simon Lucey, Jason M. Saragih, Jeffrey F. Cohn, Iain Matthews, and S. Sridharan. In the pursuit of effective affective computing: The relationship between features and registration. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 42(4):1006–1016, 2012. 13, 16, 52, 83

Ira Cohen, Nicu Sebe, Larry Chen, Ashutosh Garg, and Thomas S. Huang. Facial expression recognition from video sequences: Temporal and static modelling. *Computer Vision and Image Understanding*, 91(1):160–187, 2003. 9, 19

Timothy F. Cootes, Christopher J. Taylor, David H. Cooper, and Jim Graham. Active shape models-their training and application. *Computer Vision and Image Understanding*, 61(1): 38–59, 1995. xv, 10, 11, 47, 52

Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 681–685, 1998. 9, 10, 13, 65

Timothy F. Cootes, Gavin V. Wheeler, Kevin N. Walker, and Christopher J. Taylor. View-based active appearance models. *Image and Vision Computing Journal*, 20(9):657–664, 2002. xv, 25, 49

Navneet Dalal and Bill Triggs. Histograms of Oriented Gradients for Human Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 886–893, 2005a. xv, 17

Navneet Dalal and Bill Triggs. Histograms of Oriented Gradients for Human Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 886–893, 2005b. 49

Charles Darwin. *The expression of the emotions in man and animals.* Oxford University Press, 1998. 22

Matthew Day. Emotion recognition with boosted tree classifiers. In *Proceedings of the ACM on International Conference on Multimodal Interaction (ICMI)*, pages 531–534, 2013. 19, 45

Abhinav Dhall and Roland Goecke. Group expression intensity estimation in videos via gaussian processes. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pages 3525–3528, 2012. 81

Abhinav Dhall, Akshay Asthana, and Roland Goecke. Facial expression based automatic album creation. In *Proceedings of the Neural Information Processing. Models and Applications (ICONIP)*, pages 485–492, 2010. 5, 15, 30, 63, 70, 71

Abhinav Dhall, Akshay Asthana, and Roland Goecke. A ssim-based approach for finding similar facial expressions. In *Proceedings of the IEEE International Conference on Automatic Faces and Gesture Recognition and Workshop FERA*, pages 815–820, 2011a. 5, 83

Abhinav Dhall, Asthana Asthana, Roland Goecke, and Tom Gedeon. Emotion recognition using PHOG and LPQ features. In *Proceedings of the IEEE Conference Automatic Faces & Gesture Recognition workshop FERA*, pages 878–883, 2011b. 15, 19, 22, 63, 97

Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Static Facial Expression Analysis In Tough Conditions: Data, Evaluation Protocol And Benchmark. In *Proceedings of the IEEE International Conference on Computer Vision and Workshops BEFIT*, pages 2106–2112, 2011c. xvi, xxi, 5, 10, 17, 32, 36, 37, 40, 48, 58, 104, 107

Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Acted Facial Expressions in the Wild Database. In *Technical Report*, 2011d. 18

Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Collecting large, richly annotated facial-expression databases from movies. *IEEE Multimedia*, 19(3):0034, 2012a. xv, 2, 5, 24, 27, 32, 35, 37, 42, 87, 103

Abhinav Dhall, Jyoti Joshi, Ibrahim Radwan, and Roland Goecke. Finding Happiest Moments in a Social Context. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 613–626, 2012b. xv, 6, 24, 30, 81, 85, 87, 105

Abhinav Dhall, Roland Goecke, Jyoti Joshi, Michael Wagner, and Tom Gedeon. Emotion recognition in the wild challenge 2013. In *Proceedings of the ACM on International Conference on Multimodal Interaction (ICMI)*, pages 509–516, 2013. xv, xxiii, 6, 22, 32, 36, 41, 42, 103

Abhinav Dhall, Karan Sikka, Gwen Littlewort, Roland Goecke, and Marian Bartlett. A Discriminative Parts Based Model Approach for Fiducial Points Free and Shape Constrained Head Pose Normalisation In The Wild. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 1–8, 2014. 47

Hamdi Dibeklioğlu, Albert Ali Salah, and Theo Gevers. Are you really smiling at me? spontaneous versus posed enjoyment smiles. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 525–538, 2012. 24

Ellen Douglas-Cowie, Roddy Cowie, and Marc Schröder. A new emotion database: Considerations, sources and scope. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000. 23, 24

Samira Ebrahimi, Chris Pal, Xavier Bouthillier, Pierre Froumenty, Sbastien Jean, Kishore Reddy Konda, Pascal Vincent, Aaron Courville, and Yoshua Bengio. Combining modality specific deep neural networks for emotion recognition in video. In *Proceedings of the ACM on International Conference on Multimodal Interaction (ICMI)*, pages 543–550, 2013. 19, 44, 103

Marcin Eichner and Vittorio Ferrari. We Are Family: Joint Pose Estimation of Multiple Persons. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 228–242, 2010. 90

Paul Ekman. Facial expression and emotion. *American Psychologist*, 48(4):384, 1993. 22

Paul Ekman and W.V. Friesen. The facial action coding system: A technique for the measurement of facial movement. In *Consulting Psychologists*, 1978. xv, 10, 22

Heiner Ellgring. *Nonverbal communication in depression.* Cambridge University Press, 2008. 29

Mark Everingham, Josef Sivic, and Andrew Zisserman. Hello! My name is... Buffy" – Automatic Naming of Characters in TV Video. In *Proceedings of the British Machine and Vision Conference (BMVC)*, pages 899–908, 2006. 14, 83, 96

Florian Eyben, Martin Wollmer, and Bjorn Schuller. Openearintroducing the munich open-source emotion and affect recognition toolkit. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–6, 2009. 43

Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the ACM International Conference on Multimedia (MM)*, pages 1459–1462, 2010. 43

Beat Fasel and Juergen Luettin. Automatic Facial Expression Analysis: A Survey. *Pattern Recognition*, 36:259–275, 2003. 9

Pedro F Felzenszwalb and Daniel P Huttenlocher. Pictorial Structures for Object Recognition. *International Journal on Computer Vision*, 61(1):55–79, 2005. 13, 47, 49

Juliet Fiss, Aseem Agarwala, and Brian Curless. Candid portrait selection from video. *ACM Transaction on Graphics*, page 128, 2011. 28

Joseph P Forgas. Affect in social judgments and decisions: A multiprocess model. *Advances in experimental social psychology*, 25:227–275, 1992. 83

Yoav Freund and Robert E Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *Computational learning theory*, pages 23–37. Springer, 1995. 10

Andrew C Gallagher and Tsuhan Chen. Understanding Images of Groups of People. In *Proceedings of the IEEE Confernece on Computer Vision and Pattern Recognition (CVPR)*, pages 256–263, 2009. 4, 27

Weina Ge, Robert T. Collins, and Barry Ruback. Vision-based analysis of small groups in pedestrian crowds. *IEEE Transaction on Pattern Analysis & Machine Intelligence*, 34(5): 1003–1016, 2012. 27

Tobias Gehrig and Hazim Kemal Ekenel. Facial action unit detection using kernel partial least squares. In *Proceedings of the IEEE International Confernece on Computer Vision and Workshops (ICCW)*, pages 2092–2099, 2011. 93

Roland Goecke. The Audio-Video Australian English Speech Data Corpus AVOZES. Documentation, Australian National University, 2004. 64

Roland Goecke and J Bruce Millar. The Audio-Video Australian English Speech Data Corpus AVOZES. In *Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP)*, pages 2525–2528, 2004. xv, 64, 77

Jacob Goldberger, Sam T. Roweis, Geoffrey E. Hinton, and Ruslan Salakhutdinov. Neighbourhood Components Analysis. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, 2004. 64

Michael Grimm, Kristian Kroschel, and Shrikanth Narayanan. The vera am mittag german audio-visual emotional speech database. In *Proceeding of the IEEE International Confernce on Multimedia & Expo (ICME)*, pages 865–868, 2008. 24, 86

Ralph Gross, Iain Matthews, Jeffrey F. Cohn, Takeo Kanade, and Simon Baker. Multi-PIE. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–8, 2008a. xvi, 3, 23, 24, 48, 64, 68, 86, 103

Ralph Gross, Iain Matthews, Jeffrey F. Cohn, Takeo Kanade, and Simon Baker. Multi-PIE. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–8, 2008b. 26, 56, 57

Hatice Gunes and Maja Pantic. Automatic, dimensional and continuous emotion recognition. *International Journal of Synthetic Emotions*, 1(1):68–99, 2010. 23

Hatice Gunes, Björn Schuller, Maja Pantic, and Roddy Cowie. Emotion representation, analysis and synthesis in continuous space: A survey. In *Proceedings of the IEEE Conference on Automatic Face & Gesture Recognition and Workshops (FG)*, pages 827–834, 2011. 23

Guodong Guo and Guowang Mu. Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 657–664, 2011. 91

Munawar Hayat, Mohammed Bennamoun, and Amar A El-Sallam. Clustering of video-patches on grassmannian manifold for facial expression recognition from 3d videos. In *Proceedings of the Workshop on Applied Computer Vision*, pages 83–88, 2013. 19

Javier Hernandez, Mohammed Ehsan Hoque, Will Drevo, and Rosalind W Picard. Mood meter: counting smiles in the wild. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 301–310, 2012. 27

Erik Hjelmås and Boon Kee Low. Face detection: A survey. *Computer Vision and Image Understanding*, 83(3):236–274, 2001. 10

Mohammed Ehsan Hoque and Rosalind W Picard. Acted vs. natural frustration and delight: Many people smile in natural frustration. In *Proceedings of the IEEE International Conference on Automatic Faces & Gesture Recognition (FG)*, pages 354–359, 2011. 31

Yuxiao Hu, Zhihong Zeng, Lijun Yin, Xiaozhou Wei, Xi Zhou, and Thomas S. Huang. Multi-view facial expression recognition. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–6, 2008. 25, 26

Chang Huang, Haizhou Ai, Yuan Li, and Shihong Lao. Vector boosting for rotation invariant multi-view face detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 446–453, 2005. 10

Di Huang, Caifeng Shan, Mohsen Ardabilian, Yunhong Wang, and Liming Chen. Local binary patterns and its application to facial image analysis: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, pages 1 –17, 2011. 39, 40

Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007. 24, 31, 73, 74, 105

Vidit Jain, Erik G. Learned-Miller, and Andrew McCallum. People-lda: Anchoring topics to people using face recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1–8, 2007. 94, 95

Bihan Jiang, Michel Valstar, and Maja Pantic. Action unit detection using sparse appearance descriptors in space-time video volumes. In *Proceedings of the IEEE International Conference on Autmatic Face and Gesture Recognition (FG)*, pages 314–321, 2011. 18

Yu-Gang Jiang, Chong-Wah Ngo, and Jun Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *Proceedings of the ACM international conference on Image and Video Retrieval (CIVR)*, pages 494–501, 2007. 95

Hideo Joho, Jacopo Staiano, Nicu Sebe, and Joemon M Jose. Looking at the viewer: analysing facial activity to detect personal highlights of multimedia contents. *Multimedia Tools and Applications*, 51(2):505–523, 2011. 29

Michael Jones and Paul Viola. Fast multi-view face detection. *Mitsubishi Electric Research Lab TR-20003-96*, 3:14, 2003. 10

Jyoti Joshi, Abhinav Dhall, Roland Goecke, Michael Breakspear, and Gordon Parker. Neural-net classification for spatio-temporal descriptor based depression analysis. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pages 2634–2638, 2012. 14, 29

Susanne Kaiser and Thomas Wehrle. Automated coding of facial behavior in human-computer interaction with FACS. *Journal of Nonverbal Behavior*, pages 67–84, 1992. 20

Ioannis A. Kakadiaris, George Passalis, George Toderici, Yunliang Lu, Nikos Karampatziakis, Najam Murtuza, and Theoharis Theoharis. Expression-invariant multispectral face recognition: you can smile now! *Biometric Technology for Human Identification III*, 6202(1): 620204, 2006. 68

Takeo Kanade, Jeffrey F. Cohn, and Yingli Tian. Comprehensive database for facial expression analysis. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 46–53, 2000. xv, 3, 20, 21, 23, 24

Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331, 1988. 11

Janice R Kelly and Sigal G Barsade. Mood and emotions in small groups and work teams. *Organizational behavior and human decision processes*, 86(1):99–130, 2001. 27, 82

Ira Kemelmacher-Shlizerman, Aditya Sankar, Eli Shechtman, and Steven M. Seitz. Being john malkovich. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 341–353, 2010. 29, 78, 79

Asim Khwaja, Akshay Asthana, and Roland Goecke. Illumination and Expression Invariant Face Recognition Using SSIM Based Sparse Representation. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pages 4028–4031, 2010. 64

Andrea Kleinsmith and Nadia Bianchi-Berthouze. Affective body expression perception and recognition: a survey. *IEEE Transactions on Affective Computing*, 4(1):15–33, 2013. 106, 107

Filip Korc and David Schneider. Annotation tool. Technical Report TR-IGG-P-2007-01, University of Bonn, Department of Photogrammetry, 2007. 88

Tarun Krishna, Ayush Rai, Shubham Bansal, Shubham Khandelwal, Shubham Gupta, and Dushyant Goel. Emotion recognition using facial and audio features. In *Proceedings of the ACM on International Conference on Multimodal Interaction (ICMI)*, pages 557–564, 2013. 45

Christian Küblbeck and Andreas Ernst. Face detection and tracking in video sequences using the modifiedcensus transformation. *Image Vision Computing*, 24(6):564–572, 2006. 27

Ivan Laptev, Marcin Marszałek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008. 4, 24, 31

Fabio Lavagetto and Roberto Pockaj. The Facial Animation Engine: towards a high-level interface for the design of MPEG-4 compliant animated faces. *IEEE Transactions on Circuits and Systems for Video Technology*, 9(2):277–289, 1999. 15

Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2169–2178, 2006. 57

Yong Jae Lee and Kristen Grauman. Face discovery with social context. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 1–11, 2011. 28

FeiFei Li and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 524–531, 2005. 94

Zisheng LI, Jun ichi IMAI, and Masahide Kaneko. Facial-component-based bag of words and PHOG descriptor for facial expression recognition. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pages 1353–1358, 2009. 18, 19, 20

Jenn-Jier James Lien. *Automatic Recognition of Facial Expressions Using Hidden Markov Models and Estimation of Expression Intensity.* PhD thesis, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, April 1998. 21

Dahua Lin and Xiaoou Tang. Quality-Driven Face Occlusion Detection and Recovery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–7, 2007. 90, 91

Mengyi Liu, Shaoxin Li, Shiguang Shan, and Xilin Chen. Enhancing expression recognition in the wild with unlabeled reference data. In *Proceedings of the Asian Confernce on Computer Vision (ACCV)*, pages 577–588, 2012. 22

Mengyi Liu, Ruiping Wang, Zhiwu Huang, Shiguang Shan, and Xilin Chen. Partial least squares regression on grassmannian manifold for emotion recognition. In *Proceedings of the ACM on International Conference on Multimodal Interaction (ICMI)*, pages 525–530, 2013. 44

Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition and Workshops (CVPRW)*, pages 94–101, 2010. xv, 23, 24, 31, 38, 41, 42, 86, 103

Patrick Lucey, Jeffrey F Cohn, Kenneth M Prkachin, Patricia E Solomon, and Iain Matthews. Painful data: The unbc-mcmaster shoulder pain expression archive database. In *Proceedings of the IEEE Conference on Automatic Face & Gesture Recognition and Workshops (FG)*, pages 57–64, 2011. 29

Michael J. Lyons, Shigeru Akamatsu, Miyuki Kamachi, and Jiro Gyoba. Coding facial expressions with gabor wavelets. In *Proceedings of the IEEE International Conference on Automatic Face Gesture Recognition and Workshops (fg)*, pages 200–205, 1998. 20, 23, 31, 36

Ohil K. Manyam, Neeraj Kumar, Peter N. Belhumeur, and David J. Kriegman. Two faces are better than one: Face recognition in group photographs. In *Proceedings of the International Joint Conference on Biometrics (IJCB)*, pages 1–8, 2011. 28

Gary McKeown, Michel François Valstar, Roderick Cowie, and Maja Pantic. The semaine corpus of emotionally coloured character interactions. In *Proceedings of the IEEE International Confernece on Multimedia & Expo (ICME)*, pages 1079–1084, 2010. 23, 24, 41

Hongying Meng and Nadia Bianchi-Berthouze. Naturalistic affective expression classification by a multi-stage approach based on hidden markov models. In *Proceedings of the International Confernece on Affective Computing and Intelligent Interaction (ACII)*, pages 378–387, 2011. 19

Hongying Meng, Bernardino Romera-Paredes, and Nadia Bianchi-Berthouze. Emotion recognition by two view svm_2k classifier on dynamic facial expression features. In *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG)*, pages 854–859, 2011. 19

Sascha Meudt, Dimitri Zharkov, Markus Kächele, and Friedhelm Schwenker. Multi classifier systems and forward backward feature selection algorithms to classify emotional coloured speech. In *Proceedings of the ACM on International Conference on Multimodal Interaction (ICMI)*, pages 551–556, 2013. 45

Stephen Moore and Richard Bowden. Local binary patterns for multi-view facial expression recognition. *Computer Vision and Image Understanding*, 115(4):541–558, 2011. 25, 26, 49

Ana C. Murillo, Iljung S. Kwak, Lubomir Bourdev, David J. Kriegman, and Serge Belongie. Urban tribes: Analyzing group photos from a social perspective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition and Workshops (CVPRW)*, pages 28–35, 2012. 27

Jun-yong Noh and Ulrich Neumann. Expression cloning. In *Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 277–288, 2001. 74, 75

Timo Ojala, Matti Pietikinen, and Topi Menp. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002. xv, 16, 79, 107

Ville Ojansivu and Janne Heikkilä. Blur Insensitive Texture Classification Using Local Phase Quantization. In *Proceedings of the Image and Signal Processing (ICISP)*, pages 236–243, 2008. 16, 17, 18

Margarita Osadchy, Yann Le Cun, and Matthew L Miller. Synergistic face detection and pose estimation with energy-based models. *Journal of Machine Learning Research*, 8:1197–1215, May 2007. 10

Alice J. O'Toole, Joshua Harms, Sarah L. Snow, Dawn R. Hurst, Matthew R. Pappas, Janet H. Ayyad, and Herve Abdi. A video database of moving faces and people. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):812–816, 2005. 24

Takahiro Otsuka and Jun Ohya. Recognition of facial expressions using hmm with continuous output probabilities. In *Proceedings of the IEEE International Workshop on Robot and Human Communication (RHC*, pages 323–328, 1996. 20

Takahiro Otsuka and Jun Ohya. Recognizing multiple persons' facial expressions using hmm based on automatic extraction of significant frames from image sequences. In *Proceedings of the IEEE International Conference on Image Processing (ICIP*, pages 546–549, 1997. 21

Marco Paleari, Ryad Chellali, and Benoit Huet. Bimodal emotion recognition. In *Proceedings of the International Conference on Social Robotics, November 23-24, 2010, Singapore)*, pages 305–314, 2010. 23, 24

Maja Pantic and Leon Rothkrantz. Expert system for automatic analysis of Facial Expression. *Image and Vision Computing Journal*, 18(11):881–905, July 2000. ISSN 0262-8856. 15

Maja Pantic and L.J.M. Rothkrantz. Facial action recognition for facial expression analysis from static face images. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 34(3):1449–1461, 2004. 9, 19

Maja Pantic, Ioannis Patras, and L Rothkruntz. Facial action recognition in face profile image sequences. In *Proceedings of the IEEE International Conference on Multimedia & Expo (ICME)*, pages 37–40, 2002. 21

Maja Pantic, Ioannis Patras, and Michel Valstar. Learning spatiotemporal models of facial expressions. In *Int'l Conf. Measuring Behaviour 2005*, pages 7–10, August 2005a. URL `http://pubs.doc.ic.ac.uk/Pantic-MB05/`. 21

Maja Pantic, Michel François Valstar, Ron Rademaker, and Ludo Maat. Web-based database for facial expression analysis. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pages 5–pp, 2005b. 23, 24, 31, 41, 42

Devi Parikh and Kristen Grauman. Relative attributes. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 503–510, 2011. 94

Devi Parikh, C. Lawrence Zitnick, and Tsuhan Chen. From appearance to context-based recognition: Dense labeling n small images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008. 28

Robert Clay Prim. Shortest connection networks and some generalizations. *Bell system technical journal*, 36(6):1389–1401, 1957. 89

Carl Edward Rasmussen. *Gaussian processes for machine learning*. MIT Press, 2006. xv, 26, 51

Matthias Richter, Tobias Gehrig, and Hazim Kemal Ekenel. Facial expression classification on web images. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pages 3517–3520, 2012. 23

Mark Rosenblum, Yaser Yacoob, and Larry S Davis. Human expression recognition from motion using a radial basis function network architecture. *IEEE Transactions on Neural Networks*, 7(5):1121–1138, September 1996. 20

Roman Rosipal. *Chemoinformatics and Advanced Machine Learning Perspectives: Complex Computational Methods and Collaborative Techniques*, chapter Nonlinear Partial Least Squares: An Overview. ACCM, IGI Global, 2011. 91, 92

Ognjen Rudovic and Maja Pantic. Shape-constrained gaussian process regression for facial-point-based head-pose normalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1495–1502, 2011. 26, 48, 49, 51, 52, 58, 59, 60, 104

Ognjen Rudovic, Ioannis Patras, and Maja Pantic. Coupled gaussian process regression for pose-invariant facial expression recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 350–363, 2010a. 25, 26, 48, 49, 50, 52, 56, 104

Ognjen Rudovic, Ioannis Patras, and Maja Pantic. Regression-based multi-view facial expression recognition. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pages 4121–4124, 2010b. 25, 26, 48, 49, 50, 51, 52, 53, 56, 58, 59, 60, 104, 105

Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, and William T. Freeman. Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision*, 77(1-3):157–173, 2008. 88

Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. A semi-automatic methodology for facial landmark annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 896–903, 2013. 13

Jason Saragih and Roland Göcke. Iterative error bound minimisation for aam alignment. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pages 1192–1195, 2006. xv, 13

Jason Saragih and Roland Göcke. A nonlinear discriminative approach to aam fitting. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1–8, 2007. 13

Jason Saragih and Roland Goecke. A Nonlinear Discriminative Approach to AAM Fitting. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1–8, 2007. 11, 12

Jason Saragih and Roland Goecke. Learning AAM fitting through simulation. *Pattern Recognition*, 42(11):2628–2636, 2009. xv, 14, 65, 69

Jason M. Saragih, Simon Lucey, and Jeffrey Cohn. Face alignment through subspace constrained mean-shifts. In *International Conference of Computer Vision (ICCV)*, September 2009. 10, 13, 83

Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian A Müller, and Shrikanth S Narayanan. The interspeech 2010 paralinguistic challenge. In *INTERSPEECH*, pages 2794–2797, 2010. 43

Björn Schuller, Michel Valstar, Florian Eyben, Gary McKeown, Roddy Cowie, and Maja Pantic. Avec 2011–the first international audio/visual emotion challenge. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 415–424, 2011. 41, 42

Björn Schuller, Michel Valster, Florian Eyben, Roddy Cowie, and Maja Pantic. Avec 2012: the continuous audio/visual emotion challenge. In *Proceedings of the ACM on International Conference on Multimodal Interaction (ICMI)*, pages 449–456, 2012. 41

William Robson Schwartz, Aniruddha Kembhavi, David Harwood, and Larry S. Davis. Human detection using partial least squares analysis. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 24–31, 2009. 91

William Robson Schwartz, Huimin Guo, and Larry S. Davis. A Robust and Scalable Approach to Face Identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 476–489, 2010. 91

Nicu Sebe, Ira Cohen, Theo Gevers, and Thomas S Huang. Emotion recognition based on joint visual and audio cues. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pages 1136–1139, 2006. 19

Nicu Sebe, Michael S Lew, Yafei Sun, Ira Cohen, Theo Gevers, and Thomas S Huang. Authentic facial expression analysis. *Image and Vision Computing*, 25(12):1856–1863, 2007. 15, 29

Caifeng Shan, Shaogang Gong, and Peter W. McOwan. Conditional mutual information based boosting for facial expression recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 1–8, 2005. 19

Karan Sikka, Tingfan Wu, Josh Susskind, and Marian Bartlett. Exploring bag of words architectures in the facial expression domain. In *Proceedings of the European Conference on Computer Vision and Workshops (ECCVW)*, pages 250–259, 2012. 15, 19, 22

Karan Sikka, Abhinav Dhall, and Marian Bartlett. Weakly supervised pain localization using multiple instance learning. *Proceedings of the IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8, 2013a. 29, 106

Karan Sikka, Karmen Dykstra, Suchitra Sathyanarayana, Gwen Littlewort, and Marian Bartlett. Multiple kernel learning for emotion recognition in the wild. In *Proceedings of the ACM on International Conference on Multimodal Interaction (ICMI)*, pages 517–524, 2013b. 44, 103

Terence Sim, Simon Baker, and Maan Bsat. The CMU Pose, Illumination, and Expression Database. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1615–1618, 2003. xv, 24

Abu Sayeed Md Sohail and Prabir Bhattacharya. Classification of facial expressions using k-nearest neighbor classifier. In *MIRAGE07*, pages 555–566, 2007. 20, 64

Mingli Song, Zicheng Liu, and Baining Guo. Real-time facial expression mapping for high resolution 3d meshes. In *Computer Graphics International*, pages 277–287, 2006. 74, 75

Zak Stone, Todd Zickler, and Trevor Darell. Autotagging facebook: Social network context improves photo annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008. 28

Robert W Sumner and Jovan Popović. Deformation transfer for triangle meshes. In *Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 399–405, 2004. 75

Usman Tariq, Kai-Hsiang Lin, Zhen Li, Xi Zhou, Zhaowen Wang, Vuong Le, Thomas S Huang, Xutao Lv, and Tony X Han. Recognizing emotions from an ensemble of features. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 42(4):1017–1026, 2012. 15

Ying-Li Tian, Takeo Kanade, and Jeffrey Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(1):97 – 115, February 2001. 19

Ying-Li Tian, Takeo Kanade, and Jeffrey F Cohn. Facial expression analysis. In *Handbook of face recognition*, pages 247–275, 2005. 1

Antonio Torralba and Pawan Sinha. Statistical context priming for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 763–770, 2001. 28

Grace Tsai, Changhai Xu, Jingen Liu, and Benjamin Kuipers. Real-time indoor scene understanding using bayesian filtering with motion cues. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 121–128, 2011. 94

Michel Valstar and Maja Pantic. Fully automatic facial action unit detection and temporal analysis. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, pages 149–149, 2006. 15, 16, 21

Michel Valstar, Bihan Jiang, Marc Mehu, Maja Pantic, and Scherer Klaus. The first facial expression recognition and analysis challenge. In *Proceedings of the IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG)*, pages 314–321, 2011. 40, 41, 42

Michel F Valstar, Ioanis Patras, and Maja Pantic. Facial action unit detection using probabilistic actively learned support vector machines on tracked facial point data. In *Proccedings of the Conference on Computer Vision and Pattern Recognition-Workshops (CVPRW)*, pages 76–76, 2005. 15, 19, 21, 103

Michel François Valstar, Marc Mehu, Bihan Jiang, Maja Pantic, and Klaus Scherer. Meta-analysis of the first facial expression recognition challenge. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 42(4):966–979, 2012. 22

Paul A. Viola and Michael J. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages I–511, 2001. 9, 10, 57, 65, 96

Carl Vondrick, Aditya Khosla, Tomasz Malisiewicz, and Antonio Torralba. Hoggles: Visualizing object detection features. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1–8, 2013. 17, 18

Frank Wallhoff. Facial expressions and emotion database, 2006. `http://www.mmk.ei.tum.de/~waf/fgnet/feedtum.html`. 3, 24, 64, 66, 68, 73, 74, 79, 86, 105

Gang Wang, Andrew C. Gallagher, Jiebo Luo, and David A. Forsyth. Seeing people in social context: Recognizing people and social relationships. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 169–182, 2010a. 28

Shangfei Wang, Zhilei Liu, Siliang Lv, Yanpeng Lv, Guobing Wu, Peng Peng, Fei Chen, and Xufa Wang. A natural visible and infrared facial expression database for expression recognition and emotion inference. *IEEE Transactions on Multimedia*, 12(7):682–691, 2010b. 24

Yubo Wang, Haizhou Ai, Bo Wu, and Chang Huang. Real time facial expression recognition with adaboost. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pages 926–929, 2004a. 19, 20

Z Wang and E P Simoncelli. Translation insensitive image similarity in the complex wavelet domain. In *ICASSP*, volume II, pages 573–576, Philadelphia, PA, 18-23 Mar 2005. IEEE Sig Proc Society. 68

Zhou Wang, Alan Conrad Bovik, Hamid Rahim Sheikh, Student Member, Eero P. Simoncelli, and Senior Member. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004b. xvi, 15, 63, 64, 67, 105

Kilian Q. Weinberger and Lawrence K. Saul. Distance Metric Learning for Large Margin Nearest Neighbor Classification. *Journal of Machine Learning Research*, 10:207–244, 2009. 22, 64, 69

Jacob Whitehill, Gwen Littlewort, Ian R. Fasel. Marian Stewart Bartlett, and Javier R. Movellan. Toward Practical Smile Detection. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 31(11):2106–2111, 2009. 19, 24, 28, 87

Matthias Wimmer, Bruce A MacDonald, Dinuka Jayamuni, and Arpit Yadav. Facial expression recognition for human-robot interaction–a prototype. In *Robot Vision*, pages 139–152. Springer, 2008. 29

Jianxin Wu, S Charles Brubaker, Matthew D Mullin, and James M Rehg. Fast asymmetric learning for cascade face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(3):369–382, 2008. 10

Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 532–539, 2013. 43

Liefei Xu and Philippos Mordohai. Automatic facial expression recognition using bags of motion words. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 1–13, 2010. 19

Yaser Yacoob and Larry Davis. Computing spatio-temporal representations of human faces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 70–75, 1994. 20

Peng Yang, Qingshan Liu, and Dimitris Metaxas. Similarity features for facial event analysis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 685–696, 2008. 21, 22

Peng Yang, Qingshan Liu, and Dimitris N Metaxas. Rankboost with l1 regularization for facial expression recognition and intensity estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1018–1025, 2009. 22

Peng Yang, Qingshan Liu, and Dimitris N. Metaxas. Exploring facial expressions with compositional features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2638–2644, 2010. 22

Yi Yang and Deva Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1385–1392, 2011. 49

Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J. Rosato. A 3d facial expression database for facial behavior research. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 211–216, 2006. 26

Kaimin Yu, Zhiyong Wang, Li Zhuo, Jiajun Wang, Zheru Chi, and Dagan Feng. Learning realistic facial expressions from web images. *Pattern Recognition*, 2013. 24, 25

Zhihong Zeng, M. Pantic, G.I. Roisman, and T.S. Huang. A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009. 9

Cha Zhang and Zhengyou Zhang. A survey of recent advances in face detection. Technical report, Technical Report, Microsoft Research, 2010. 10

Cha Zhang, John C Platt, and Paul A Viola. Multiple instance boosting for object detection. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, pages 1417–1424, 2005a. 29

Ligang Zhang, Dian Tjondronegoro, and Vinod Chandran. Evaluation of texture and geometry for dimensional facial expression recognition. In *Proceedings of the International Conference on Digital Image Computing Techniques and Applications (DICTA)*, pages 620–626, 2011. 15

Qingshan Zhang, Zicheng Liu, Baining Guo, Demetri Terzopoulos, and Heung-Yeung Shum. Geometry-driven photorealistic facial expression synthesis. *IEEE Transactions on Visualization and Computer Graphics*, 12(1):48–60, 2006. 74, 75

Shiliang Zhang, Qi Tian, Qingming Huang, Wen Gao, and Shipeng Li. Utilizing affective analysis for efficient movie browsing. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 1853–1856, 2009. 28

Wenchao Zhang, Shiguang Shan, Wen Gao, Xilin Chen, and Hongming Zhang. Local Gabor binary pattern histogram sequence (LGBPHS): A novel non-statistical model for face representation and recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 786–791, 2005b. xv, 18

Guoying Zhao and Matti Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 29(6):915–928, 2007. xv, 14, 15, 16, 18, 29

Feng Zhou, Jonathan Brandt, and Zhe Lin. Exemplar-based graph matching for robust facial landmark localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1–8, 2013. 14

Jun Zhu, Amr Ahmed, and Eric P. Xing. Medlda: maximum margin supervised topic models for regression and classification. In *Proceedings of the International Conference on Machine Learning (ICML)*, page 158, 2009. 96

Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2879–2886, 2012. 4, 6, 14, 43, 47, 48, 49, 50, 57, 83, 115