Journal of Biomechanics 55 (2017) 1–10

Contents lists available at ScienceDirect

Journal of Biomechanics

journal homepage: www.elsevier.com/locate/jbiomech
www.JBiomech.com

A learning-based markerless approach for full-body kinematics estimation *in-natura* from a single image

Ami Drory[a,*], Hongdong Li[a,c], Richard Hartley[a,b,c]

[a] Australian National University, Canberra, Australia
[b] Data61, CSIRO, Canberra, Australia
[c] Australian Centre for Robotic Vision, Australia

ARTICLE INFO

Article history:
Accepted 14 January 2017

Keywords:
Pose estimation
Skeletal kinematics
Markerless motion capture
Mixture of parts
Cycling

ABSTRACT

We present a supervised machine learning approach for markerless estimation of human full-body kinematics for a cyclist from an unconstrained colour image. This approach is motivated by the limitations of existing marker-based approaches restricted by infrastructure, environmental conditions, and obtrusive markers. By using a discriminatively learned mixture-of-parts model, we construct a probabilistic tree representation to model the configuration and appearance of human body joints. During the learning stage, a Structured Support Vector Machine (SSVM) learns body parts appearance and spatial relations. In the testing stage, the learned models are employed to recover body pose via searching in a test image over a pyramid structure. We focus on the movement modality of cycling to demonstrate the efficacy of our approach. *In natura* estimation of cycling kinematics using images is challenging because of human interaction with a bicycle causing frequent occlusions. We make no assumptions in relation to the kinematic constraints of the model, nor the appearance of the scene. Our technique finds multiple quality hypotheses for the pose. We evaluate the precision of our method on two new datasets using loss functions. Our method achieves a score of 91.1 and 69.3 on mean Probability of Correct Keypoint (PCK) measure and 88.7 and 66.1 on the Average Precision of Keypoints (APK) measure for the frontal and sagittal datasets respectively. We conclude that our method opens new vistas to robust user-interaction free estimation of full body kinematics, a prerequisite to motion analysis.

1. Introduction

1.1. Motivation - Deficiencies of marker based mocap

Characterizing the non-linear behaviour of human motion enhances the understanding of neuromuscular coordination patterns and dysfunction. Using inverse dynamics or dynamic optimisation, resultant compressive and shear loads and muscle contributions to segment and joint accelerations can be estimated based on the measured kinetics, inertial properties and skeletal kinematics. Obtaining skeletal kinematics is currently limited mostly to marker-based motion capture systems. This is unsatisfactory because the approach is constrained by expansive laboratory infrastructure with camera array, control of lighting and environmental conditions, and the obtrusive use of markers requiring palpation. Inherent to the use of surface mounted markers are output errors caused by a critical reliance on a strong assumption of rigid linkage skeletal system and ignoring surface deformation (Challis, 1995; Hatze, 2002; Cappozzo et al., 2005). In particular, the effect of Soft Tissue Artefacts (STA) causing movement impediment has received extensive attention in the literature (Leardini et al., 2005; Cutti et al., 2005; Riemer et al., 2008; Camomilla et al., 2009; Andersen et al., 2010; Peters et al., 2010; Rosario et al., 2012; Li et al., 2012; Miranda et al., 2013; Grimpampi et al., 2014; Camomilla et al., 2015), as has the precision of anatomical landmark determination (Lu and O'connor, 1999; Della Croce et al., 2005; Taylor et al., 2005; Ehrig et al., 2006; Taylor et al., 2010). Consequently, the development of evidence-based decision support tools for diagnosis and treatment is inhibited. Hence, the development of a markerless solution for acquisition of full body kinematics has attracted significant research efforts.

1.2. Previous work

1.2.1. Kinematics estimation from images

Estimation of the full body human kinematics from monocular images remains an open problem. The difficulties stem from

[*] Corresponding author.
E-mail address: ami.drory@anu.edu.au (A. Drory).

http://dx.doi.org/10.1016/j.jbiomech.2017.01.028
0021-9290/© 2017 Elsevier Ltd. All rights reserved.

background clutter, scene illumination and the weak local appearance support, which is further hindered by out-of-plane motion and severe occlusions caused by the motion of the articulated body (Gupta et al., 2008). Since 2D intensity images remain the most readily obtainable for capture of unrestricted motion *in-natura*, feature tracking via direct manual digitization has formed the most common form of analysis. Krosshaug and Bahr (2005) reconstructed motion kinematics from uncalibrated images using manual annotation of anatomical landmark locations that was matched across camera views and applied to a subject-specific scaled anatomical model with joint constraints. Likewise, Sanders et al. (2016) have shown high repeatability of manual 3D marker trajectories digitised from multi view swimming images. Magalhaes et al. (2013) attempted to automatically track surface mounted markers underwater using optical flow with limited success. Using textured clothing to replace surface mounted markers approach Lerasle et al. (1997) tracked low level image features of a cycling leg using a Kalman filter. Similarly, Sandau et al. (2014) used a texture enhanced clothing aided by background subtraction to achieve point correspondences for surface reconstruction in a calibrated multi-view camera setup. They fitted an articulated model to the 3D surface reconstruction using a patch matching technique, which enforces local photometric consistency and global visibility constraints.

1.2.2. Computer vision and machine learning approaches

In generative approaches, pose estimation is formulated as an optimisation problem whose objective function is a discrepancy between a parametric prior body model and the input observation (Baak et al., 2013; Fastovets et al., 2013; Salzmann et al., 2007 (for review, see Yang et al. (2014))). This approach, however, suffers from local minima and solution multiplicity due to its often highly non-convex nature. For instance, Corazza et al. (2006) fitted prior articulated model to a 3D surface visual hull reconstruction using patch matching with high accuracy. They used body part segmentation and least-squares optimisation to identify the location of joint centres under the assumption of rigid links connected by pivot joints (Corazza et al., 2007) and to estimate the centre of mass (Corazza and Andriacchi, 2009). The same method was modified to use adaptive Gaussian mixture models to enhance background subtraction for the pose estimation in a water environment (Ceseracciu et al., 2011). Notably, the visual hull approach tends to overestimate the volume of the subject and fails to reconstruct cavities in the subject's surface. Whilst less obtrusive than marker-based methods, the method critically relies on background subtraction and a constrained capture space. This requires considerable control over lighting and environmental conditions, and remains unsuitable for estimation of kinematics in realistic natural environments.

In contrast, discriminative approaches seek a mapping from image observation space to a set of body pose parameters space, from which the kinematics can be estimated (Agarwal and Triggs, 2006). The pictorial structures framework uses a probabilistic graph model to model the appearance and configuration of body parts. Pose estimation can then be formulated as a statistical inference problem, where the model parameters are learned from training examples using maximum likelihood estimation (Felzenszwalb and Huttenlocher, 2005). This powerful framework allows for efficient inference and captures large variations in posture and appearance. The inter-part relative deformation term makes this framework invariant to some global transformation. Additionally, the overall decision is made with no assumptions being made about the initial location of parts. For these reasons, the approach has been popular for simultaneous human detection and pose estimation tasks (Andriluka et al., 2009; Eichner et al., 2012; Sun et al., 2012; Pishchulin et al., 2013; Yang and Ramanan, 2013; Cherian et al., 2014).

1.2.3. Deformable part-based methods

Variants of the approach have been proven to outperform single object templates in detecting humans in images. In Felzenszwalb and Huttenlocher (2005) a discriminatively trained, multiscale Deformable Parts Model (DPM) approach is introduced for pedestrian detection. The DPM model consists of a coarse root filter, a mixture of body parts filters, and part deformation relative to the root model to represent a person. The models are trained offline on a positive and negative image set using Support Vector Machines (SVM). In inference, the learned model is used for object search in a new image over a pyramid of image features, for instance, an appearance representation based on Histogram of Oriented Gradients (HOG) features (Dalal and Triggs, 2005). An object proposal is calculated from a unary data term representing the scores of each appearance filter at their respective locations and a deformation cost that depends on the position of each part with respect to the root.

Recently, approaches that use Convolutional Neural Networks (CNNs) have outperformed pictorial structures in pose estimation tasks (Chu et al., 2016; Chen and Yuille, 2014). However, CNNs require prohibitively large datasets for training, or risk overfitting a model to the data. Consequently, the approach also requires extensive computing resources and training time. Furthermore, due to its intractable nature, a CNN remains largely a 'black box' approach, which provides little insight or intuition to its performance. These limitations justify our decision to adopt the pictorial structures framework.

2. Method

2.1. Problem scope and contributions

Motivated by the limitations of existing approaches, we address in this paper the problem of estimating full-body kinematics from challenging monocular images that contain severe occlusions in unconstrained environments. We opt for a discriminative part-based approach that requires an offline learning of a model that recovers pose estimates from observable image metrics. To demonstrate the efficacy of our approach, we focus our experiments on the movement modality of cycling. Our motivation stems from the observation that this movement modality is especially challenging due to the human interaction with an object (i.e. the bicycle), which induces severe occlusions, the similarity of the posture in the frontal plane to normal human gait, and the severely occluded sagittal plane posture, for which a pose estimation method was not found in the literature. We use images captured in natural environment and a variety of resolutions. Importantly, We make no assumptions about the anthropometric proportions nor the kinematic constraints of the human model, nor the appearance of the scene. Our technique finds multiple good hypotheses for the human posture rather than just a single best solution. This is advantageous for cases where imprecision in the model may result in the desired match not being the one with the minimum energy.

2.2. Method overview

In this section we introduce our framework for the estimation of a cyclist's posture from unconstrained images. Given a monocular image with one or more cyclists, we aim to simultaneously detect and estimate the cyclists' posture characterised by the joints' spatial locations and limbs' orientations in the image. Our method learns disparate appearance and geometry models of a cyclist offline, and estimates the human posture in a new image. Specifically, our work builds on the deformable mixture of parts framework of Yang and Ramanan (2013) and Desai and Ramanan (2012), who used local part *mixtures* that capture spatial relations between parts and local appearance. We provide a diagrammatic overview of our learning and inference frameworks in Figs. 1 and 2 respectively.

2.3. Mixture of parts human model

We model the human body as a collection of the body's articulations (joints) whose spatial location is represented as a point in the 2D plane, and local appearance filters. We model the articulations as ball-and-socket joints expressed in Joint Coordinate System (JCS) following Wu et al. (2002, 2005). We express a human model as a tree-structured undirected graph $G = (V, E)$, where the vertices

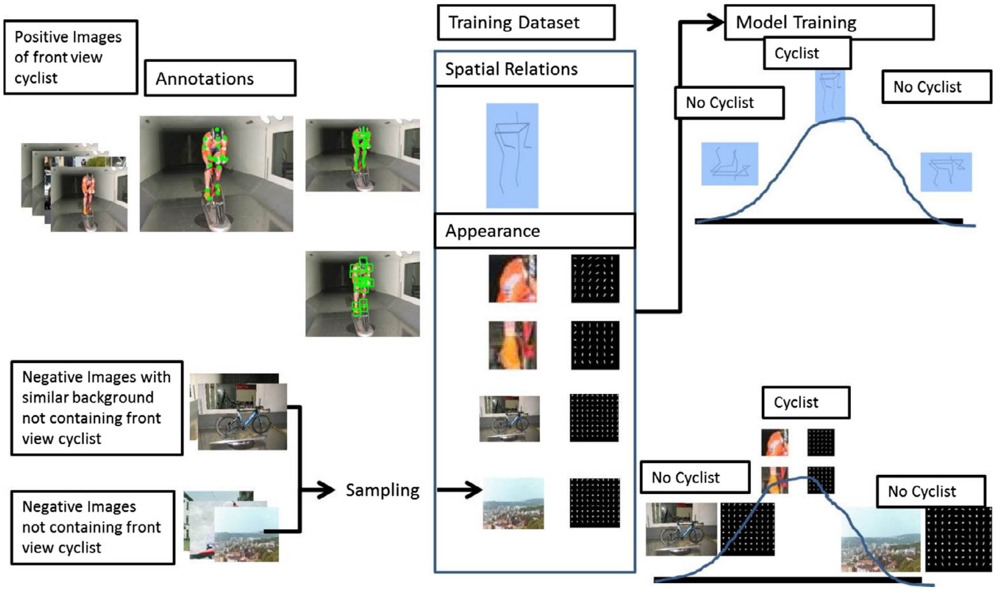
Positive Images of front view cyclist
Annotations
Training Dataset
Spatial Relations
Appearance
Model Training
Cyclist
No Cyclist
No Cyclist
Negative Images with similar background not containing front view cyclist
Negative Images not containing front view cyclist
Sampling
Cyclist
No Cyclist
No Cyclist

Fig. 1. An overall learning framework of the flexible mixture of parts approach applied to a cycling model. Appearance and spatial relations features are extracted from both positive (object is present) and negative (object is not present) images to train an object model.

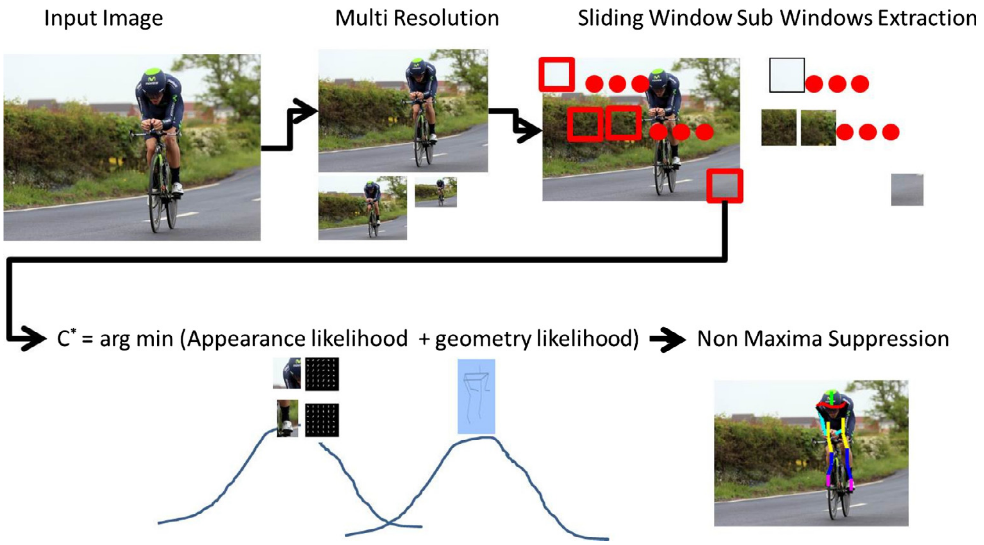
Input Image
Multi Resolution
Sliding Window Sub Windows Extraction
C* = arg min (Appearance likelihood + geometry likelihood)
Non Maxima Suppression

Fig. 2. An overall inference framework of the flexible mixture of parts approach. Sub windows are extracted from a new image at multiple resolutions using a sliding window scheme. A probabilistic object hypothesis is tested against the appearance and geometry models for each sub window resulting in a likelihood score. Finally, a non maxima suppression is applied to overlapping object proposals.

$V = \{v_1, \ldots, v_n\}$ correspond to n body joints, and an edge $(v_i, v_j) \in E$ for each pair of connected body joints v_i and v_j corresponding to the body's segments (Fig. 3). An instance of a full body in the image I is given by a configuration of body parts $\boldsymbol{L} = \{\mathbf{l}_1, \ldots, \mathbf{l}_n\} \in \mathbf{R}^{n\times 2}$, where $\mathbf{l}_i = (x_i, y_i)$ denotes the location of part v_i.

2.3.1. Appearance model

We represent the appearance of body joint v_i by a concatenated HOG (Dalal and Triggs, 2005) feature vector $\phi(I, \mathbf{l}_i) \in \mathbb{R}^{5\times5\times32}$. HOG is an edge orientation histogram based feature descriptor, which we compute on a dense grid of uniformly spaced cells over an image patch of size 32 × 32 pixels centred at $\mathbf{l}_i$. In training, we learn a full body appearance model $\boldsymbol{W} = \{\mathbf{w}_1, \ldots, \mathbf{w}_n\}$ where $\mathbf{w}_i \in \mathbb{R}^{800}$ is the template feature vector for body joint v_i (see a part visualisation in Fig. 5 and a full body visualisation in Fig. 4a).

2.3.2. Spatial relations model

To encode the spatial relations between adjacent joints, we represent the spatial relations for an edge (v_i, v_j) by a quadratic deformation vector $\psi(\mathbf{l}_i, \mathbf{l}_j) = [dx, dy, dx^2, dy^2]^T$ from the relative position of the connected joints v_i and v_j. This term is often interpreted as a negative spring energy resulting from pulling body part j from a relative position with respect to body part i (Felzenszwalb and Huttenlocher, 2005).[1] In training, we learn the spatial relations model $\mathbf{w}_{ij} \in \mathbb{R}^4$ for each edge (v_i, v_j). This parameter can be viewed as indicating the spring's position at equilibrium and rigidity. It also encodes implicit relations to distal parts through connected edges.

Thus, a score S associated with a particular configuration of body parts in an image I is a function of the parts' appearance and deformation and can be written as

$$S(I, \boldsymbol{L}) = \sum_{i=1}^{n} m_i(I, \mathbf{l}_i) + \sum_{(v_i, v_j)\in E} d_{ij}(\mathbf{l}_i, \mathbf{l}_j) \qquad (1)$$

where $m_i(I, \mathbf{l}_i) = \langle \mathbf{w}_i, \phi(I, \mathbf{l}_i)\rangle$ is a unary scalar term measuring the appearance discrepancy for each part v_i at location $\mathbf{l}_i$ in the image I with the local template $\mathbf{w}_i \in \mathbb{R}^{800}$, and is based on convolving the image I with a family of underlying linear local templates $\boldsymbol{W}$, and $\langle \cdot, \cdot \rangle$ is the inner product operator. Similarly,

[1] Appendix A provides further interpretation and generalisation of the parts deformation cost.

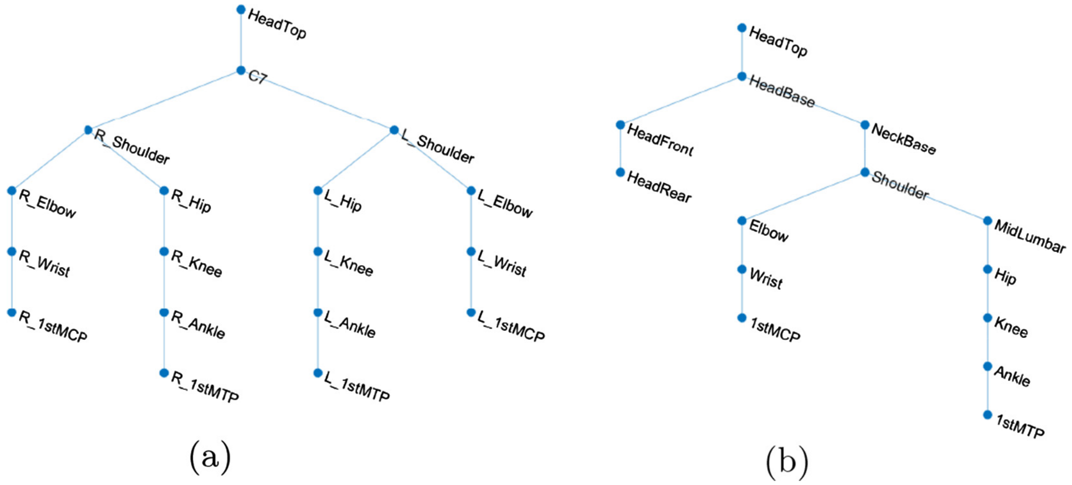

Fig. 3. Tree structured graph models of a cyclist in frontal (a) and sagittal (b) views. Note that for the frontal view the model is an acyclical approximation of a natural representation that links the two hip nodes with a pelvic edge. The approximation simplifies the graph model and enables exact inference.

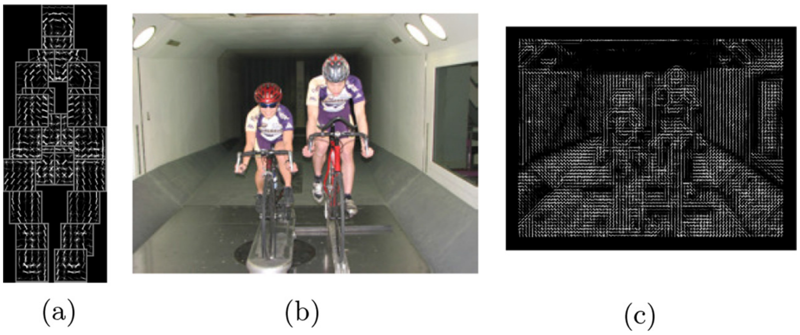

Fig. 4. (a) A visualisation of the learnt full body frontal cyclist model, where parts' appearance models characterised by HOG (Dalal and Triggs, 2005) filters and their relative spatial location relative to a root filter. Since our model also learns co-occurrence probabilities for part mixture components, the displayed model shows one example of such co-occurrence. In inference, a feature representation of a new image (b) is computed at multiple image resolutions. (c) shows an example feature representation of the image at one resolution level. The feature vector is then convolved with each of the learnt appearance model parts' filters (a) to yield a local response.

$d_{ij}(\mathbf{l}_i, \mathbf{l}_j) = \langle \mathbf{w}_{ij}, \psi(\mathbf{l}_i, \mathbf{l}_j) \rangle$ is a pairwise scalar term measuring the deformation cost for a given pair of connected parts, that is of part v_i at location $\mathbf{l}_i$ and part v_j at location $\mathbf{l}_j$. A low negative d_{ij} score indicates that a body part's location and orientation with respect to its parent (proximal body segment) is close to the learnt prior spatial relations model. For clarity, the terms are summarised as follows:

- $\mathbf{w}_i \in \mathbb{R}^{800}$ is the learnt HOG feature appearance template for joint v_i
- $\phi(I, \mathbf{l}_i) \in \mathbb{R}^{800}$ is the concatenated HOG feature descriptor of a 32×32 pixels sized patch of image I centred at $\mathbf{l}_i$
- $m_i(I, \mathbf{l}_i) = \langle \mathbf{w}_i, \phi(I, \mathbf{l}_i) \rangle$ is a scalar measuring how well the feature at the image patch centred at $\mathbf{l}_i$ matches the template of v_i
- $\mathbf{w}_{ij} \in \mathbb{R}^4$ is the learnt spatial deformation model of joint v_j with respect to its parent v_i
- $\psi(\mathbf{l}_i, \mathbf{l}_j) \in \mathbb{R}^4$ is the spatial deformation between two points $\mathbf{l}_i$ and $\mathbf{l}_j$ in the image I
- $d_{ij}(\mathbf{l}_i, \mathbf{l}_j) = \langle \mathbf{w}_{ij}, \psi(\mathbf{l}_i, \mathbf{l}_j) \rangle$ is a scalar measuring how well the deformation between two points in the image match the learnt deformation model for joints v_u and v_j

2.3.3. Mixture of parts

Notwithstanding the activity-specific application, the appearance of body parts is highly variable. For instance, an appearance patch for the first metatarsophalangeal (1stMTP) joint looks different at the top-dead-centre of the cycling stroke than at the bottom-dead-centre. Likewise, the helmet and the hand position varies between a road cyclist, a sprint track cyclist and a mountain bike rider. Therefore, to encode a richer family of appearances for each body part, we model the appearance of each body joint v_i by a mixture of templates, instead of a single fixed appearance template. We write $t^i \in \{1, \ldots, T\}$ for a latent variable denoting an appearance mixture component for part v_i, and model the appearance of each body *joint* v_i by a mixture $\mathbf{\Phi}_i = \{\boldsymbol{\phi}_{i1}, \ldots, \boldsymbol{\phi}_{iT}\} \in \mathbb{R}^{800 \times T}$, where $\phi_{it}(I, \mathbf{l}_i)$ is a feature vector component t^i centred at $\mathbf{l}_i$ following Desai and Ramanan (2012). An assignment of mixtures for a full body model can then be denoted by $\mathbf{t} = \{t_1, \ldots, t_n\}$. In training, we learn a unary term $b_i(t^i)$ that supports a particular mixture component assignment for the body joint v_i (see a visualisation in Fig. 5).

2.3.4. Co-occurrence model

Our intuition is that support for a part's particular mixture component assignment depends somewhat on the full-body pose. That is, two joints connected by a rigid limb are likely to present consistent pairing of appearance representations based on the limb's global orientation, for example the elbow and shoulder joints connected by the upper arm when elevated versus externally rotated. To capture the dependency of global pose on local appearance variations, we also learn a pairwise term $b_{ij}(t^i, t^j)$ that supports a particular mixture components co-assignment for the body joints v_i and v_j.

We write $\mathbf{z}_i = (\mathbf{l}_i, t^i) \in \mathbb{R}^3$ for the pixel location and mixture component for part v_i. We can then write the full score S associated with a particular configuration of body parts in an image I as

$$S(I, \mathbf{z}) = \sum_{i=1}^{n} m_i(I, \mathbf{z}_i) + \sum_{(v_i, v_j) \in E} d_{ij}(\mathbf{z}_i, \mathbf{z}_j) \tag{2}$$

where

$$\boldsymbol{M}_i(I, \mathbf{z}_i) = \langle \mathbf{w}_i(t^i), \phi(I, \mathbf{l}_i) \rangle + b_i(t^i)$$

$$d_{ij}(\mathbf{z}_i, \mathbf{z}_j) = \langle \mathbf{w}_{ij}(t^i, t^j), \psi(\mathbf{l}_i, \mathbf{l}_j) \rangle + b_{ij}(t^i, t^j)$$

The pairwise term $b_{ij}(t^i, t^j)$ favours consistent co-occurence of mixture components for the corresponding parts v_i and v_j, such that a positive $b_{ij}(t^i, t^j)$ score reflects consistent pose assignments, and a negative score reflects the alternative. Thus, an instance of a human in the image I indicates which mixture component from each body joint is used and its relative location.

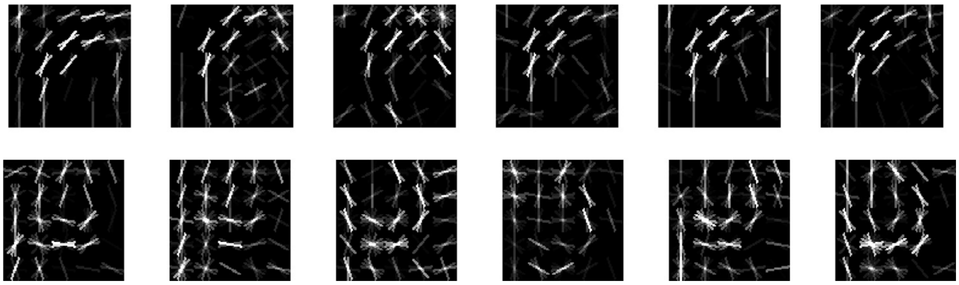

Fig. 5. Each joint is modelled by 6 possible templates learnt during the training phase and represent alternative HOG filter appearance representations of an image patch of size 32 × 32 pixels centred at the joint. The filters capture variance in part appearance due to small changes in view point, background, shape, illumination and texture (e.g. changes induced by the pedalling cycle). Above is a visualisation of the filter mixtures for the right shoulder (top) and left 1stMTP joint (bottom) in the frontal view. Predictably, in cycling the appearance variability at the foot is higher than at the shoulder. A particular full body model has one mixture component for each joint, for instance the second component for the shoulder and the fifth component for the 1stMTP.

2.4. Inference

Our goal is to detect and estimate the posture of a cyclist from test images. In inference, we produce candidate pose proposals by using a sliding window detection scheme over an image pyramid. The optimal match of a model to an image is found by maximising (2) over $\mathbf{z}$

$$C(z) = \max_{z} \left(\sum_{i=1}^{n} m_i(\mathbf{z}_i) + \sum_{(v_i,v_j)\in E} d_{ij}(\mathbf{z}_i,\mathbf{z}_j) \right). \tag{3}$$

Conveniently, the tree graph structure leads to an efficient and tractable inference such as sampling or belief propagation. We use our models to compute the score for each part v_i, at every pixel location of image I, and for all appearance mixture components $\mathbf{t}^i$, which includes messages from the children nodes of v_i by

$$s_i(\mathbf{z}_i) = m_i(I,\mathbf{z}_i) + r_i \tag{4}$$

where r_i is the sum of messages passed by the children of v_i. We provide a score heatmap visualisation for a representative part v_i for all its mixture components $\mathbf{t}^i$ in Fig. 6. A message from a child part to its parent computes the best location and mixture component for the child part.

Upon arrival of all messages at the root node, its score represents the optimal pose at its location. Retaining the indices of the best scoring part proposal, it is then possible to track back to find the location and mixture of each body part that is optimal for the pose. The principle that underpins this approach is that a collection of weak classifiers, collectively creates a strong class classifier. We retain the q best-scoring candidate cyclist poses using a threshold and apply non maxima suppression to prune overlapping proposals and, for each, select the one with top cyclist score. This enable the retention of multiple instances of cyclists in the image (Fig. 7).

2.5. Learning

We adopt the supervised learning paradigm of Kumar et al. (2009) and train a part based detector for the human (cyclist) using manually annotated positive examples of joint locations and negative examples. We separately trained classifiers for the frontal and saggital view on 141 and 144 sagittal images of cyclists, respectively, whose pose was manually annotated for each of our models' body joint keypoints (see Fig. 3). Our negative set contains 1217 images of people but not cyclists, and background scenery images.

The learning problem can be cast as obtaining a weight vector $\boldsymbol{\gamma} = (\mathbf{w}_i, \mathbf{w}_{i,j}, b_i, b_{i,j})$ and scalar bias ξ, such that the learnt model parameters are able to discriminate between positive and negative examples in terms of their energy value. Learning the most discriminative model parameters is equivalent to solving the optimisation problem

$$(\gamma^\star, \xi^\star) = \arg\min_{\gamma,\xi\geqslant 0} \frac{1}{2}\|\gamma\|^2 + C\left(\sum_n \xi_n\right) \tag{5}$$

such that

$$\gamma \cdot \phi(I_n, \mathbf{z}_n) \geqslant 1 - \xi_n \forall n \in \text{ positive images},$$

$$\gamma \cdot \phi(I_n, \mathbf{z}) \leqslant -1 + \xi_n \forall \mathbf{z}, \forall n \in \text{ negative images}.$$

where the $\|\cdot\|$ operator represents the standard l_2-norm, n the number of images, and $C \geqslant 0$ is a constant, which specifies the trade-off between accuracy and regularisation of the weights vector. The constraints ensure that positive samples score better than 1, and the negative samples score less than −1, violations of which are penalised by the objective function using the slack variables ξ_n.

Whilst convex, this learning problem cannot be solved efficiently due to the number of constraints. Kumar et al. (2009) have reduced this problem to an equivalent problem with a polynomial number of constraints, from which an optimal solution can be reached. Known as a structural SVM, this learning problem has many efficient solvers. We use the dual-coordinate-descent quadratic programming solver of Yang and Ramanan (2013) (see visualisation in Fig. 4a).

2.5.1. Part mixture components determination

We assume that the appearance of a body joint depends on its position relative to its parent (proximal body segment) in E. Therefore, we can use this relative position as criterion to cluster the part's appearance instances in our training data into consistent relative orientation. We define the mixture label for a part based on cluster membership achieved via K-means with $K = T$.

2.6. Implementation

We represent the human pose using a tree-structured graph with 26 nodes comprised of 18 and 14 keypoint nodes representing joints and 8 and 12 secondary mid-limb nodes for the frontal and sagittal models respectively, with the base of the neck at joint $C7$ as its root and the limbs and head as its extremities (Fig. 1). We justify supplementing the number of joint keypoints by secondary mid-limb nodes by the experiments of Yang and Ramanan (2013), who showed that 26 nodes provide a good trade-off of performance vs. computation. In the sagittal case, we consider the base, top, front and rear edges of the bicycle helmet as key points to enable helmet orientation characterisation relative to the cyclist's trunk for future analysis.

A sagittal view of a cyclist presents severe occlusion of most or all of the limbs on the far side. Thus, it presents a large and significant view change that the framework of Yang and Ramanan (2013), who impose a fixed number of keypoint nodes to model *small* changes in foreshortening, is unable to handle. Instead, in agreement with Felzenszwalb and Huttenlocher (2005) we handle the severe occlusions induced by these rotations by explicitly encoding out-of-plane rotations by using a separate model with a different number of keypoint nodes for each view.

3. Results

To evaluate the performance of our pose estimation method, we apply our method to a set of challenging test images comprised of frontal and sagittal views of human cycling in unconstrained environment. In this section we report on both qualitative and quantitative results.

3.1. Quantitative results

We conducted experiments on new task-specific datasets containing 141 frontal and 144 sagittal images of cyclists, whose pose was manually annotated for each of our models' body joint keypoints. The datasets contain images that have been either taken by the authors, downloaded from on-line repositories with a licence search criteria set to creative common (Flickr), labelled for reuse (Google Images), or provided by the University of Washington's windtunnel.

To train our models, we have split the datasets into a standard 60%, 20% and 20% for model learning, cross-validation and test sets respectively. Our negative set contains 1217 images comprised of

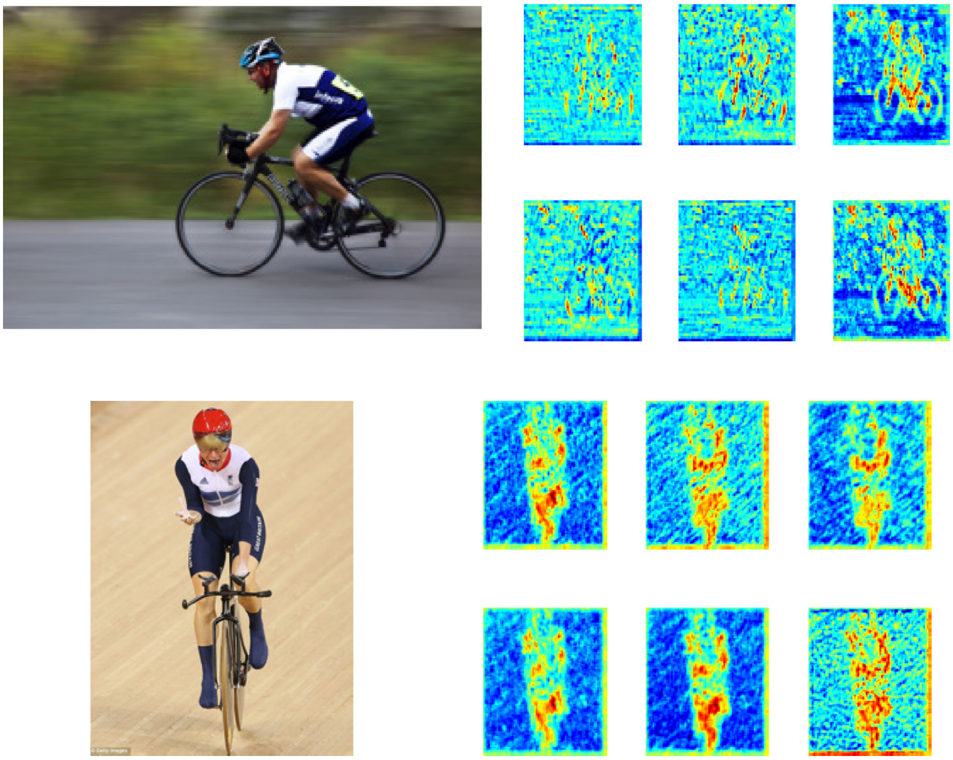

Fig. 6. In inference, a feature representation of a new image is convolved with the learnt model's mixture parts appearance filters. The computation yields a local response as part appearance scores represented as heatmaps for the ankle body part for each mixture component (example from one indicative pyramid level). A hot colour represents high score/low discrepancy signal with respect to the body part appearance mixture component model. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

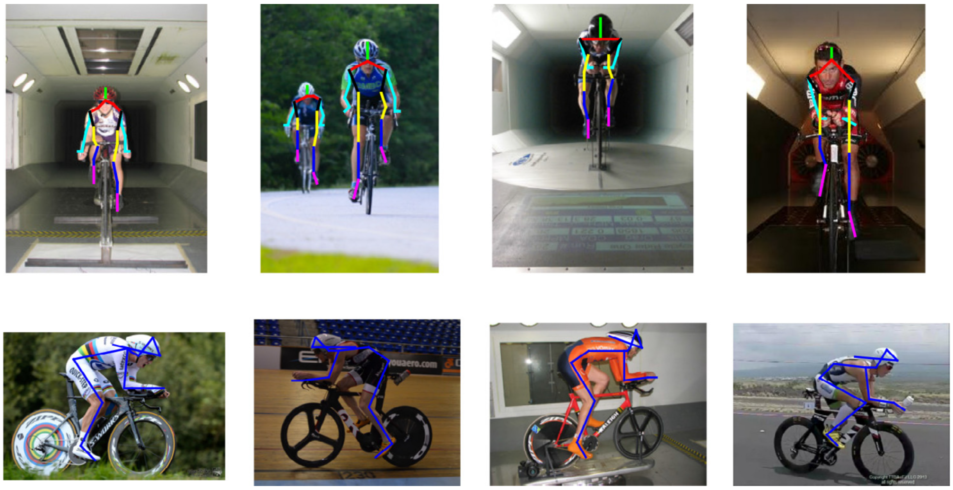

Fig. 7. Our method yields pose estimation from a monocular image of a cyclist in the frontal and sagittal views and is robust to changes in scene, illumination and can yield multiple instances in an image.

the positive and negative training images from the INRIA Person (Dalal and Triggs, 2005) and Parse (Ramanan, 2007) datasets. In both datasets, the positive training images contain images of people with images of cyclists removed, whilst the negative sets contain mostly background scenery images. Using images that contain people in our negative set ensures that our cyclist model discriminates well between people in normal gait and cyclists for our specific task.

We measure the pose prediction of our method using loss functions, the Probability of Correct Keypoint (PCK) and Average Precision of Keypoints (APK) (Yang and Ramanan, 2013). A prediction is considered correct if it resides within a small distance from the annotated 'ground truth' point. For a given part at the annotated location i_*, the loss for prediction $\hat{i}$ is defined by

$$\triangle^p(i_*,\hat{i}) = I\left(\|i_* - \hat{i}\| > \alpha \max(h,w)\right), \tag{6}$$

where I is the indicator function, and h and w are the vertical and horizontal distances respectively, and α is a detection region threshold parameter. The results are presented in Table 1.

Table 1
Probability of Correct Keypoint (PCK) and Average Precision of Keypoints (APK) results.

Dataset	Mean PCK	Mean APK
Frontal view	91.1	88.7
Sagittal view	69.3	66.1

3.2. Qualitative results

In contrast to the test images datasets, which contain images of a single cyclist with no other human objects in the scene, here we aim to qualitatively investigate the performance of our methods on unannotated images that contain multiple cyclists, severe occlusions and additional objects and humans in the scene (see results in Fig. 8). Whilst our approach is robust to small changes in orientation, view point and scale of the human object in images, the pictorial structure framework assumes that all body parts have a fixed scale. Hence, it may suffer local failure in such images that present cases where certain body parts experience severe foreshortening effect due to change in view point, where the problem becomes ill-posed. Fig. 9 presents such cases.

4. Discussion

The presented results demonstrate the utility of estimating human kinematics via a fully supervised learning approach to reconstruction of human posture from unconstrained images as an alternative to marker-based motion capture. Our method is underpinned by a spatial relations model and parts' appearance and co-occurrence models. Therefore, like other markerless approaches it does not suffer from the deficiencies of marker-based system. For instance, our method is not constrained to an expensive laboratory setting with controlled lighting, reflectance and other environmental conditions, nor does it have any direct sensitivity to STA. likewise, it does not suffer the inter-trial and inter-tester repeatability and reliability of direct digitisation approaches (Krosshaug and Bahr, 2005; Sanders et al., 2016).

For our task, we applied the flexible mixture of parts method (Yang and Ramanan, 2013) to the estimation of cyclists' posture and achieved robust results with a task-specific trained classifier. It is, however, impossible to design a discriminative classifier for the general case because of the high variability in the appearance of human ambulation. The performance of the approach critically relies on time consuming and costly process and the availability of a large quantity of training samples. Further, the generalisation of the approach can only be achieved through the addition of adequate training samples, as its adaptability to unseen body postures is low, as typically manifested by poor performance were occlusions exist. Thus, to adapt our method to a different movement modality, a task-specific classifier needs to be trained on an appropriate training set.

With respect to our specific cycling task, the coupling of the human and object reduces the search space of likely poses represented by the spatial relations between parts in our model as a result, for instance, of the predictable position of the hands on the handlebars and feet on the pedals. This may be viewed as an advantage of our task of choice when compared to unconstrained activity such as running. On the other hand, the human-object interaction, in fact, complicates the search for a body part's appearance. This is

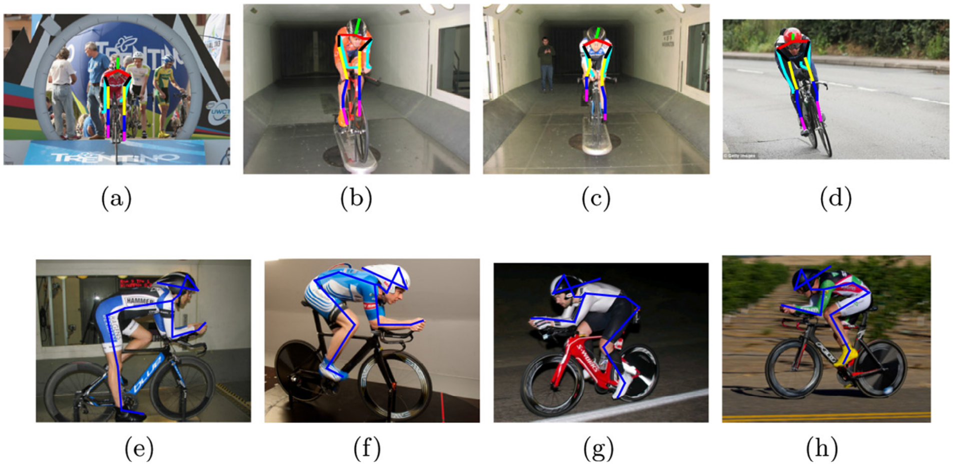

Fig. 8. Qualitative results in the frontal (top) and sagittal (bottom) planes. The discriminative power of the classifier is demonstrated by rejection of cyclist present in the scene but not in a riding position in (a), presence of people in (a) and (c), and robustness to small changes in object's view (b, f, g) and orientation (d, e).

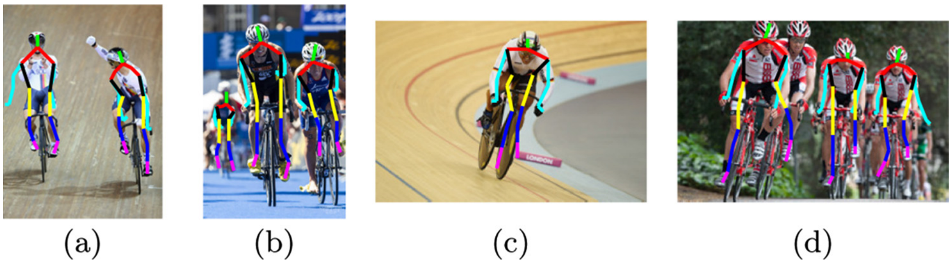

Fig. 9. Examples of local failures; (a) shows local failure at the wrists due to part unlearned spatial location. (b) shows failure at the ankle of the cyclist on the right due to unlearnt part appearance (barefoot). (c) shows failure at the legs due to hip occlusion. (d) shows failure at the left ankle of the lead cyclist due to part appearance similarity of the second cyclist's ankle.

due to increased incidence of body part occlusions and appearance similarity or ambiguity. Therefore, a trade-off exists between improved spatial relations and less discriminative appearance. Since the apriori location of the human is unknown in a new image, the search space for parts' appearance is an image pyramid over the entire input image. Hence, overall it is less challenging to estimate the pose in running than when a human-object interaction exist. Normal gait and running have been addressed often previously (Eichner et al., 2009, 2012; Andriluka et al., 2009, 2014, and others). Typically, these approaches are aided by segmentation and background modelling algorithms that separate the human (foreground) and background scene prior to pose estimation by reducing the image search space to the foreground alone. We avoid pre processing steps, and directly estimate the pose from the parts' appearance and spatial relations simultaneously.

This is consistent with our aforestated motivation, which stems from the observation that the cycling is particularly challenging due to the human interaction with the bicycle, which induces severe occlusions, the similarity of the posture in the frontal plane to normal human gait, which presents a discrimination challenge, and the severely occluded sagittal plane posture, for which a pose estimation method was not found in the literature. In follow-up work that is outside the scope of this paper, we will investigate whether modelling the human-object interaction explicitly improves the pose estimation.

A principal limitation of the pictorial structures framework is that exact inference is only possible with tree structured graph model. However, in certain situations, such as temporal and occlusion models, it is advantageous to have non-tree models. Kiefel and Gehler (2014) used a binary random variable to model occlusions at every possible location, scale and orientation and graph topology that is not restricted to a tree structure, with modest improved performance. Cherian et al. (2014) generated pose candidates in each time frame by enforcing temporal constancy between instances of a tree model. They decomposed the body parts to generate temporally smoothed body part sequences, followed by recomposition of the body pose. A non-tree extension to our work that models occlusions and allows for temporal constancy to be enforced would make the approach very attractive for a variety of applications.

Despite significant progress, accurate inference remains an extremely challenging problem principally due to occlusions and self-occlusions in the image. Consequently, in our framework, we model the human body only for the visible body parts for each view. This results in separate models for the frontal and sagittal views. The challenge is further compounded by inter-object interaction, such as the interaction between a cyclist and the bicycle. Therefore, a natural extension to our approach would exploit advances in modelling the human-object interaction within the framework. Moreover, in this work we avoided imposing explicit kinematic constraints on the pose proposals. Introducing such constraints will significantly reduce the search space of probable poses and improve the accuracy.

We note that whilst manual annotation of point location remains the standard ground truth for performance evaluation of pose estimation techniques in the computer vision domain, a higher level of accuracy is often expected in the biomechanics domain. The reason stems from error propagation with subsequent calculations and double differentiation required for inverse dynamics optimisation. The accuracy required is commonly achieved through the obtrusive use of surface mounted reflective markers. Nevertheless, our choice to validate our model estimation against manual annotation is justified since the use of reflective markers would contaminate the model in learning and the image data in inference. It will result in appearance models that are tuned for the presence of a marker in the image patch. Consequently, performance evaluation would be grossly overestimated. Furthermore, marker-based systems require control of scene, lighting and environmental conditions, which would unfairly penalise our method.

The estimation of pose using our method can be used for the reconstruction of the human body's geometry. Using our cycling task as an example, the geometric shape of the cyclist can be used to enhance the understanding of the relationship between a cyclist's posture and aerodynamic drag. The task then becomes a problem of extracting the cyclist's shape from the background in an image in a segmentation pipeline. Thus, the extracted skeletal pose can be used as a necessary prior foreground seed for segmentation techniques such as graph cuts (Boykov et al., 2001; Veksler, 2008), and the exterior to a convex hull that contains all part patches as its background seed.

5. Conclusion

In this paper, we investigated the challenging problem of markerless estimation of a human full body kinematics from monocular images. We proposed a discriminative part-based approach that develops a probabilistic prior model based on learned measurements. In learning, a structured SVM solver learns spatial relations of skeletal segment orientation and co-occurrence relation between parts appearance. In inference, the model detects the human in the image and recovers pose estimates. We applied our approach to images of cyclists captured in natural environment with no assumptions in relation to kinematic constraints, nor the appearance of the scene. Our technique finds multiple good hypotheses for the human pose rather than just a single best solution. This is advantageous for cases where imprecision in the model resulting in the desired match not being the one with the minimum energy. Our method yields a robust user-interaction-free approach for estimation of full body kinematics, which serves as a crucial evidence base pre-requisite to motion analysis. Based on skeletal kinematics, joint forces, torques, power and efficiency of motion can then be determined. Furthermore, our pose estimate can be viewed as a necessary first step in a segmentation pipeline aimed at characterising the geometry of human motion.

Conflict of interest statement

None.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. HL's research is partly funded by an ARC Discovery grant DP120103896.

Appendix A. Part Spatial Relation and Deformation Cost

In Section 2.3, we represent the spatial relations between adjacent body parts (joints) by a quadratic deformation vector $\psi(\mathbf{l}_i, \mathbf{l}_j) = [dx, dy, dx^2, dy^2]^T$ from the relative position of the connected joints v_i at $\mathbf{l}_i$ and v_j at $\mathbf{l}_j$, such that $dx = x_i - x_j$ and $dy = y_i - y_j$, consistent with Yang and Ramanan (2013) and others. We mention that this term can be interpreted as a negative spring energy resulting from pulling joint j from a relative position with respect to joint i. For clarity, in this section we provide context and alternate interpretations of this force. In order to avoid notational clutter, we omit the subscript ij for the remainder of this section, where the interpretation is obvious.

Consider a multivariate Gaussian distribution over the set of *relative* positions of the adjacent joints (v_i, v_j) in our training data, where $\boldsymbol{\mu} = (\mu_x, \mu_y)$ is the mean of the distribution. We wish to impose a penalty over a relative joint location hypothesis that deviates from the mean of the distribution. Consequently, a body part location hypothesis would favour a relative location proposal that is in agreement with the learnt prior spatial relations model. Thus, we define a spring force that represents the distance of the test point from the distribution's centre of mass, which we define by

$$d_{ij}(\mathbf{l}_i, \mathbf{l}_j) = A(dx - \mu_x)^2 + B(dy - \mu_y)^2 \tag{7}$$

where, A and B are arbitrary real-valued constants. This can be expanded to

$$d_{ij}(\mathbf{l}_i, \mathbf{l}_j) = Adx^2 + Bdy^2 - 2A\mu_x dx - 2B\mu_y dy + (A\mu_x^2 + B\mu_y^2). \tag{8}$$

The last term does not depend on $\mathbf{l}_i$ or $\mathbf{l}_j$ and can be replaced with a constant F. After rearrangement the force can be written as the inner product of two vectors plus a constant F

$$\begin{aligned} d_{ij}(\mathbf{l}_i, \mathbf{l}_j) &= \langle (dx, dy, dx^2, dy^2), (-2A\mu_x, -2B\mu_y, A, B) \rangle + F \\ &= \langle (dx, dy, dx^2, dy^2), w_{ij} \rangle \rangle, \end{aligned} \tag{9}$$

where only the first vector depends on the proposed body part locations $\mathbf{l}_i$ or $\mathbf{l}_j$ and therefore corresponds to our deformation vector $\psi(\mathbf{l}_i, \mathbf{l}_j)$, and w_{ij} is the vector of weight coefficients to be learnt. Eq. (8) can also be viewed as a special case of the general canonical form of an arbitrarily oriented ellipse

$$Adx^2 + Bdy^2 + Cdxdy + Ddx + Edy + F = 0,$$

where $C = 0$. Therefore, a generalisation of our approach may represent the deformation cost by $\psi^{\star}(\mathbf{l}_i, \mathbf{l}_j) = [dx, dy, dxy, dx^2, dy^2]^T \in \mathbb{R}^5$ with the addition of a cross dimension term dxy.

More generally, an arbitrarily oriented ellipsoid in $\mathbb{R}^n$, centred at $\boldsymbol{\mu}$ satisfies

$$(\mathbf{l} - \boldsymbol{\mu})^T \boldsymbol{S} (\mathbf{l} - \boldsymbol{\mu}) = 1,$$

where $\boldsymbol{S}$ is a positive definite matrix and $\mathbf{l}$ and $\boldsymbol{\mu}$ are vectors. After substitution we get

$$\begin{aligned} &[\,dx \quad dy \quad 1\,] \begin{bmatrix} S & -S\mu \\ -\mu^T S & -\mu^T S\mu \end{bmatrix} \begin{bmatrix} dx \\ dy \\ 1 \end{bmatrix} = 0 \\ &[\,dx \quad dy \quad 1\,] \begin{bmatrix} A & C/2 & D/2 \\ C/2 & B & E/2 \\ D/2 & E/2 & 1 \end{bmatrix} \begin{bmatrix} dx \\ dy \\ 1 \end{bmatrix} = 0, \end{aligned} \tag{10}$$

where the five upper triangular elements of the symmetric matrix here correspond to the five elements of (A, B, C, D, E).

For a probability distribution we can define a dissimilarity measure between two random vectors $\mathbf{l}$ and $\boldsymbol{\mu}$ of the same distribution as

$$\boldsymbol{D}(\mathbf{l}, \boldsymbol{\mu}) = \sqrt{(\mathbf{l} - \boldsymbol{\mu})\boldsymbol{S}^{-1}((\mathbf{l} - \boldsymbol{\mu})},$$

where $\boldsymbol{S}$ is the covariance matrix. In our application, this distance, the Mahalanobis distance, represents the distance of the test point $\mathbf{l}$ from the distribution's centre of mass $\boldsymbol{\mu}$. The eigenvectors of $\boldsymbol{S}$ define the principal axes of the ellipsoid and the eigenvalues of $\boldsymbol{S}$ are the reciprocals of the squares of the semi-axes. This representation also provides an elegant link to Principal Component Analysis (PCA). Indeed, this is the form of the deformation cost that was used in early pictorial structures work (e.g. Felzenszwalb and Huttenlocher, 2005).

References

Agarwal, A., Triggs, B., 2006. Recovering 3d human pose from monocular images. IEEE Trans. Pattern Anal. Mach. Intell. 28 (1), 44–58.

Andersen, M.S., Benoit, D.L., Damsgaard, M., Ramsey, D.K., Rasmussen, J., 2010. Do kinematic models reduce the effects of soft tissue artefacts in skin marker-based motion analysis? An in vivo study of knee kinematics. J. Biomech. 43 (2), 268–273.

Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B., 2014. 2d human pose estimation: new benchmark and state of the art analysis. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 3686–3693.

Andriluka, M., Roth, S., Schiele, B., 2009. Pictorial structures revisited: People detection and articulated pose estimation. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009).

Baak, A., Müller, M., Bharaj, G., Seidel, H.-P., Theobalt, C., 2013. A data-driven approach for real-time full body pose reconstruction from a depth camera. In: Consumer Depth Cameras for Computer Vision. Springer, pp. 71–98.

Boykov, Y., Veksler, O., Zabih, R., 2001. Fast approximate energy minimization via graph cuts. IEEE Trans. Pattern Anal. Mach. Intell. 23 (11), 1222–1239.

Camomilla, V., Bonci, T., Dumas, R., Cheze, L., Cappozzo, A., 2015. A model of the soft tissue artefact rigid component. J. Biomech. 48 (10), 1752–1759.

Camomilla, V., Donati, M., Stagni, R., Cappozzo, A., 2009. Non-invasive assessment of superficial soft tissue local displacements during movement: a feasibility study. J. Biomech. 42 (7), 931–937.

Cappozzo, A., Della Croce, U., Leardini, A., Chiari, L., 2005. Human movement analysis using stereophotogrammetry: part 1: theoretical background. Gait & Posture 21 (2), 186–196.

Ceseracciu, E., Sawacha, Z., Fantozzi, S., Cortesi, M., Gatta, G., Corazza, S., Cobelli, C., 2011. Markerless analysis of front crawl swimming. J. Biomech. 44 (12), 2236–2242.

Challis, J.H., 1995. A procedure for determining rigid body transformation parameters. J. Biomech. 28 (6), 733–737.

Chen, X., Yuille, A., 2014. Articulated pose estimation by a graphical model with image dependent pairwise relations. In: Advances in Neural Information Processing Systems (NIPS).

Cherian, A., Mairal, J., Alahari, K., Schmid, C., 2014. Mixing body-part sequences for human pose estimation. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition.

Chu, X., Ouyang, W., Li, H., Wang, X., 2016. Structured feature learning for pose estimation. In: CVPR.

Corazza, S., Andriacchi, T.P., 2009. Posturographic analysis through markerless motion capture without ground reaction forces measurement. J. Biomech. 42 (3), 370–374.

Corazza, S., Mündermann, L., Andriacchi, T., 2007. A framework for the functional identification of joint centers using markerless motion capture, validation for the hip joint. J. Biomech. 40 (15), 3510–3515.

Corazza, S., Mündermann, L., Chaudhari, A.M., Demattio, T., Cobelli, C., Andriacchi, T. P., 2006. A markerless motion capture system to study musculoskeletal biomechanics: visual hull and simulated annealing approach. Ann. Biomed. Eng. 34 (6), 1019–1029.

Cutti, A.G., Paolini, G., Troncossi, M., Cappello, A., Davalli, A., 2005. Soft tissue artefact assessment in humeral axial rotation. Gait Posture 21 (3), 341–349.

Dalal, N., Triggs, B., June 2005. Histograms of oriented gradients for human detection. In: Schmid, C., Soatto, S., Tomasi, C. (Eds.), International Conference on Computer Vision & Pattern Recognition. vol. 2. INRIA Rhône-Alpes, ZIRST-655, av. de l'Europe, Montbonnot-38334, pp. 886–893.

Della Croce, U., Leardini, A., Chiari, L., Cappozzo, A., 2005. Human movement analysis using stereophotogrammetry: part 4: assessment of anatomical landmark misplacement and its effects on joint kinematics. Gait Posture 21 (2), 226–237.

Desai, C., Ramanan, D., 2012. Detecting actions, poses, and objects with relational phraselets. In: ECCV (4). pp. 158–172.

Ehrig, R.M., Taylor, W.R., Duda, G.N., Heller, M.O., 2006. A survey of formal methods for determining the centre of rotation of ball joints. J. Biomech. 39 (15), 2798–2809.

Eichner, M., Ferrari, V., Zurich, S., 2009. Better appearance models for pictorial structures. In: BMVC. vol. 2. p. 5.

Eichner, M., Marin-Jimenez, M., Zisserman, A., Ferrari, V., 2012. 2d articulated human pose estimation and retrieval in (almost) unconstrained still images. Int. J. Comput. Vision 99, 190–214.

Fastovets, M., Guillemaut, J.-Y., Hilton, A., June 2013. Athlete pose estimation from monocular tv sports footage. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 1048–1054.

Felzenszwalb, P.F., Huttenlocher, D.P., 2005. Pictorial structures for object recognition. Int. J. Comput. Vision 61 (1), 55–79.

Grimpampi, E., Camomilla, V., Cereatti, A., de Leva, P., Cappozzo, A., 2014. Metrics for describing soft-tissue artefact and its effect on pose, size, and shape of marker clusters. IEEE Trans. Biomed. Eng. 61 (2), 362–367.

Gupta, A., Chen, T.P., Chen, F., Kimber, D., Davis, L.S., 2008. Context and observation driven latent variable model for human pose estimation. In: 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR 2008, 24–26 June 2008. Anchorage, Alaska, USA.

Hatze, H., 2002. The fundamental problem of myoskeletal inverse dynamics and its implications. J. Biomech. 35 (1), 109–115.

Kiefel, M., Gehler, P., 2014. Human pose estimation with fields of parts. In: Computer Vision – ECCV 2014. Lecture Notes in Computer Science, vol. LNCS 8693. Springer International Publishing, pp. 331–346.

Krosshaug, T., Bahr, R., 2005. A model-based image-matching technique for three-dimensional reconstruction of human motion from uncalibrated video sequences. J. Biomech. 38 (4), 919–929.

Kumar, M.P., Zisserman, A., Torr, P.H., 2009. Efficient discriminative learning of parts-based models. In: 2009 IEEE 12th International Conference on Computer Vision. IEEE, pp. 552–559.

Leardini, A., Chiari, L., Della Croce, U., Cappozzo, A., 2005. Human movement analysis using stereophotogrammetry: part 3. Soft tissue artifact assessment and compensation. Gait Posture 21 (2), 212–225.

Lerasle, F., Rives, G., Dhome, M., Garcier, J., Van Praagh, E., 1997. Leg cycling tracking by dynamic vision. J. Biomech. 30 (8), 837–840.

Li, K., Zheng, L., Tashman, S., Zhang, X., 2012. The inaccuracy of surface-measured model-derived tibiofemoral kinematics. J. Biomech. 45 (15), 2719–2723.

Lu, T.-W., O'connor, J., 1999. Bone position estimation from skin marker coordinates using global optimisation with joint constraints. J. Biomech. 32 (2), 129–134.

Magalhaes, F.A., Sawacha, Z., Di Michele, R., Cortesi, M., Gatta, G., Fantozzi, S., 2013. Effectiveness of an automatic tracking software in underwater motion analysis. J. Sports Sci. Med. 12 (4), 660.

Miranda, D.L., Rainbow, M.J., Crisco, J.J., Fleming, B.C., 2013. Kinematic differences between optical motion capture and biplanar videoradiography during a jump–cut maneuver. J. Biomech. 46 (3), 567–573.

Peters, A., Galna, B., Sangeux, M., Morris, M., Baker, R., 2010. Quantification of soft tissue artifact in lower limb human motion analysis: a systematic review. Gait Posture 31 (1), 1–8.

Pishchulin, L., Andriluka, M., Gehler, P., Schiele, B., 2013. Poselet conditioned pictorial structures. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Ramanan, D., 2007. Learning to parse images of articulated bodies. Adv. Neural Inf. Process. Syst. 19, 1129.

Riemer, R., Hsiao-Wecksler, E.T., Zhang, X., 2008. Uncertainties in inverse dynamics solutions: a comprehensive analysis and an application to gait. Gait Posture 27 (4), 578–588.

Rosario, H., Page, A., Besa, A., Mata, V., Conejero, E., 2012. Kinematic description of soft tissue artifacts: quantifying rigid versus deformation components and their relation with bone motion. Med. Biol. Eng. Comput. 50 (11), 1173–1181.

Salzmann, M., Pilet, J., Ilic, S., Fua, P., 2007. Surface deformation models for nonrigid 3d shape recovery. IEEE Trans. Pattern Anal. Mach. Intell. 29 (8), 1481–1487.

Sandau, M., Koblauch, H., Moeslund, T.B., Aanæs, H., Alkjær, T., Simonsen, E.B., 2014. Markerless motion capture can provide reliable 3d gait kinematics in the sagittal and frontal plane. Med. Eng. Phys. 36 (9), 1168–1175.

Sanders, R.H., Gonjo, T., McCabe, C.B., 2016. Reliability of three-dimensional angular kinematics and kinetics of swimming derived from digitized video. J. Sports Sci. Med. 15 (1), 158.

Sun, M., Telaprolu, M., Lee, H., Savarese, S., June 2012. An efficient branch-and-bound algorithm for optimal human pose estimation. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Taylor, W., Kornaropoulos, E., Duda, G., Kratzenstein, S., Ehrig, R., Arampatzis, A., Heller, M., 2010. Repeatability and reproducibility of ossca, a functional approach for assessing the kinematics of the lower limb. Gait Posture 32 (2), 231–236.

Taylor, W.R., Ehrig, R.M., Duda, G.N., Schell, H., Seebeck, P., Heller, M.O., 2005. On the influence of soft tissue coverage in the determination of bone kinematics using skin markers. J. Orthop. Res. 23 (4), 726–734.

Veksler, O., 2008. Star shape prior for graph-cut image segmentation. In: European Conference on Computer Vision. Springer, pp. 454–467.

Wu, G., Siegler, S., Allard, P., Kirtley, C., Leardini, A., Rosenbaum, D., Whittle, M., DLima, D.D., Cristofolini, L., Witte, H., Schmid, O., Stokes, I., 2002. {ISB} recommendation on definitions of joint coordinate system of various joints for the reporting of human joint motionpart i: ankle, hip, and spine. J. Biomech. 35 (4), 543–548.

Wu, G., van der Helm, F.C., Veeger, H.D., Makhsous, M., Roy, P.V., Anglin, C., Nagels, J., Karduna, A.R., McQuade, K., Wang, X., Werner, F.W., Buchholz, B., 2005. {ISB} recommendation on definitions of joint coordinate systems of various joints for the reporting of human joint motionpart ii: shoulder, elbow, wrist and hand. J. Biomech. 38 (5), 981–992.

Yang, S.X., Christiansen, M.S., Larsen, P.K., Alkjær, T., Moeslund, T.B., Simonsen, E.B., Lynnerup, N., 2014. Markerless motion capture systems for tracking of persons in forensic biomechanics: an overview. Comput. Methods Biomech. Biomed. Eng.: Imaging Visual. 2 (1), 46–65.

Yang, Y., Ramanan, D., 2013. Articulated human detection with flexible mixtures of parts. IEEE Trans. Pattern Anal. Mach. Intell. 35 (12), 2878–2890.