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.

© 2017 Elsevier Ltd. All rights reserved.

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

Download English Version:

https://daneshyari.com/en/article/5032092

Download Persian Version:

https://daneshyari.com/article/5032092

Daneshyari.com