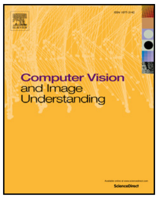
Computer Vision and Image Understanding 000 (2016) 1–16

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu



A local-global coupled-layer puppet model for robust online human pose tracking

Miao Ma[a,b,*], Naresh Marturi[b,c], Yibin Li[a], Rustam Stolkin[b], Ales Leonardis[b]

[a] Shandong University, Jinan, Shandong, 250061, PR China
[b] University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK
[c] KUKA robotics UK Ltd., Wednesbury Great Western Street, WS10 7LL, UK

ARTICLE INFO

Article history:
Received 31 August 2015
Revised 28 May 2016
Accepted 22 August 2016
Available online xxx

Keywords:
Human pose tracking
Human tracking
Video tracking
Pose estimation
Coupled-layer model

ABSTRACT

This paper addresses the problem of online tracking of articulated human body poses in dynamic environments. Many previous approaches perform poorly in realistic applications: often future frames or entire sequences are used anticausally to mutually refine the poses in each individual frame, making online tracking impossible; tracking often relies on strong assumptions about *e.g.* clothing styles, body-part colours and constraints on body-part motion ranges, limiting such algorithms to a particular dataset; the use of holistic feature models limits the ability of optimisation-based matching to distinguish between pose errors of different body parts. We overcome these problems by proposing a coupled-layer framework, which uses the previous notions of deformable structure (DS) puppet models. The underlying idea is to decompose the global pose candidate in any particular frame into several local parts to obtain a refined pose. We introduce an adaptive penalty with our model to improve the searching scope for a local part pose, and also to overcome the problem of using fixed constraints. Since the pose is computed using only current and previous frames, our method is suitable for online sequential tracking. We have carried out empirical experiments using three different public benchmark datasets, comparing two variants of our algorithm against four recent state-of-the-art (SOA) methods from the literature. The results suggest comparatively strong performance of our method, regardless of weaker constraints and fewer assumptions about the scene, and despite the fact that our algorithm is performing online sequential tracking, whereas the comparison methods perform mutual optimisation backwards and forwards over all frames of the entire video sequence.

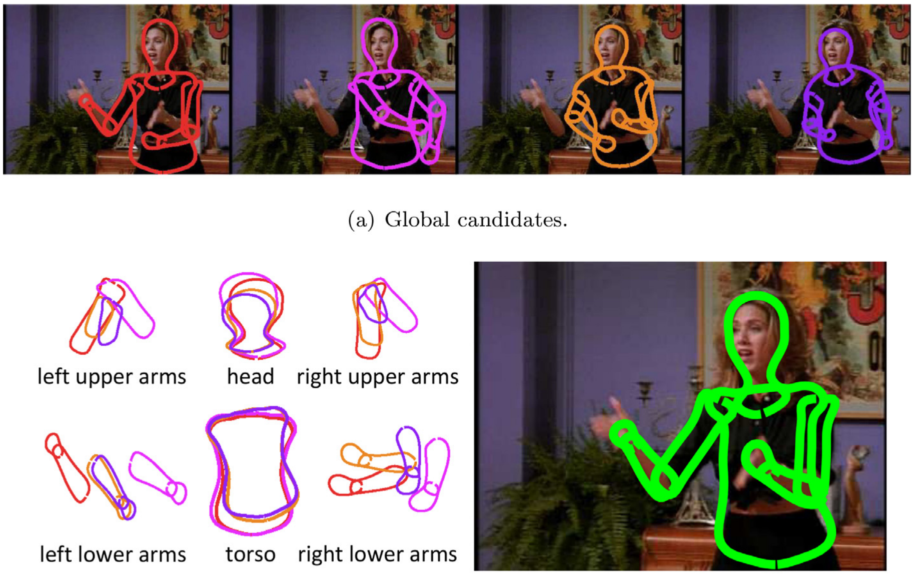
1. Introduction

Human pose estimation and tracking are increasingly popular research areas in computer vision, and have been studied for well over 30 years in the literature, *e.g.* Hogg (1983). There is growing interest in such algorithms for a variety of applications including activity recognition (Yao et al., 2011), video understanding (Burgos-Artizzu et al., 2012), gesture analysis (Charles et al., 2013), human-robot interaction (Agarwal et al., 2012), and others. Significant advances were made in recent years, however even state-of-the-art (SOA) methods often rely on strong assumptions and constraints in representing human bodies, such as visual appearance (Charles et al., 2013), scale (Sapp et al., 2011), lighting conditions, occlusions, and the ranges of motion of limbs and limb-parts. In this work, our goal is to sequentially track human body poses in monocular video frames obtained under variable conditions, where people move freely and interact with each other. Typical examples include videos of TV series or movies, where human appearance is unconstrained (*e.g.* variable background, any colour and type of clothing, no fixed scale, etc.). Many recent efforts have been devoted to track and estimate human poses from monocular video frames. Even though most of them perform well on certain body parts such as torsos and heads, their performance for arms is still not convincing. Within this context, we are most closely interested in tracking upper body poses, which include head, torso and arms, and in particular, improving the pose accuracy of lower arms. Nevertheless, our approach is not constrained for human upper body and can be easily adapted to the entire body. Our method is initialised from a single frame, and does not require any prior knowledge of the human clothing style, background scene or other conditions.

A variety of methods have been proposed in recent years to track and estimate the poses of articulated human bodies. However, many methods make use of the entire image sequence to

[*] Corresponding author.
E-mail address: mamiaosdu@hotmail.com (M. Ma).

http://dx.doi.org/10.1016/j.cviu.2016.08.010
1077-3142/© 2016 Elsevier Inc. All rights reserved.



(a) Global candidates.

(b) Local candidates

(c) Refined human pose

Fig. 1. Proposed coupled-layer model. (a) Different global pose candidates. (b) Local parts obtained by decomposing the global pose candidates. (c) Recomposed global pose.

mutually refine the poses in each individual frame, *e.g.* Cherian et al. (2014); Zuffi et al. (2013), rendering them only suitable for offline applications. In contrast, our method relies only on the previous frame information at any point in time, with computation only in the temporal direction, enabling online tracking applications. Since this reduction in available temporal information affects the overall performance, our method makes use of additional information from the spatial domain. For estimating articulated human pose, the overall information associated with the target makes the state space too large to compute. In this case, we exploit a local-global coupled-layer method, which uses the entire human body as a global layer and uses decomposed parts as a local layer (see Fig. 1). This type of methodology not only reduces the computational space and cost, but also improves the overall accuracy.

In this paper, we present an on-line coupled-layer method using discrete-structure puppets (Zuffi et al., 2012) for estimating the upper human body pose information. Recently published human pose estimation methods predominantly use an evaluation function to evaluate a candidate pose for the entire human body (Dantone et al., 2013; Yang and Ramanan, 2013). However, such methods can become prone to local convergence problems. For example, if one candidate pose suggests a correct left arm position, and an erroneous right arm position, and an alternative candidate pose is vice versa, then both candidates may generate similar evaluation scores. In this paper, we address this problem by decomposing the entire body into smaller parts and by estimating the pose separately for each of them. Nevertheless, if enough constraints are not provided, this decomposition method will also be unreliable, *e.g.* left and right arms may erroneously swap places and converge on each other's true image locations. To resolve this issue we introduce an adaptive penalty policy (Section 4.3.3) with our coupled-layer method to improve the scope of local parts pose searching. It also assists in tackling variable body scales and tuning any propagated erroneous poses.

The remainder of this paper is organized as follows. The methods that are closely related to our work are presented in Section 2. The proposed coupled-layer model is presented in Section 3, where we detail the model and explain the relationship between its local and global layers. Section 4 explains the tracking and estimation procedure, using the coupled-layer model. Section 5 presents experiments conducted using three different public benchmark datasets, where we compare the performance of our method against four other SOA pose estimation techniques. In this section, we also investigate the robustness of our method to various different levels of initialization error. Section 6 concludes the paper and the proposed method.

2. Related work

Numerous human pose estimation techniques, developed for a variety of applications, are available in the literature. In this section, we discuss the work most closely related to our proposed method.

The well-known *pictorial structures* (PS) model, proposed by Fischler and Elschlager (1973), is still drawing significant attention from researchers for its efficient tree-based inference algorithm (Dantone et al., 2013; Eichner et al., 2012; Park and Ramanan, 2011; Pishchulin et al., 2012; Yang and Ramanan, 2013). A key limitation of PS, and some extended models, is that the parts are treated as rigid templates and are represented as rectangular (or polygonal) regions. Later methods, such as *contour people* (Freifeld et al., 2010) and *deformable structures* (DS) model (Zuffi et al., 2012), that are derived from 3D human models, can better capture the 2D shape as non-rigid, deformable parts. However, due to the holistic nature of these models, several problems can arise e.g. in the case of rapid part motions or occlusions.

Several methods from the literature use some kind of hierarchical methodology or coarse-to-fine scheme for inference. For example, Wu and Huang (1999) used a two-layer model for hand motion tracking, where the palm motion is represented in the global model and the fingers motion in the local model. Paul et al. (2011) used a two-layer model which searches for the coarse location of the human body regions over the image sequence in one layer, and then estimates and refines detailed human body part poses over the image sequence in another layer. Lee and Nevatia (2006) proposed a three-layer model. An alternative strategy is to model each part separately (Felzenszwalb and Huttenlocher, 2005; Ferrari et al., 2009; Ramanan et al., 2005) and impose different constrains on different parts (Sigal and Black, 2006a). However, these methods estimate and evaluate the entire body