Computer Vision and Image Understanding xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

2D–3D pose consistency-based conditional random fields for 3D human pose estimation

Ju Yong Chang[a], Kyoung Mu Lee[*,b]

[a] Department of Electronics and Communications Engineering, Kwangwoon University, 20 Kwangwoon-ro, Nowon-gu, Seoul 01897, Republic of Korea
[b] Department of Electrical and Computer Engineering, Automation and Systems Research Institute, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea

ARTICLE INFO

Keywords:
Human pose estimation
Conditional random fields
Deep learning

ABSTRACT

This study considers the 3D human pose estimation problem in a single RGB image by proposing a conditional random field (CRF) model over 2D poses, in which the 3D pose is obtained as a byproduct of the inference process. The unary term of the proposed CRF model is defined based on a powerful heat-map regression network, which has been proposed for 2D human pose estimation. This study also presents a regression network for lifting the 2D pose to 3D pose and proposes the prior term based on the consistency between the estimated 3D pose and the 2D pose. To obtain the approximate solution of the proposed CRF model, the N-best strategy is adopted. The proposed inference algorithm can be viewed as sequential processes of bottom-up generation of 2D and 3D pose proposals from the input 2D image based on deep networks and top-down verification of such proposals by checking their consistencies. To evaluate the proposed method, we use two large-scale datasets: Human3.6M and HumanEva. Experimental results show that the proposed method achieves the state-of-the-art 3D human pose estimation performance.

1. Introduction

Human pose estimation is one of the most actively investigated problems in computer vision. Its goal is to infer the configuration of the human body from images or videos. Recently, single-image 2D human pose estimation has considerably advanced as a result of publicly available benchmark datasets (Andriluka et al., 2014; Johnson and Everingham, 2010; Sapp and Taskar, 2013) and discriminative methods such as deformable part models (Andriluka et al., 2009; Dantone et al., 2013; Felzenszwalb et al., 2008; Pishchulin et al., 2013; Yang and Ramanan, 2011) and convolutional neural networks (CNNs) (Carreira et al., 2016; Chen and Yuille, 2014; Newell et al., 2016; Tompson et al., 2015; 2014; Toshev and Szegedy, 2014; Wei et al., 2016; Yang et al., 2016). However, 3D human pose estimation from single images remains extremely challenging due to inherent ambiguities (Lee and Chen, 1985) in recovering 3D information from a 2D image. Other difficulties include large appearance variations, various types of body shape, (self-)occlusions, and huge solution space. Recent single-image 3D human pose estimation approaches can be broadly classified into two categories: prediction-based approaches and optimization-based approaches. The *prediction-based approaches* (Brau and Jiang, 2016; Ionescu et al., 2014a; 2014b; Li and Chan, 2014; Tekin et al., 2016) exploit training data to find a regression function that can directly generate a 3D pose from an input 2D image. The *optimization-based approaches* (Bogo et al., 2016; Kostrikov and Gall, 2014; Li et al., 2015; Simo-Serra et al., 2013; 2012; Yasin et al., 2016; Zhou et al., 2016) attempt to minimize an energy function including the prior terms that are usually based on 3D pose statistics.

In this study, we propose a new 3D human pose estimation method based on a conditional random field (CRF) framework with a *high-order 2D–3D pose consistency prior*. Our CRF defines the probability distribution over 2D human poses rather than 3D poses. The unary likelihood term is defined by using the 2D joint heat maps that are produced by conventional CNN-based 2D human pose estimation approaches (Insafutdinov et al., 2016; Pfister et al., 2015; Pishchulin et al., 2016; Tompson et al., 2015; 2014). The high-order 2D–3D pose consistency term is defined by the following steps. First, we directly estimate 3D pose from 2D pose using the 2D-to-3D pose-lifting network that can be obtained by training with ground-truth 2D and 3D pose data. Second, we re-project the estimated 3D pose onto the 2D image and then compare the re-projected 2D pose with the original 2D pose to measure the consistency by computing their differences. If the input 2D pose is normal and probable, then this consistency should be high. If otherwise, the consistency should be low. By inferring the maximum a

* Corresponding author.
E-mail addresses: jychang@kw.ac.kr (J.Y. Chang), kyoungmu@snu.ac.kr (K.M. Lee).

https://doi.org/10.1016/j.cviu.2018.02.004
Received 9 March 2017; Received in revised form 9 November 2017; Accepted 8 February 2018
1077-3142/ © 2018 Elsevier Inc. All rights reserved.

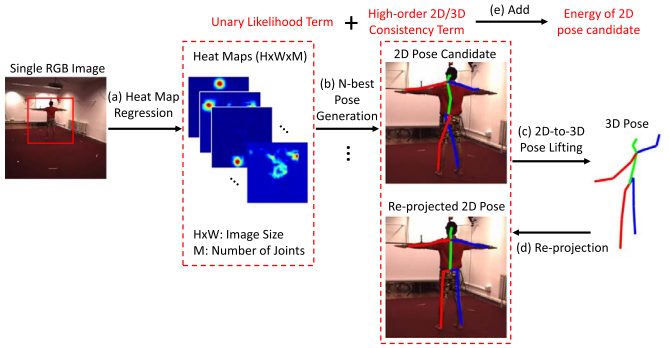

Fig. 1. Overview of the proposed method.

posteriori (MAP) estimate of the proposed CRF, we can find the most probable 2D pose and its corresponding 3D pose as a byproduct.

Therefore, the 2D-to-3D pose-lifting network plays a key role in the proposed method. Previous methods for 2D-to-3D pose lifting (Akhter and Black, 2015; Ramakrishna et al., 2012; Wang et al., 2014; Zhou et al., 2015) are usually based on time-consuming 3D reconstruction processes and assume orthographic camera projection, which results in suboptimal performance. Therefore, we propose the use of a multilayer perceptron (MLP) that is a simple feedforward neural network. This network directly regresses the 3D pose from the input 2D pose with high efficiency and can produce more accurate estimates by considering perspective projection.

Our CRF model requires the CNN-based heat map regression and MLP-based 2D-to-3D pose-lifting networks. Both are independently trained using the ground-truth RGB image, 2D pose, and 3D pose data. Inferring the exact MAP estimate of the proposed CRF model is a highly difficult task because of high-order prior term. Thus, we adopt the N-best strategy (Cherian et al., 2014; Park and Ramanan, 2011) to find the approximate solution. An overview of the proposed method is illustrated in Fig. 1. From an input RGB image, per-joint heat maps are generated using the heat map regression network, as shown in Fig. 1(a). The heat maps serve as the unary term describing the likelihood of each joint occurring in the 2D spatial location. Then, 2D pose candidates are obtained by applying the N-best pose generating procedure to the heat maps, as shown in Fig. 1(b). For each 2D pose candidate, we use the 2D-to-3D pose-lifting network to produce the 3D pose estimate, which is then re-projected onto the 2D image space as shown in Fig. 1(c) and (d). The re-projected 2D pose is compared to the original 2D pose candidate, which results in a high-order 2D–3D consistency term. By adding the unary likelihood term and 2D–3D consistency term, we can obtain the energy of the 2D pose candidate, as shown in Fig. 1(e). Finally, we find the minimum energy among the 2D pose candidates to obtain the optimal 2D pose and its corresponding 3D pose.

The main contributions of this work are as follows:

- In this study, we propose a new CRF model with a novel 2D pose prior term. Unlike the conventional priors that explicitly model the probability distribution of the 2D pose, our CRF model implicitly measures the plausibility of the 2D pose by computing the point estimate of the 3D pose and the consistency between the 2D and 3D poses. Our new 2D–3D pose consistency-based CRF can be combined with the N-best strategy to obtain the approximate solution with high efficiency because the optimization process relies only on two feedforward networks and simple arithmetic operations.
- We propose a simple but powerful 2D-to-3D pose-lifting method based on the MLP, which has two roles. First, it is used to constructively define our 2D–3D pose consistency prior. Second, by computing the prior of a 2D pose, its corresponding 3D pose can be automatically obtained as a byproduct through the proposed pose-lifting network. The proposed network does not require the assumption of orthographic projection and involved optimization processes but achieves state-of-the-art 2D-to-3D pose-lifting performance.
- We have conducted thorough experiments on two real datasets: Human3.6M (Ionescu et al., 2014b) and HumanEva (Sigal et al., 2010). We compare the proposed approach with recent 3D human pose estimation methods and show that ours produces the state-of-the-art results.

The remainder of this paper is organized as follows. We review the related works in Section 2. The proposed CRF model and inference procedure are presented in Sections 3 and 4, respectively. We provide the experimental results in Section 5 and the concluding remarks in Section 6.

2. Related works

2.1. 3D human pose estimation from 2D pose

A group of methods have tried to recover 3D human poses from 2D image landmarks. From the 2D images, the landmarks are usually given by manual annotation or automatic extraction using 2D human pose estimation methods. All recent 2D-to-3D pose-lifting methods (Akhter and Black, 2015; Ramakrishna et al., 2012; Wang et al., 2014; Zhou et al., 2015) use the 3D shape prior enforcing that valid 3D human shape variations should be represented by a linear combination of basis vectors. The 3D shape and viewpoint (i.e., camera extrinsic parameters) are then obtained by adopting a 3D-to-2D shape fitting process in which 2D re-projection errors are minimized. In Ramakrishna et al. (2012), a greedy orthogonal matching pursuit algorithm is proposed to reconstruct the shape and viewpoint from manually labeled 2D joints while encouraging anthropometric regularity. In Wang et al. (2014), an alternating direction method, which alternately updates the 3D shape and camera parameters, is presented and applied to the inaccurate 2D joints detected by a 2D pose estimator. In Akhter and Black (2015), a physically motivated prior based on pose-dependent joint angle limits is learned from a new dataset that includes an extensive variety of stretching poses. In Zhou et al. (2015), the authors focus on non-convexity in joint optimization of 3D shape and viewpoint, and propose a convex relaxation approach in which the joint estimation problem is formulated as a convex program and an efficient algorithm is developed on the basis of the alternating direction method of multipliers. All these methods assume orthographic camera projection and obtain the 3D human pose by estimating the 3D shape and viewpoint separately. Our method does not require the orthographic assumption and directly generates the 3D human pose in camera coordinates.

2.2. 3D human pose estimation from single image

Numerous studies have focused on 3D human pose estimation from single images. Early approaches perform automatic discriminative prediction of 3D pose from various image features (Agarwal and Triggs, 2006; Bo and Sminchisescu, 2009; 2010; Rosales and Sclaroff, 2001; Sminchisescu et al., 2007) or build a 3D model and then compute 3D pose by a generative model-image alignment process (Andriluka et al., 2010; Deutscher et al., 2000; Gall et al., 2010; Li et al., 2006; Sidenbladh et al., 2000; Sigal et al., 2004; Sminchisescu and Triggs, 2001; 2003; 2005). Recent methods tend to rely on the CNN architectures or successful 2D body joint detectors, both of which are usually discriminatively trained on a large amount of data. We classify the recent methods into two classes (i.e., prediction-based and optimization-based methods) and review them in the following paragraphs.

The *prediction-based approaches* (Brau and Jiang, 2016; Ionescu et al., 2014a; 2014b; Li and Chan, 2014; Tekin et al., 2016) directly estimate the 3D human pose from a 2D image. In Ionescu et al. (2014b), the authors present a large-scale structured prediction method that leverages the Fourier approximation of 2D histogram of oriented

Download English Version:

https://daneshyari.com/en/article/6937370

Download Persian Version:

https://daneshyari.com/article/6937370

Daneshyari.com