Contents lists available at ScienceDirect

Neurocomputing

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

Latent variable pictorial structure for human pose estimation on depth images

Li He [a], Guijin Wang [a,*], Qingmin Liao [b], Jing-Hao Xue [c]

[a] *Department of Electronic Engineering, Tsinghua University, Beijing 100084, China*
[b] *Department of Electronic Engineering/Graduate School at Shenzhen, Tsinghua Campus, Xili university town, Shenzhen 518055, China*
[c] *Department of Statistical Science, University College London, London WC1E 6BT, UK*

ARTICLE INFO

Article history:
Received 5 May 2015
Received in revised form
22 December 2015
Accepted 5 April 2016
Communicated by Xu Zhao
Available online 6 May 2016

Keywords:
Pose estimation
Pictorial structure
Latent variable
Body silhouette
Regression forest
Depth images

ABSTRACT

Prior models of human pose play a key role in state-of-the-art techniques for monocular pose estimation. However, a simple Gaussian model cannot represent well the prior knowledge of the pose diversity on depth images. In this paper, we develop a latent variable-based prior model by introducing a latent variable into the general pictorial structure. Two key characteristics of our model (we call Latent Variable Pictorial Structure) are as follows: (1) it adaptively adopts prior pose models based on the estimated value of the latent variable; and (2) it enables the learning of a more accurate part classifier. Experimental results demonstrate that the proposed method outperforms other state-of-the-art methods in recognition rate on the public datasets.

1. Introduction

Human pose estimation [1,2] is widely applied in human-computer interaction, smart video surveillance, and health care. Although a lot of efforts have been devoted to the research of pose estimation, it remains a very challenging problem in computer vision because of occlusion, high dimensionality of the search space and high variability in people's appearance.

The depth image obtained by the depth sensor [3–5] can provide 2.5D scene geometry, which facilitates both the segmentation of human body from background and the disambiguation of similar poses. Recently, the focus of pose estimation [6–9] has been shifted toward pose estimation on depth images. Most of these works can be divided into two categories: generative methods and discriminative methods.

Typical generative methods include the proposals in [9–12], in which a kinematic chain and a 3D surface mesh are built as the human body model. They treat the depth image as a point cloud over 3D space and apply a model-fitting algorithm, such as the iterative closest point (ICP), to the human body model to fit the 3D point cloud. Ye et al. [11], Ganapathi et al. [12] and Baak et al. [9] combine dataset searching and model fitting to approach the problem of 3D pose estimation. Ganapathi et al. [10] extend the ICP to an articulated model by enforcing constraints over the pose space. Although such methods do not need a training step, they suffer many drawbacks. For example, the accuracy depends on the surface mesh level [13] and the fitting usually needs long processing and inconvenient setups.

Compared with the generative methods, the discriminative methods do not iteratively fit models to the observed data. Rather they directly estimate the parameters about pose. Thus they can estimate the pose quickly and adapt to various conditions. They regard the human pose as a collection of different parts/joints and learn discriminative classifiers for the part/joint detection [6–8,14]. The most famous works on depth images are those based on random forest [6–8]. Shotton et al. [6] formulate the pose estimation as a classification task and use the random forests to learn the classifiers. Girshick et al. [8] convert the classification task to the regression problem for the estimation of the occluded parts. In [7], Sun et al. incorporate temporary states of the object, such as person's height and facing direction, to boost the performance of the classifiers. However, these methods infer locations of body joints either independently [6,8] or relying on some global information [7], neglecting the dependence between body joints.

It is natural to boost the pose estimation performance by adding constraints among joints. One of the most widely used approach in this direction is to use graph model-based prior

[*] Corresponding author. Tel.: +86 18911389502; fax: +86 62770317.
E-mail addresses: l-he10@mails.tsinghua.edu.cn (L. He), wangguijin@tsinghua.edu.cn (G. Wang), liaoqm@tsinghua.edu.cn (Q. Liao), jinghao.xue@ucl.ac.uk (J.-H. Xue).

http://dx.doi.org/10.1016/j.neucom.2016.04.009
0925-2312/© 2016 Published by Elsevier B.V.

structure, which was first proposed in [15] for general computer vision problems and later applied to the pose estimation problem in [16]. It assumes that the relationships among joints are state-constrained among the body parts. Two important components are defined in the model: one is the appearance model which represents the probability of a body part at a particular location in the given image; the other is the prior model which represents the probability distribution over pose space. To make a trade-off between computational efficiency and estimation accuracy, tree-structured models with a single Gaussian prior are commonly used [15–18]. However, as the diversity of human pose increases, a simple Gaussian prior usually leads to a poor model of human articulation, which cannot be applied well to the tasks on the depth images. This is mainly due to two reasons. One is that it is not an easy work to find a proper kernel number for the Gaussian model in a large dataset. A small number may cause a poor fitting of the prior, while a large number will cost extra computation and is prone to over-fitting. The other is that the method always applies the same prior model to test samples, even when they are of distinct poses. This limits the adaptability of the method. The works in [19,20] cluster poses into sub-clusters and learn a GMM for each sub-cluster to enhance the adaptability of prior model. However, at the inference stage, they need to infer all possible poses and select one as the final output. This makes the inference complex.

In this paper, we propose a novel framework called Latent Variable Pictorial Structure (LVPS) for pose estimation on depth images. We construct and estimate a latent variable based on the human silhouette. At the inference stage, our model rebuilds the appearance model and the prior model based on the values of the latent variable and then infers human poses. We shall show its effectiveness through experiments on public datasets. Compared with the state-of-the-art methods, our proposal can significantly increase the accuracy of pose estimation.

The rest of the paper is organized as follows. We overview the proposal in Section 2. Our LVPS model is introduced in Section 3 and its application to the pose estimation in Section 4. We present experiments and discussions in Section 5 and draw conclusions in Section 6.

2. Overview of the proposed method

Fig. 1 shows the framework of our LVPS. It consists of two main processes: the training stage indicated by green arrows and the inference stage indicated by blue arrows.

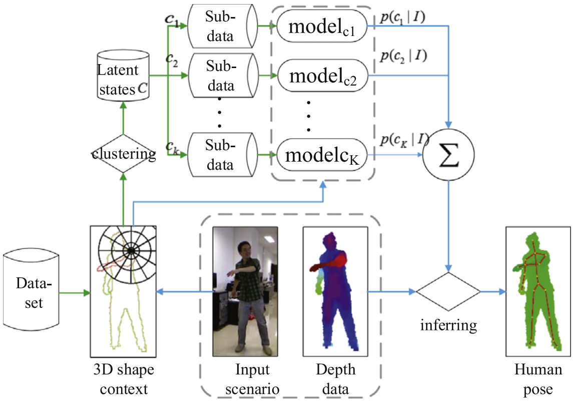

Fig. 1. The flowchart of the proposed method: the process with green arrows is the training stage and that with blue arrows is the inference stage. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

The training stage: The keys of the training stage involve generation and selection of the latent variable and the training of models. In our work, we extract silhouette features of poses, obtain their distributions, quantize the distributions into a set of states C, and use the state label as the latent variable. According to the value of the latent variable, all the training samples are partitioned into subsets. After that, we attach the value of the latent variable to each sample and treat each sample as a two-labels object: a body part label and a latent variable state. Samples with labels are then input into classifiers to learn appearance models and prior models. As a result, the diversity of the appearance and prior in each cluster would be reduced and the prior model can be better learned and the discrimination ability of the appearance model can be largely enhanced.

The inference stage: As the blue arrows indicate, to estimate one body pose on depth image I, we shall first evaluate its latent state. This is, the likelihood $p(c_i | I)$ is estimated. After that we rebuild our prior model and appearance model by assembling the learned models of individual clusters according to the likelihoods. As a result, our proposal adapts the models based on the specific test image.

3. Latent variable pictorial structure

A classical pictorial structure model of the human body was proposed in [15]. It assumes that the dependences between body joints can be expressed by a predefined graph, $G = (V, E)$, as shown in Fig. 2, where V and E denote the sets of nodes and edges in the graph G, respectively. We use $X = \{x_1, x_2, \ldots\}$ to denote the pose, in which x_i denotes the position of joint i. For the detection of an articular object, the objective function to be maximized when given image I can be written as

$$p^{PS}(X | I) \propto \left\{ \prod_{i \in V} \phi(x_i | I) \right\} \left\{ \prod_{(i,j) \in E} \phi(x_i, x_j) \right\}, \tag{1}$$

where $\phi(x_i | I)$ denotes the appearance likelihood, which models the probability of a part at a particular location and orientation given the input image I, and the factor $\phi(x_i, x_j)$ denotes a prior, which models the probability distribution over pose space. In this paper, the factor $\phi(x_i, x_j)$ describes the distribution of relative position between joint i and joint j.

In the most existing methods based on the general pictorial structure model, only one tree-structured Gaussian prior is used to speed up the inference, and the appearance models of individual

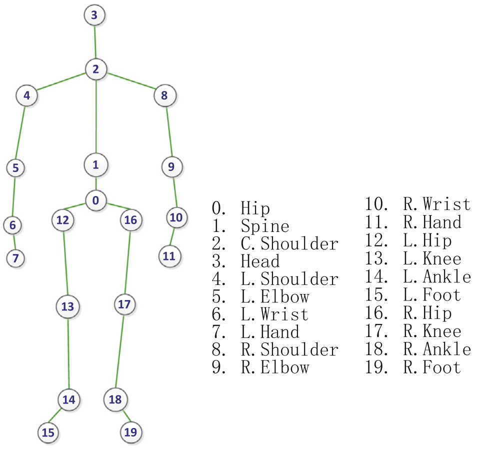

Fig. 2. The graph model on human pose. The circle with a number is a vertex in V, which presents a joint/part of the body; the line between two joints is an edge in E, which indicates that the connected joints/parts are dependent.