Neurocomputing 164 (2015) 210–219

Contents lists available at ScienceDirect

Neurocomputing

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

Depth-images-based pose estimation using regression forests and graphical models

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 2014
Received in revised form
23 January 2015
Accepted 27 February 2015
Communicated by Xu Zhao
Available online 9 March 2015

Keywords:
3D local shape context
Graphical models
Pose estimation
Regression forests

ABSTRACT

Depth-images-based human pose estimation is facing two challenges: how to extract features which are discriminative to variations in human poses and robust against noise, and how to reliably learn body joints based on their dependence structure. To tackle the first problem, we propose a novel 3D Local Shape Context feature extracted from human body silhouette to characterise the local structure of body joints. To tackle the second problem, we incorporate a graphical model into regression forests to exploit structural constrains. Experiments demonstrate that our method can efficiently learn local body structures and localise joints. Compared with the state-of-the-art methods, our method significantly improves the accuracy of pose estimation from depth images.

1. Introduction

Accurate estimation of human poses is a key step for many visual applications, such as human computer interaction, smart video surveillance, character animation and augmented reality. A nice review on this topic can be found in [1]. Although considerable research efforts have been devoted to it, pose estimation is still a challenging task due to cluttered background, occlusion and variation in appearance and pose [2]. Most techniques address these challenges from two aspects: one seeks discriminative and robust features to fight against noise and variations in appearance and pose, and the other designs graphical models to utilise structural information to constrain the distributions of body joints.

With respect to the features for pose estimation, a variety of discriminative features have been developed [3]. Recently, with the development of depth sensor techniques (such as Kinect or time-of-flight sensors), many works focus on extracting features from depth images [4,5]. A depth image represents depth measurements of the scene [6–8]. Compared with RGB images, depth images supply much richer geometrical information, facilitating both the separation of human body from background and the disambiguation of similar poses. Generally, appearance and shape are the commonly used features for pose estimation. As to depth-appearance-based features, Plagemann et al. [4] proposed a geodesic-distance-based feature, which costs a large amount of computation for iteratively calculating points of interest, and Shotton et al. [5] proposed DCF (depth comparison features), which describe body parts by depth differences at a sequence of random offsets. Their works yielded state-of-the-art results. Their features are effective and efficient on depth images, from which many of later works [9–12] benefited. As to depth-shape-based features, Li et al. [13] proposed a shape-based feature, termed 3DSC, which utilised depth information to obtain an edge-point mask and calculated silhouette histograms on this 2D mask to detect end-points of interest. As extracted on the mask images, their features lack the 3D information. Furthermore, their framework only processes limited endpoints (e.g. head, hand and foot). Baak et al. [14] and Ye et al. [15] used point cloud matching techniques for pose estimation, which are computationally demanding. To the best of our knowledge, it seems that none of the shape-based features have achieved a performance comparable to DCF yet. In our work, we aim to propose a novel depth-shape-based feature which can attain satisfactory results.

With respect to the models of human pose, the pictorial structure model [16] is one of the most popular models, for its effective representation of articulated objects and its efficient inference algorithm. It is trained to learn the spatial relationship between pairs of joints, since the location of a joint is well constrained by its connected joints. At its inference stage, the likelihood of each body joint is evaluated over the 2D/3D space restricted by the trained model. Many improvements of this model have been made, and the most relevant

[*] 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.2015.02.068
0925-2312/© 2015 Published by Elsevier B.V.

work goes in either of three directions: to build more reliable body part (or joint) detectors [17–21], to introduce richer body models [22–27] or to perform inference [24,28] by imposing temporary constraints. In the first direction, many methods tend to be finely tuned to a specific dataset. In the other two directions, complex models and inference require extensive computation. As we know, most of these methods could hardly provide a real-time output due to the complexity of part detection and inference on RGB images. In recent years, some joint detection algorithms using random forests give real-time state-of-the-art results [29–33]. However, they infer locations of body joints either independently [5,10] or relying on some global latent variables [9], neglecting the dependence between body joints. Dantone et al. [21] designed two-layers regression forests to learn more reliable joint detectors and modelled the constraints by using Gaussian distributions for efficient inference on RGB images. Yu et al. [34] integrated action detection and cross-modality regression forests for the estimation of 3D human pose.

In this paper, we propose a novel framework for human pose recognition. It mainly consists of two modules. Firstly, we propose a new depth-shape-based feature, termed 3D Local Shape Context feature (3DLSC), by extending the 2D Shape Context (2DSC) [35] to 3D space, to characterise the location cues between human silhouette and joints. Different from 3DSC [13], our 3DLSC captures relative position information of silhouette points in 3D space. Thus our feature is body-size invariant and efficiently adaptive to persons with different heights. Experiments demonstrate that our shape-based features could achieve comparable results with the widely used DCF for pose estimation on depth images. Secondly, we propose a combined learning scheme by incorporating a data-dependent pictorial structure into regression forests. More specifically, depending on the training data arriving at the leaf nodes of the regression forests, our model can learn distributions of each joint and spatial constrains between adjacent joints. Different from the general pictorial structure [16], our proposal models relative distributions according to the specific test image. Compared with the state-of-the-art methods, our proposal can significantly increase the accuracy of pose estimation.

The rest of the paper is organised as follows. In Section 2, we present the construction of our 3DLSC feature, which consists of two steps: silhouette extraction and histogram binning. The details of our graphical models and regression forests are presented in Section 3. Finally, experiments and discussion are shown in Section 4 and conclusion and future work are given in Section 5.

2. 3D local shape context

In this section we present our 3DLSC feature. In [35] the 2DSC feature was first proposed for shape matching. It has been applied to pose estimation as it efficiently encodes local information of human silhouette by using histograms at logarithmic polar (log-polar) coordinates [36,37,13]. However, it faces two problems: (1) it is usually noisy in the body silhouette obtained by motion detection and it is difficult to extract inner edges due to the ambiguity on clothing texture [36]; (2) it is ill-conditional to recover 3D poses from 2D silhouettes due to lack of depth information. To mitigate these problems, we extract our features from depth images, which not only supply 3D information of human body but also facilitate the extraction of inner edges. In a similar spirit to [38] but using a different strategy and targeting a different task, we develop novel 3D local features by computing feature histograms at regularly spaced points on the edges of body silhouette extracted from depth images. Therefore, our feature construction consists of two steps: silhouette extraction and histogram binning.

2.1. Silhouette extraction

Given depth image **I**, we assume that the foreground of human body is already known. What we need to do is to extract outer and inner edges from the depth image. To reduce the influence of noise from depth sensors, we first use a Gaussian filter to smooth the extracted body shape.

The Gaussian filtering for depth d_i at pixel p_i is defined as

$$\hat{d}_i = \frac{1}{S} \sum_{j \in N(i)} \mathcal{G}(\mathrm{dst}(i,j); 0, \sigma^2) d_j, \tag{1}$$

where $\mathcal{G}(\cdot)$ is a Gaussian smooth function with mean zero and variance σ^2, $N(i)$ is the 3×3 neighbourhood of p_i, $\mathrm{dst}(i,j)$ indicates the distance between points p_i and p_j in 3D space, and S is a normalising constant. The Gaussian filter can effectively reduce noise in depth measurements; Fig. 1 shows the effect of smoothing on silhouette extraction.

A body silhouette on the depth image is a point set of edge points. In order to extract silhouette points, depth values of background pixels are set to ∞. As a result, the set of silhouette points **E** is obtained by using a local depth extrema function:

$$\mathbf{E} = \left\{ p_i : \max_{j \in N(i)} (\hat{d}_j - \hat{d}_i) > t_d \right\}, \tag{2}$$

where parameter t_d is a depth threshold set to 4 cm in our experiments.

There are often many thousands of points in **E**, which not only are too dense for shape description but also cost a large amount of computation. Hence, we uniformly down-sample **E** to a subset $\mathbf{E}'$ of N points with N=300–500.

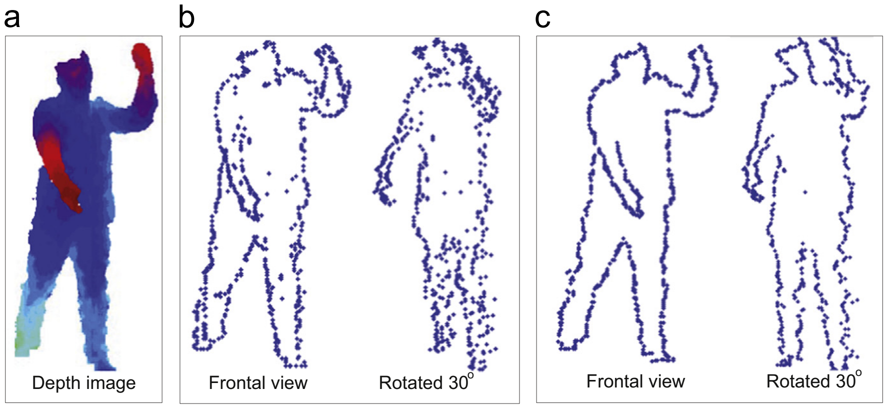

Fig. 1. Human body silhouette extraction of human body: (a) a human body depth image; (b) silhouette points without smoothing; (c) silhouette points after smoothing.