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/

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.

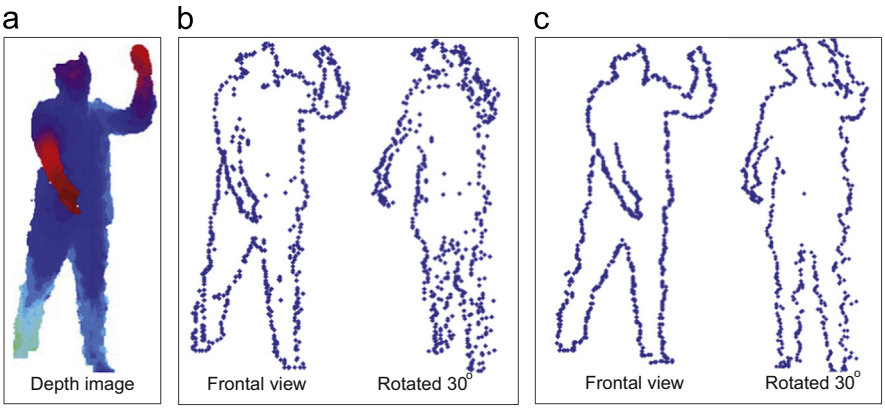


Fig. 1. Human body silhouette extraction of human body: (a) a human body depth image; (b) silhouette points without smoothing; (c) silhouette points after smoothing.

2.2. Histogram binning

To capture 3D information, we transform each point in $\mathbf{E}'$ to 3D space and then calculate a 3D feature histogram for it.

Specifically, for each point p_i in subset $\mathbf{E}'$, we could calculate 3D spatial parameters $\{\rho_{ij}, \theta_{ij}, \varphi_{ij}\}$, which include the radius ρ, the azimuth angle θ and the zenith angle φ, relative to other $N-1$ points. The specific meanings of spatial parameters are illustrated in Fig. 2. Then we project them to a 3D histogram h_i. The histogram h_i consists of 360 bins, $h_i(k)$ for $k=1,\ldots,360$, containing 5 bins of radius ρ, 12 bins of angle θ and 6 bins of angle φ. The value of its kth bin $h_i(k)$ is calculated by

$$h_i(k) = \frac{1}{C_i} \sum_{j=1, j\neq i}^{N} v_{ij}(k), \tag{3}$$

where $C_i = \sum_{k,j\neq i} v_{ij}(k)$ is a normalisation constant, $v_{ij}(k)$ is the weight projected to the kth bin of histogram h_i from point p_j and its value is

$$v_{ij}(k) \propto \frac{1}{\left\|\rho_{ij}-\rho(k)\right\| \left\|\theta_{ij}-\theta(k)\right\| \left\|\varphi_{ij}-\varphi(k)\right\|}, \tag{4}$$

where parameters $\{\rho(k), \theta(k), \varphi(k)\}$ are the middle values of the kth bin. We distribute v_{ij} only to the two bins with the closest centres

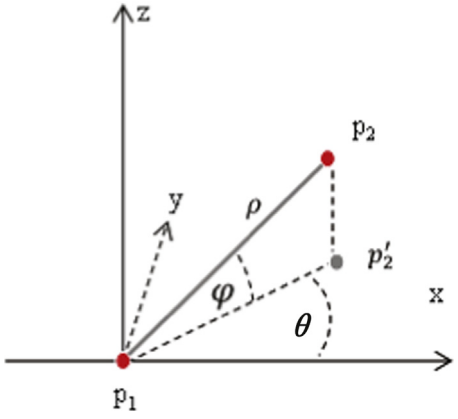


Fig. 2. Calculation of 3DLSC parameters (ρ, θ, φ) of p_2 relative p_1: p_2' is the projected point of p_2 on XY plane, ρ is the distance of two points, θ is the azimuth angle of p_2 on the XY plane and φ is the zenith angle of p_2 to the XY plane.

in each of the three dimensions, i.e. eight bins in total. This can be seen as a kind of smoothing for quantising histograms, which is widely used in the HOG and SIFT features, and can reduce the influence of outliers on depth silhouette. The whole shape or a pose is thus encoded by N histograms denoted by $\mathcal{H}$.

To make the feature local and invariant to different scales of body shapes such as those for adults and children, for each h_i, we only consider the points that fall within the range of diameter $\overline{\rho}$ ($\overline{\rho}$ is the mean distance of all ρ_{ij}) and normalise radius ρ_{ij} by $\overline{\rho}$. Besides, to strengthen discrimination in the near region and weaken the influence of the far region, we convert radius to the log-polar space. Meanwhile, as with [36], we do not normalise shape contexts with respect to their dominant local orientations.

Our 3DLSC is a robust local descriptor of human pose. It contains a great deal of pose information, since locations of body joints can be easily identified when the 3D shape of a person is given. We shall evaluate features in our experiments.

3. Graphical models and regression forests

There have been some methods proposed to learn mappings from shape feature to human pose [36,37,13]. However, they are either easily affected by ambiguous shape for single pose estimation [36,37] or designed for some specific end-points detection [13]. This makes them unfit for joints detection. Recently, regression forests have proved to be an efficient algorithm for pose recognition [9,10]. They can handle high-dimensional feature vectors and are of low computational complexity. Therefore, we would take the advantages of random forests to learn regressors for pose estimation. However, in previous work [36,37], distributions of body joints were learnt independently, in spite of the fact that there is strong dependence between connected joints such as if the position of the elbow is given along with some information of nearby shapes, the position of hand will be strongly constrained.

Therefore, as [21] we incorporate the pictorial structure into regression forests to learn a mapping from the 3DLSC feature space to human poses. We learn two kinds of weak regressors based on the training samples arriving at a leaf node: one is an independent estimator about joint x_i, and the other is a spatial constraint estimator for a pair of joints (x_i, x_j) if they are connected in the topological graph defined by Fig. 3(a). We estimate the distributions of joints x_i and pairs (x_i, x_j) by using the Gaussian Parzen density estimators. As a result, our method can provide structured outputs of joints' positions.

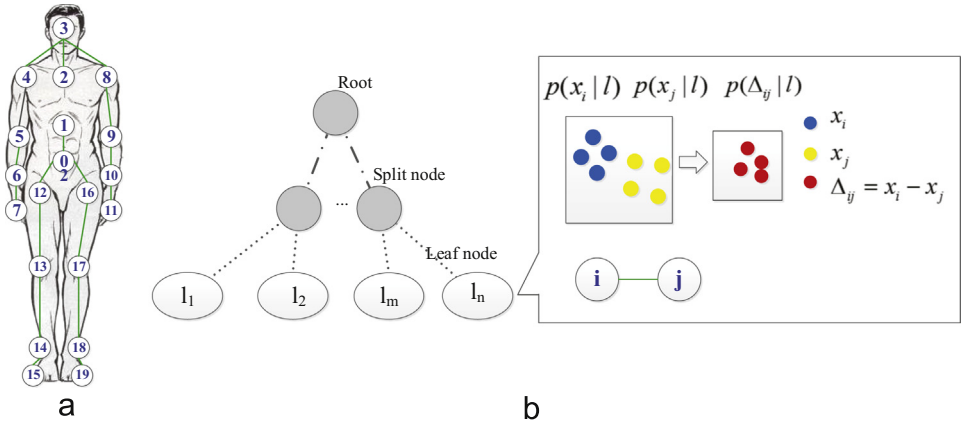


Fig. 3. (a) The graphical model we define for human body joints, with green lines indicating the dependence between connected joints. Our graph is divided into two subgraphs, mainly for two motivations: we experientially assume that there is no strong dependency between upper and lower body parts and the division can accelerate inference by parallelisation. (b) Our regression forest which learns independent and joint distributions of joints at leaf nodes: blue/yellow dots in squares indicate the votes in 3D space of joint x_i/x_j; red dots indicate their relative positions. (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this paper.)

3.1. Graphical models

We use a classical pictorial structure to model the human body [16]. Assume that the dependence between body joints can be expressed by a predefined graph, $G=(V,E)$, as shown in Fig. 3(a), where V and E denote the sets of nodes and edges in the graph G. The nodes $i=1,2,\ldots,J\in V$ correspond to the joints located at x_i of human body, and edges $(i,j)\in E$ describe their relations. Our graphical model can be divided into two subgraphs: one illustrates the structure of the upper part of human body, and the other for the lower part.

The human pose configuration $\mathbf{X}=\{x_1,x_2,\ldots,x_J\}$ can be estimated by maximising the posterior distribution given an observed silhouette or a bag of 3DLSCs $\mathcal{H}$:

$$p(\mathbf{X}|\mathcal{H})\propto p(\mathcal{H}|\mathbf{X})p(\mathbf{X}), \tag{5}$$

where $p(\mathcal{H}|\mathbf{X})$ is the likelihood and often factorised as $\prod_{i\in V}\phi(x_i)$, and $p(\mathbf{X})$ is the prior on pose configurations and often defined as $\prod_{i,j\in E}\phi(x_i,x_j)$ by applying the constraints of connected joints in G. As a result, the posterior of a human pose can be rewritten as

$$p(\mathbf{X}|\mathcal{H})\propto \prod_{i\in V}\phi(x_i)\cdot\prod_{i,j\in E}\phi(x_i,x_j), \tag{6}$$

where $\phi(x_i,x_j)$ is the structural constraint between joints x_i and x_j and its specific expression would be discussed in Section 3.3.

In contrast to the work of [9,10], in which body joints are treated independently, our model possesses an additional term $\prod_{i,j\in E}\phi(x_i,x_j)$ to impose spatial constraints when localising x_i.

In our work, we utilise regression forests to learn the parameters of $\phi(x_i)$ and $\phi(x_i,x_j)$. At each leaf of regression forests, we use a compact representation [10] to store the distribution of each body joint of interest and the joint distribution of each pair of connected joints there. Then we estimate $\phi(x_i)$ and $\phi(x_i,x_j)$ using the Gaussian Parzen density estimation.

3.2. Learning

This section presents how to learn the structure of a regression forest and the corresponding parameters.

3.2.1. Learning tree structures

A regression forest, denoted by $\mathbf{T}=\{T_1,\ldots,T_t\}$, consists of t regression trees and each tree can be grown by a sequence of learnt splitting functions $\{f^*\}$ as below [39,40].

Let $\mathbf{S}=\{p_i\}$ denote a set of sampled points collected from training images. At the root node of a tree, training data $\mathbf{S}$ can be split into two subsets, the left $\mathbf{S}^L$ and right $\mathbf{S}^R$ subsets, by evaluating a splitting function f. The splitting function f compares the value of the kth feature $h_p(k)$ of 3DLSC of point p with threshold η:

$$f(p;k,\eta)=\begin{cases}0 & \text{if } h_p(k)<\eta,\\ 1 & \text{otherwise.}\end{cases} \tag{7}$$

At each node, there is a group of candidate splitting functions with values randomly set for parameters $\{k,\eta\}$ and, among them, the selected function f^* is the one that minimises an entropy:

$$f^*=\underset{f}{\operatorname{argmin}}\sum_{m\in\{R,L\}}\frac{\|\mathbf{S}^m\|}{\|\mathbf{S}\|}H(\mathbf{S}^m), \tag{8}$$

where

$$H(\mathbf{S}^m)=-\sum_{j=1}^{J}\frac{\sum_{p_i\in\mathbf{S}^m}p(x_j|p_i)}{\|\mathbf{S}^m\|}\log\frac{\sum_{p_i\in\mathbf{S}^m}p(x_j|p_i)}{\|\mathbf{S}^m\|}, \tag{9}$$

in which $p(x_j|p_i)$ denotes the contribution of, or the weight carried by, p_i to the localisation of joint x_j, defined as

$$p(x_j|p_i)\propto\exp\left(-\frac{\|\Delta_j^{p_i}\|_2^2}{b}\right)\delta(\|\Delta_j^{p_i}\|_2<\overline{\rho}), \tag{10}$$

where $\Delta_j^{p_i}=x_j-p_i$ presents the 3D distance between p_i and x_j and $\delta(\cdot)$ is a delta function to make sure that only the joints x_j close to p_i will be considered. The parameter b controls the steepness of contribution of p_i to x_j; we set its value $\overline{\rho}^2$. Since a point may be associated with more than one joint, we normalise $p(x_j|p_i)$ to make $\sum_j p(x_j|p_i)=1$. Then, at each subsequent child node, the splitting procedure will be applied recursively until stopping criteria are met. The criteria usually include the minimum number of samples arriving at the leaf node and the maximum depth of the tree.

3.2.2. Regressors of $\phi(x_i)$ and $\phi(x_i,x_j)$

After the tree is constructed, at each leaf node l, we learn the expressions of the regressors of $\phi(x_i)$ and $\phi(x_i,x_j)$ based on the samples arriving at this leaf. In order to compactly present their distributions, we use the Gaussian Parzen functions. So in this paper, we learn and store their corresponding parameter set: two kinds of sets of offset centres, $\mathbf{V}_{(i)l}$ and $\mathbf{V}_{(ij)l}$.

Each element in the joint-offset $\mathbf{V}_{(i)l}$, $\Delta_i^p=x_i-p$, presents the offset from the sampled point p to the predicted joint x_i. Each element in the structure-offset $\mathbf{V}_{(ij)l}$, $\Delta_{ij}=x_i-x_j$, presents the offset from joint x_i to joint x_j. We use a mean-shift algorithm with a Gaussian kernel of a fixed bandwidth to cluster the offsets. The centres of the largest K clusters (K=1,2) are selected as the representative offsets (termed relative votes [10]), and the sizes of the clusters are used as the vote weights (termed confidence weights [10]) of the offsets. Hence, two sets $\mathbf{V}_{(i)l}$ and $\mathbf{V}_{(ij)l}$ of these two kinds of relative votes and vote weights at leaf node l are denoted by

$$\mathbf{V}_{(i)l}=\{\Delta_{k(i)l},w_{k(i)l}\}_{k=1}^{K_{(i)}} \tag{11}$$

and

$$\mathbf{V}_{(ij)l}=\{\Delta_{k(ij)l},w_{k(ij)l}\}_{k=1}^{K_{(ij)}}, \tag{12}$$

where $\mathbf{V}_{(i)l}$ is the set of joint-offsets with its centre $\Delta_{k(i)l}$ and weight $w_{k(i)l}$, and $\mathbf{V}_{(ij)l}$ is the set of structure-offsets with its centre $\Delta_{k(ij)l}$ and weight $w_{k(ij)l}$. The specific expressions of $\phi(x_i)$ and $\phi(x_i,x_j)$ will be stated at the inference stage in next Section 3.3.

3.3. Inference

Consequently, at the inference stage for a given depth image, we can sample N silhouette points $\{s_1,\ldots,s_N\}$; each point s would reach a leaf node l of every tree in the forest, through evaluating the binary splitting functions stored at splitting nodes, and can provide weighted estimators for the location of x_i:

$$\{x_i^s(k),w_{k(i)l}\}_{k=1}^{K_i}, \tag{13}$$

where $x_i^s(k)=s+\Delta_{k(i)l}$, called the absolute vote from point s to joint x_i.

We use S_i to denote the set of absolute votes obtained from all sampled points to x_i. Generally, S_i contains thousands of elements and we shall select M largest weighted elements as a subset S_i'. We shall discuss the size M of S_i' in the experiments. Then using these absolute votes, we can estimate the likelihood of joint x_i using a Gaussian Parzen density estimator as

$$\phi(x_i)=\sum_{m}^{|S_i'|}\omega_m\exp\left(-\frac{\|x_i-x_i(m)\|^2}{h_i^2}\right), \tag{14}$$

where $x_i(m)$ is the mth absolute vote for x_i in S_i', ω_m is the corresponding weight, and h_i is the empirical joint bandwidth and we set it to 0.07 m in our experiments.

In the same way, we obtain its spatial constraint vote from leaf l to the connected joints x_i and x_j, and thus for all sample points we obtain a set S_c. A subset S_c' of S_c is used to estimate $\phi(x_i, x_j)$ by using a Gaussian Parzen density estimator:

$$\phi(x_i, x_j) = \sum_{m}^{|S_c'|} \omega_m \exp\left(-\frac{\|x_i - x_j - \Delta_{ij}(m)\|^2}{h_{ij}^2}\right), \tag{15}$$

where $\Delta_{ij}(m)$ is the mth relative vote for (x_i, x_j) in S_c', ω_m is the corresponding weight. The parameter h_{ij} is the limb-length bandwidth of two connected joints and we set it to the mean of limb-length calculated by using the training data. By collecting the votes from sampling points, our spatial constraints are data-dependent and it also makes our proposed method different from the general pictorial structure [16,17].

As discussed in Section 3.1, the best configuration of human pose in a pictorial structural model is searched by using the following expression:

$$\mathbf{X}^* = \underset{\mathbf{X}}{\operatorname{argmax}} \prod_{i \in V} \phi(x_i) \cdot \prod_{i,j \in E} \phi(x_i, x_j). \tag{16}$$

Note that our location variables are continuous in 3D space, which makes searching for the best solution computationally demanding. Therefore, we give an approximate solution for the output of pose using iterative methods. Specifically, we use a mean-shift algorithm independently over the likelihood space $\phi(x_i)$ to find one or two candidates for each joint as the initial pose. Then a gradient descent algorithm is applied to (17) to get the final result:

$$G(\mathbf{X}) = \prod_{i \in V} \phi(x_i) \cdot \prod_{i,j \in E} \phi(x_i, x_j). \tag{17}$$

4. Experiments and discussion

4.1. Datasets

In this section, we evaluate our algorithm for human pose estimation on two depth datasets, the Stanford dataset [41] and our THU pose dataset. There is a similar dataset [5], but it is not publicly available yet.

The Stanford dataset consists of 28 action sequences, which include 7891 images in total with a resolution of 176×144. All the images were captured from frontal view using a ToF camera in a lab environment. Among the images, 6000 are selected for training and the rest, less than 2000, are for testing.

The THU dataset contains 4500 depth images captured by a Kinect camera, which consists of 4 persons performing general actions (mostly upper limbs movements) and a limited torso direction (-60°, 60°). Some samples are shown in Fig. 4. We use manually labelled landmarks as the ground truth. Among the images, 3600 are randomly selected for training and the rest are for testing.

4.2. Preprocessing of the training data

Since we assume that the foreground of human body is already known, some preprocessing should be done to the input samples. For the samples in the THU dataset we use the foreground obtained by Kinect SDK, while for each depth image sequence in the Stanford dataset we segment the foreground from background using a motion-based method [42]. We show some segmentation results in the Stanford dataset in Fig. 5.

4.3. Implementation details

Our implementation of regression forests adopts settings similar to those of [10]: our forest is an aggregation of 3 trees with the maximum depth of 20 and the minimum samples of 20; at each splitting node, 12 randomly selected features with 100 random thresholds each, i.e. 1200 random tests in total, were used to find the best splitting.

4.4. Performance evaluation

We compare our algorithm with the state-of-the-art methods in [10,5,41,13], using two measures: the average error and the mean of average precision (mAP). The average error for each joint is calculated by averaging the Euclidean distance between the estimate and its ground truth. The mAP is obtained as the ratio of the most confident joint hypothesis within the distance tolerance $\tau = 0.1$ m, as with [10].

Figs. 6–8 and Table 1 show the performance of our algorithm, denoted by 'ours (3DLSC + Graph Model)', the regression method of [10], denoted by 'Girshick et al.' and three other algorithms [5,41,13]. From the figures, we can find that our algorithm achieves better results than that of [10]. More specifically, our algorithm obtains 3.6 cm in the average error and 95.2% in mAP on the THU dataset, and 3.5 cm in the average error and 98.2% in mAP on the Stanford dataset, which surpass those of [10]. Moreover, the superior results can be remarkably observed at limb ends, such as elbow, wrist and hand, which we think benefits from the use of graphical models. Compared with the results reported in [5,41,13], our performance is superior as well. Note that in Table 1 we evaluate our proposed method only on three joints (hand, head and foot) and set the distance tolerance $\tau = 0.2$ m to make the evaluation consistent with [13].

Besides, we test the speed of our algorithm to process one image on the Stanford dataset. With our non-optimised code, it runs the processing at about 36 fps on our 4-cores computer.

4.5. Module performance

We further evaluate the effects of our modules separately: (1) for the first module, we replace the features with 2DSC [36] and DCF [5], and preserve the rest part; (2) for the second module, we remove the graphical model from our method. For a fair comparison, all of these experiments used the same parameters and settings of regression forests and the graphical model.

Table 2 compares the performance between our 3DLSC feature, 2DSC [36] and DCF [5] on the Stanford dataset. We can observe that our 3DLSC obtains the mAP of 98.2%, exceeding 2DSC by 10% and also moderately better than DCF. This indicates the discriminative power of our 3DLSC for pose estimation.

Furthermore, we investigate the effect of the graphical model. In experiments, we compare the results obtained from inferring each joint independently with those obtained from inferring each joint by exploiting the constraints in the graphical model. It is shown in Figs. 9 and 10 that, after adding the graphical model, our method yields a better performance. Fig. 9 shows that the average error decreases about 1.0 cm (from 4.5 cm to 3.5 cm), and mAP increases 4.0% (from 94.2% to 98.2%) at $\tau = 0.1$ m on the Stanford dataset. Fig. 10 displays specific samples. Fig. 10(1) and (2) shows the situation that the ambiguities occur when the left and right body parts are close in space; Fig. 10(3) shows the case that a body part (foot) might affect the detection results of another part (hand). To show our results clearly, we give another two results from the left-side and the top views. It can be observed that the graphical model helps us to reduce the ambiguities among similar

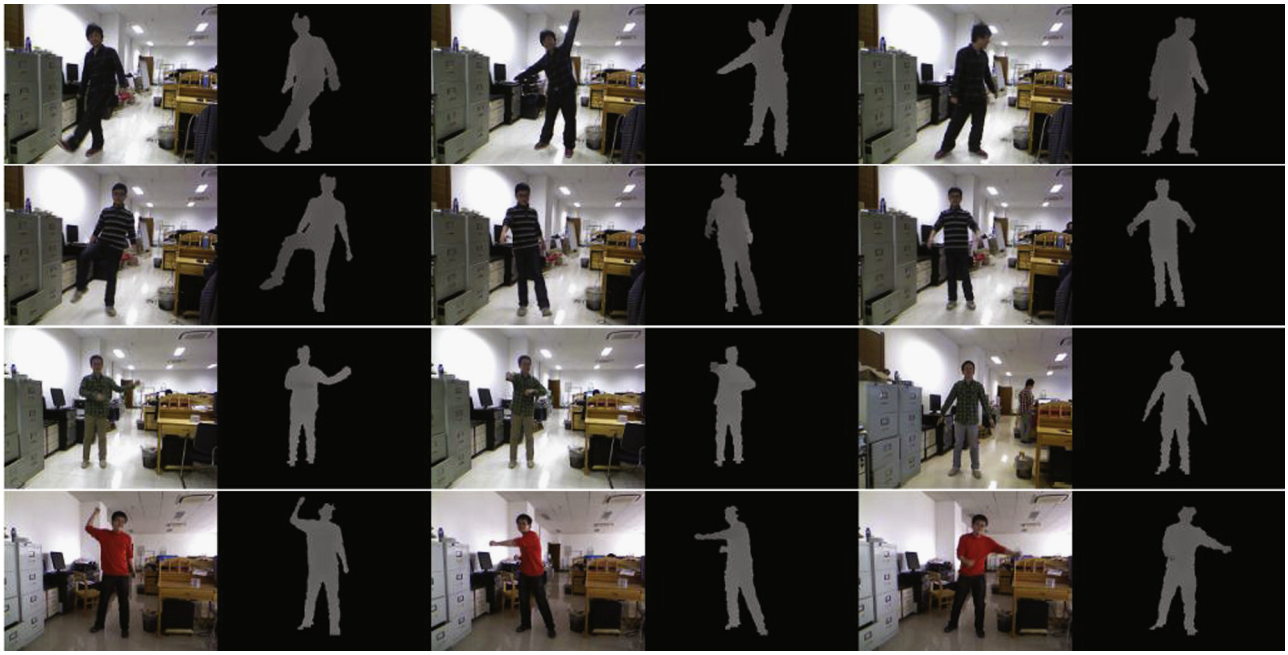


Fig. 4. Samples from the THU dataset: RGB and depth images.

Fig. 5. Results of foreground segmentation in the Stanford dataset: pairs of original and foreground images.

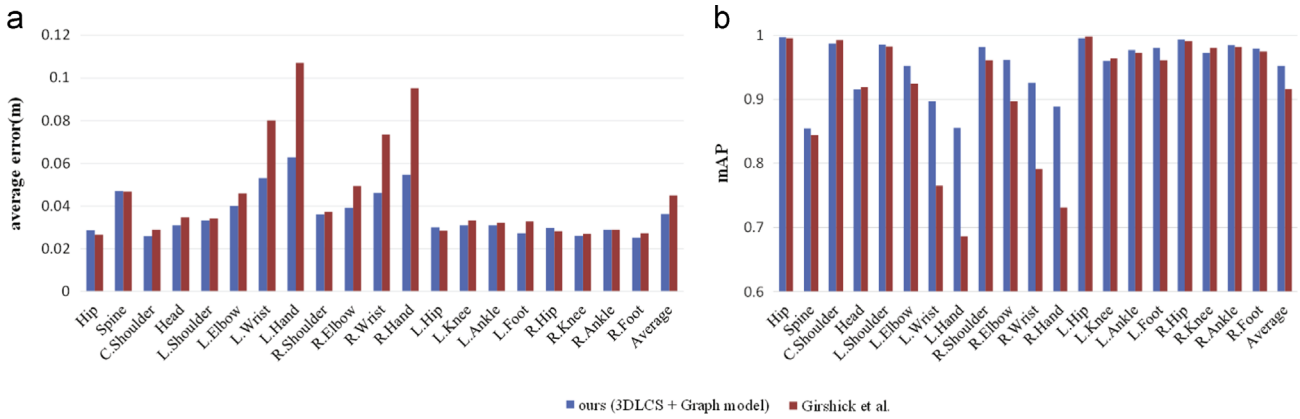


Fig. 6. Performance of two methods on the THU dataset: (a) average estimation error vs. body joint; (b) mAP vs. body joint.

joints and the influence of other body parts. This indicates the positive effect of using the graphical model for pose estimation.

4.6. Discussion

In this section we investigate the effects of five main factors that may affect the estimation accuracy of our method. These factors are the Gaussian filtering, the histogram multi-binning, the number of silhouette points, the votes normalisation factor $\overline{\rho}$ and the sizes of subsets S_i' and S_c'. At the end of this section, we also discuss the scalability of our model.

4.6.1. Gaussian filtering and histogram multi-binning

Fig. 11 illustrates the effects of two smoothing factors, the Gaussian filtering and the histogram multi-binning, that correspond to the procedures represented by (1) and (4), respectively. It shows that the pose-estimation performance will drop when there is no Gaussian filtering or multi-binning for the construction of 3DLSC; between these two factors, removing multi-binning leads to a greater drop.

Such positive effects of the Gaussian filtering and the multi-binning may be attributed to the fact that the Gaussian filtering can remove some noisy points that interfere a good localisation of human shape, and the histogram multi-binning can distribute the

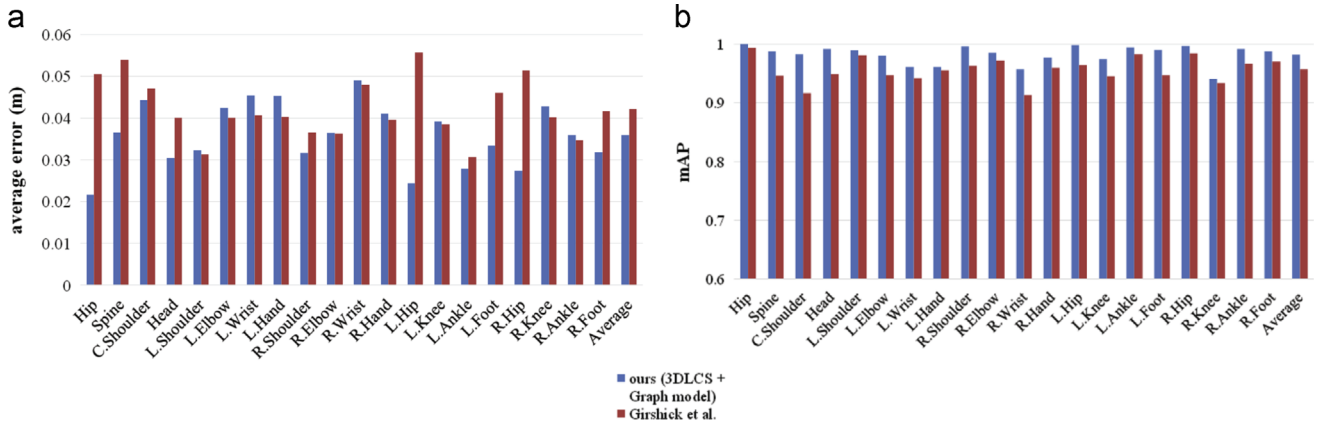


Fig. 7. Performance compared with Girshick et al. [10] on the Stanford dataset: (a) average estimation error vs. body joint; (b) mAP vs. body joint.

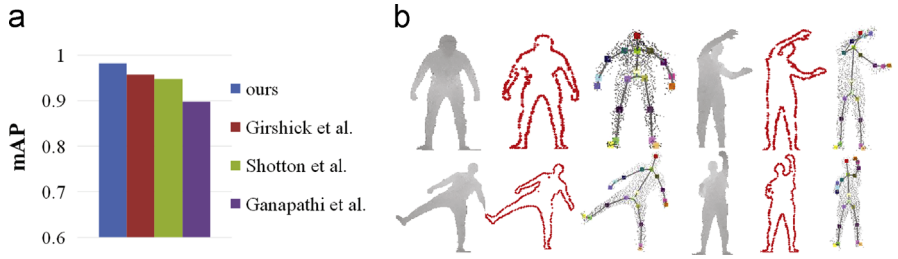


Fig. 8. Performance compared with other methods [5,41] on the Stanford dataset: (a) comparison with two other methods of [5,41]; (b) estimation results of our proposed method.

Table 1
Comparison with Li et al. [13] on the Stanford dataset. Note that we use a different distance tolerance ($\tau = 0.2$ m) to make our evaluation consistent with that of [13].

Method/mAP ($\tau = 0.2$ m)	Hand	Head	Foot
Li et al. [13]	0.870	0.800	0.990
Ours (3DLSC + Graph Model)	**0.997**	**1.000**	**0.997**

Table 2
Comparison of our 3DLSC with 2DSC [36] and DCF [5] on the Stanford dataset.

Method/Accuracy	mAP ($\tau = 0.1$ m)
2DSC	0.876
DCF	0.976
3DLSC	**0.982**

noise across a feature vector and thus weaken the negative influence of the noise.

4.6.2. Number of silhouette points

We retrain regression forests with different numbers N of sampled silhouette points, with N from 100 to 900. The results are shown in Fig. 12. When N is increased from 100 to 400, the performance will be enhanced dramatically: the average error decreases from 5.1 cm to 3.6 cm, and mAP increases from 92.8% to 95.2%. When $N \in [400, 900]$, the accuracy tends to be stable. This indicates that a human pose can be efficiently encoded by a low-dimensional silhouette representation.

4.6.3. Votes normalisation factor

We used the votes normalisation factor $\overline{\rho}$ to make our algorithm invariant to different scales of shapes. In order to investigate the effect of $\overline{\rho}$, we train new regression forests without normalising votes. That is, we use a fixed absolute distance ρ to replace $\overline{\rho}$. Considering that there seems only one person's actions available in the Stanford dataset, we carry out experiments on the THU dataset, which captures four persons' actions. In the experiments three values of ρ (0.3 m, 0.5 m and 0.8 m) were applied and their results, labelled by 'Not Normalised', are shown in Fig. 13, along with the result of our method. The results clearly indicate that our normalisation factor $\overline{\rho}$ can adapt well to different body sizes.

4.6.4. Sizes of subsets S_i' and S_c'

For simplicity, the sizes of subsets S_i' and S_c' are set the same. The results of mAP versus the subset size from 1 to 500 for the Stanford dataset are shown in Fig. 14. We can observe that mAP rises as the size increases from 1 to 100, after which it flattens out. In the experiments, we use 200 as the subset size in our experiments.

4.6.5. Scalability of model

In this section, we discuss the scalability of our model. Theoretically, the basic framework of our model is founded on random forests, which have been proved to be robust for human pose recognition even for samples of large sizes [5,9,10]. To validate this, in our real applications we increase the number of training samples from 3.6 K to 10 K. The results are illustrated in Fig. 15. From Fig. 15, it can be observed that our model can effectively adapt more complex poses as the samples increase.

5. Conclusion and future work

We have proposed a novel approach to human pose estimation from depth images, which significantly outperforms the state-of-the-art methods. Our model combines regression forests and graphical models. It considers the dependence between body joints by using a predefined graphical model. The results have shown that, by employing such a combination, the accuracy for

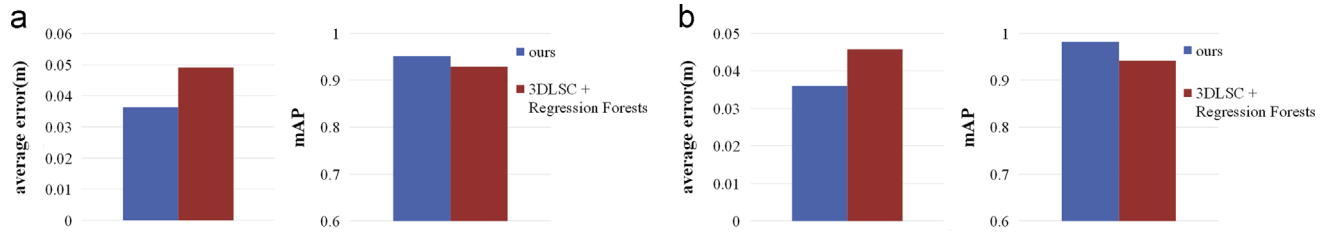


Fig. 9. Performance of the graphical model for (a) the THU dataset and (b) the Stanford dataset. Left-hand panels in (a) and (b): average estimation error vs. model. Right-hand panels in (a) and (b): mAP vs. model.

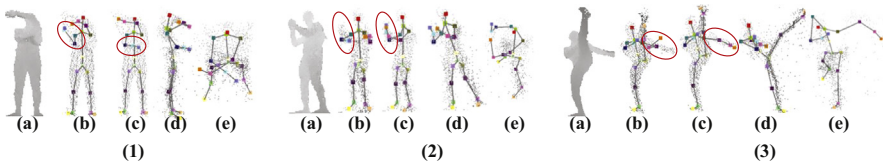


Fig. 10. Effect of the graphical model: (a) the original depth image; (b) results from the method [10]; our results from (c) the front view, (d) the left-side view and (e) the top view.

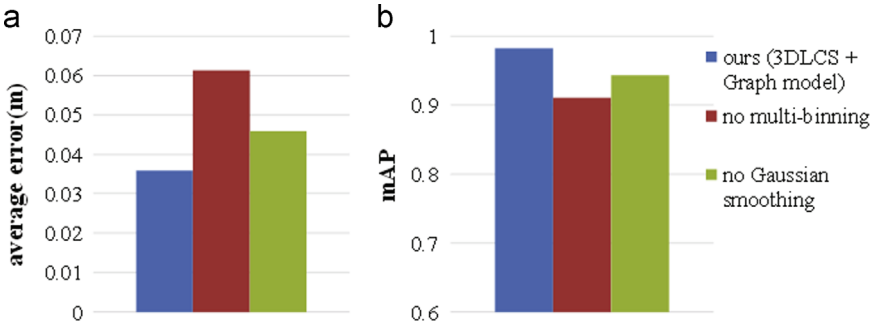


Fig. 11. Effects of Gaussian filtering and histogram multi-binning on pose estimation for (a) average estimation error and (b) mAP.

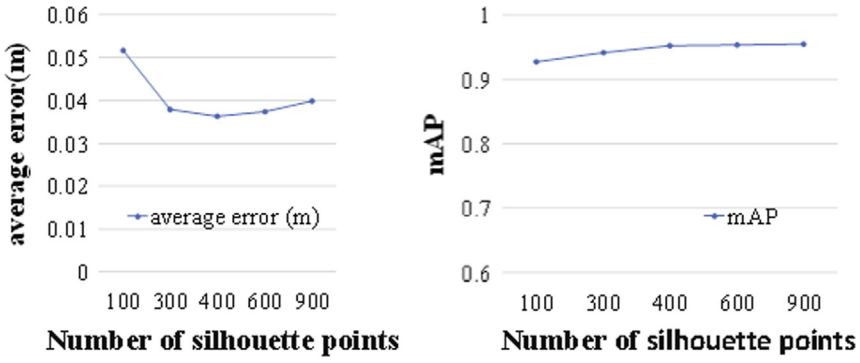


Fig. 12. Effect of the number of silhouette points.

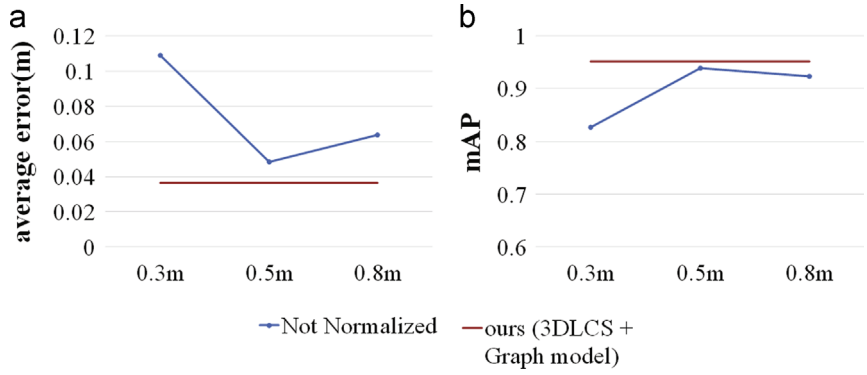


Fig. 13. Effect of the votes normalisation factor $\overline{\rho}$.

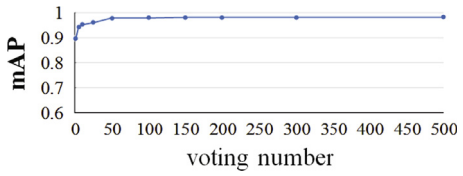


Fig. 14. Effect of the sizes of subsets S_i' and S_c'.

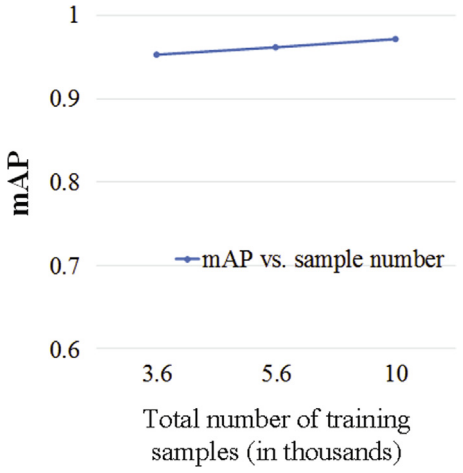


Fig. 15. Effect of the number of training samples.

human pose estimation could be dramatically improved. Furthermore, we have proposed a new 3D local shape feature called 3DLSC, which contains a sequence of histograms in log-polar bins by using 3D silhouette points. The 3DLSC feature has excellent discriminative ability for human poses.

Our experimental results also suggest some research questions, in particular how to combine colour and depth information to offer a more accurate estimation of the dependence among joints. These questions are inspiring our future work.

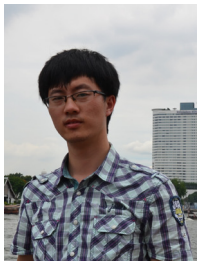
Acknowledgements

This work was partially sponsored by NSFC 61271390, by 863 Project 2015AA016304 and by the Special Foundation for the Development of Strategic Emerging Industries of Shenzhen (No. ZDSYS20140509172959989 & No. YJ20130402145002441).

References

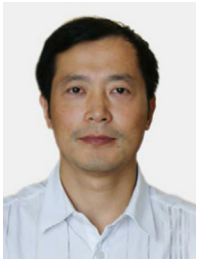
[1] T.B. Moeslund, Visual Analysis of Humans: Looking at People, Springer, 2011.
[2] Z. Hu, G. Wang, X. Lin, H. Yan, Recovery of upper body poses in static images based on joints detection, Pattern Recognit. Lett. 30 (5) (2009) 503–512.
[3] M. Andriluka, S. Roth, B. Schiele, Discriminative appearance models for pictorial structures, Int. J. Comput. Vis. 99 (3) (2012) 259–280.
[4] C. Plagemann, V. Ganapathi, D. Koller, S. Thrun, Real-time identification and localization of body parts from depth images, in: IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2010, pp. 3108–3113.
[5] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, A. Blake, Real-time human pose recognition in parts from single depth images, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2011, pp. 1297–1304.
[6] G. Wang, X. Yin, X. Pei, C. Shi, Depth estimation for speckle projection system using progressive reliable points growing matching, Appl. Opt. 52 (3) (2013) 516–524.
[7] A. Shpunt, Z. Zalevsky, Three-Dimensional Sensing using Speckle Patterns, US Patent App. 12/282,517, March 8 2007.
[8] C. Shi, G. Wang, X. Yin, X. Pei, B. He, X. Lin, High-accuracy stereo matching based on adaptive ground control points, IEEE Trans. Image Process. (2015), in press.
[9] M. Sun, P. Kohli, J. Shotton, Conditional regression forests for human pose estimation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2012, pp. 3394–3401.
[10] R. Girshick, J. Shotton, P. Kohli, A. Criminisi, A. Fitzgibbon, Efficient regression of general-activity human poses from depth images, in: IEEE International Conference on Computer Vision (ICCV), IEEE, 2011, pp. 415–422.
[11] G. Fanelli, M. Dantone, J. Gall, A. Fossati, L. Van Gool, Random forests for real time 3D face analysis, Int. J. Comput. Vis. 101 (3) (2013) 437–458.
[12] K. Buys, C. Cagniart, A. Baksheev, T. De Laet, J. De Schutter, C. Pantofaru, An adaptable system for RGB-D based human body detection and pose estimation, J. Vis. Commun. Image Represent. 25 (1) (2014) 39–52.
[13] Z. Li, D. Kulic, Local shape context based real-time endpoint body part detection and identification from depth images, in: Canadian Conference on Computer and Robot Vision (CRV), IEEE, 2011, pp. 219–226.
[14] A. Baak, M. Müller, G. Bharaj, H.-P. Seidel, C. Theobalt, A data-driven approach for real-time full body pose reconstruction from a depth camera, in: Consumer Depth Cameras for Computer Vision, Springer, 2013, pp. 71–98.
[15] M. Ye, X. Wang, R. Yang, L. Ren, M. Pollefeys, Accurate 3D pose estimation from a single depth image, in: IEEE International Conference on Computer Vision (ICCV), IEEE, 2011, pp. 731–738.
[16] P.F. Felzenszwalb, D.P. Huttenlocher, Pictorial structures for object recognition, Int. J. Comput. Vis. 61 (1) (2005) 55–79.
[17] M. Andriluka, S. Roth, B. Schiele, Pictorial structures revisited: People detection and articulated pose estimation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2009, pp. 1014–1021.
[18] B. Sapp, A. Toshev, B. Taskar, Cascaded models for articulated pose estimation, in: Computer Vision–ECCV 2010, Springer, 2010, pp. 406–420.
[19] Y. Yang, D. Ramanan, Articulated pose estimation with flexible mixtures-of-parts, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2011, pp. 1385–1392.
[20] M. Sun, S. Savarese, Articulated part-based model for joint object detection and pose estimation, in: IEEE International Conference on Computer Vision (ICCV), IEEE, 2011, pp. 723–730.
[21] M. Dantone, J. Gall, C. Leistner, L. Van Gool, Human pose estimation using body parts dependent joint regressors, in: Computer Vision and Pattern Recognition (CVPR), IEEE Conference on, IEEE, 2013, pp. 3041–3048.
[22] X. Ren, A.C. Berg, J. Malik, Recovering human body configurations using pairwise constraints between parts, in: IEEE International Conference on Computer Vision (ICCV), vol. 1, IEEE, 2005, pp. 824–831.
[23] T.-P. Tian, S. Sclaroff, Fast globally optimal 2D human detection with loopy graph models, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2010, pp. 81–88.
[24] M. Sun, M. Telaprolu, H. Lee, S. Savarese, An efficient branch-and-bound algorithm for optimal human pose estimation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2012, pp. 1616–1623.
[25] D. Tran, D. Forsyth, Improved human parsing with a full relational model, in: Computer Vision–ECCV 2010, Springer, 2010, pp. 227–240.
[26] M. Bergtholdt, J. Kappes, S. Schmidt, C. Schnörr, A study of parts-based object class detection using complete graphs, Int. J. Comput. Vis. 87 (1-2) (2010) 93–117.
[27] F. Li, S.-R. Zhou, J.-M. Zhang, D.-Y. Zhang, L.-Y. Xiang, Attribute-based knowledge transfer learning for human pose estimation, Neurocomputing 116 (2013) 301–310.
[28] B. Sapp, C. Jordan, B. Taskar, Adaptive pose priors for pictorial structures, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2010, pp. 422–429.
[29] J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A. Kipman, et al., Efficient human pose estimation from single depth images, IEEE Trans. Pattern Anal. Mach. Intell. 35 (12) (2013) 2821–2840.
[30] T.K. Ho, Random decision forests, in: International Conference on Document Analysis and Recognition, vol. 1, IEEE, 1995, pp. 278–282.
[31] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[32] R.J. Lewis, An introduction to classification and regression tree (CART) analysis, in: Annual Meeting of the Society for Academic Emergency Medicine in San Francisco, California, Citeseer, 2000, pp. 1–14.
[33] A. Liaw, M. Wiener, Classification and regression by random Forest, R News 2 (3) (2002) 18–22.
[34] T.-H. Yu, T.-K. Kim, R. Cipolla, Unconstrained monocular 3D human pose estimation by action detection and cross-modality regression forest, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2013, pp. 3642–3649.
[35] S. Belongie, J. Malik, J. Puzicha, Shape matching and object recognition using shape contexts, IEEE Trans. Pattern Anal. Mach. Intell. 24 (4) (2002) 509–522.
[36] A. Agarwal, B. Triggs, Recovering 3D human pose from monocular images, IEEE Trans. Pattern Anal. Mach. Intell. 28 (1) (2006) 44–58.
[37] C. Ek, P. Torr, N. Lawrence, Gaussian process latent variable models for human pose estimation, in: MLMI'07, 2007, pp. 132–143.
[38] M. Körtgen, G.-J. Park, M. Novotni, R. Klein, 3D shape matching with 3D shape contexts, in: The 7th Central European Seminar on Computer Graphics, vol. 3, 2003, pp. 5–17.
[39] M. Dantone, J. Gall, G. Fanelli, L. Van Gool, Real-time facial feature detection using conditional regression forests, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2012, pp. 2578–2585.

[40] H. Yang, I. Patras, Face parts localization using structured-output regression forests, in: Asian Conference on Computer Vision (ACCV), Springer, 2013, pp. 667–679.
[41] V. Ganapathi, C. Plagemann, D. Koller, S. Thrun, Real time motion capture using a single time-of-flight camera, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2010, pp. 755–762.
[42] Y. Li, G. Wang, X. Lin, G. Cheng, Real-time depth-based segmentation and tracking of multiple objects, in: IEEE Conference on Technology in Automation, Control, and Intelligent Systems (CYBER), IEEE, 2012, pp. 429–433.



Li He received the B.S. degree from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 2010. He is currently working toward the Ph.D. degree in the Department of Electronic Engineering, Tsinghua University, Beijing, China. His research interests include the applications of machine learning and pattern recognition in human pose/action recognition and tracking.

Guijin Wang was born in 1976. He received the B.S. and Ph.D. degrees (with honor) from the Department of Electronic Engineering, Tsinghua University, China in 1998 and 2003 respectively, all in Signal and Information Processing. From 2003 to 2006, he has been with Sony Information Technologies Laboratories as a researcher. From Oct., 2006, he has been with the Department of Electronic Engineering, Tsinghua University, China as an Associate Professor. He has published over 80 international journal and conference papers, and held several patents. His research interests are focused on wireless multimedia, image and video processing, depth imaging, pose recognition, intelligent surveillance, industry inspection, object detection and tracking, online learning, etc.



Qingmin Liao received the B.S. degree in radio technology from the University of Electronic Science and Technology of China, Chengdu, China, in 1984, and the M.S. and Ph.D. degrees in Signal Processing and Telecommunications from the University of Rennes 1, Rennes, France, in 1990 and 1994, respectively. Since 1995, he has been with Tsinghua University, Beijing, China. He became a Professor in the Department of Electronic Engineering of Tsinghua University, in 2002. From 2001 to 2003, he served as the Invited Professor with a tri-year contract at the University of Caen, France. Since 2010, he has been the Director of the Division of Information Science and Technology in the Graduate School at Shenzhen, Tsinghua University, Shenzhen, China. His research interests include image/video processing, transmission and analysis; biometrics; and their applications to teledetection, medicine, industry, and sports. He has published over 90 papers internationally.

Jing-Hao Xue received the B.Eng. degree in Telecommunication and Information Systems in 1993 and the Dr.Eng. degree in Signal and Information Processing in 1998, both from Tsinghua University, the M.Sc. degree in Medical Imaging and the M.Sc. degree in Statistics, both from Katholieke Universiteit Leuven in 2004, and the degree of Ph.D. in Statistics from the University of Glasgow in 2008. Since 2008, he has worked in the Department of Statistical Science at University College London, as a Lecturer and Senior Lecturer. His current research interests include statistical classification, high-dimensional data analysis, computer vision, and pattern recognition.