IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 27, 2016, accepted December 7, 2016, date of publication December 29, 2016, date of current version March 2, 2017.

Digital Object Identifier 10.1109/ACCESS.2016.2643439

Human Pose Estimation by Exploiting Spatial and Temporal Constraints in Body-Part Configurations

QINGWU LI, FEIJIA HE, TIAN WANG, LIANGJI ZHOU, AND SHUYA XI
Key Laboratory of Sensor Networks and Environmental Sensing, Hohai University, Changzhou 213022, China

Corresponding author: Q. Li (li_qingwu@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 41306089 and in part by the Science and Technology Program of Jiangsu Province under Grant BY2014041.

ABSTRACT We present an algorithm for estimating a sequence of articulated upper-body human pose in unconstrained videos. Most previous work often fails to locate forearms in those video scenes suffering from illumination varieties, background clutter, camera shake, or occlusion. In order to deal with such intractable cases, we propose a novel algorithm for addressing the problem of certain body parts localization. The proposed approach can be roughly divided into two steps: first, a spatial model is designed to capture the high-order relationship between adjacent parts and meanwhile to generate a set of configurations in each frame under the temporal context constraint; second, a competitive method is presented to select the best body parts among diverse pose configurations. In this paper, the proposed algorithm focuses on the unconstrained video scenes and improves the detection precision of certain body parts with high degree of freedom. Moreover, the proposed algorithm can be well applied to a very challenging dataset named Movies. Experimental results show that the proposed algorithm can dramatically improve performance compared with those related algorithms on two benchmark datasets (MPII and SHPED datasets) and on our Movies dataset.

INDEX TERMS Pose estimation, object detection, motion detection.

I. INTRODUCTION

Articulated human pose estimation plays a vital role in computer vision area, such as human action recognition, character animation and image understanding. The target of pose estimation is to detect the positions of human body parts in images or videos. However, there are no versatile yet efficient solutions available for all scenarios. Several factors make this task challenging, such as the variability of human appearances, illumination varieties over a diverse palette of lighting conditions and camera viewpoint, background clutter, and occlusion. An influential approach for articulated human pose estimation using part-based models is proposed in [1], which has successfully captured the pairwise spatial relationships between the parts. However, the method in [1] performed poorly on certain body parts such as forearm.

Using original pictorial structure (PS) [2], Felzenszwalb *et al.* [3] proposed a simple appearance model for human body while requiring background elimination. This structure ties body parts together based on the conditional random field (CRF) leading to tractable and efficient inference. Recently, pose estimation methods [4] employed more complex appearance models based on this structure. Many methods, such as [1] and [2], [5]–[8], decompose the entire body into rigid parts. Templates of the body parts will be obtained by learning from the large-scale dataset. Conditional constraints between body parts are generally employed to strengthen the model [10]–[14]. Although conditional constraints can help to strengthen the model, they may break the tree structure of different body parts. To deal with this problem, approximate inference methods such as sampling [10], [12], [13] are also presented.

In our approach, spatial and temporal context is utilized to restrain certain body parts. A set of plausible configurations is captured (pose set) in each frame under the restraints of both temporal context and spatial condition. The precision of complex body parts detection is improved in our process. By combining the method proposed by Dennis Park [15], the proposed approach ensures that the configurations perform best but do not overlap. However, some body parts are estimated best in this configuration while

others in another one. Therefore, how to ensure that each component in the first best configuration performs best is a problem. An approach is then proposed to select an optimal configuration of human pose from the pose set. This strategy can be briefly described as that we first break the N-best configurations into body parts and then sample best ones from the first node to the end node. By using the proposed strategy, the final human pose will be an optimal configuration with the best performance.

In terms of numerical evaluation, our model outperforms the similar methods on two benchmark datasets (MPII dataset and SHPED dataset) and a novel dataset addressed in this paper named Movies.

A. RELATED WORK

There exists a large body of works [7], [16]–[18] tackling the problem of human pose estimation in recent years. Early methods [19], [20] focused on human pose estimation in controlled environment. Although human pose estimation algorithms [1], [21]–[24] in unconstrained scenes have progressed in the last few years, they can only be applied to single images. Video based human pose estimation [15], [25]–[27] in unconstrained scenes is still an open problem.

There has been a recent thrust in methods that aim to estimate human pose in a single image. Methods [1], [5], [18], [28], [29] based on PS describe the human body as a tree structured graphical model with kinematic priors of connected limbs. These methods have performed well in the case of visible limbs. However, human body parts are too flexible and diverse to be localized, which brings considerable difficulties and challenges to the research work. In order to explore and utilize the higher dependence relationship among components, many methods [1], [29]–[35] have been proposed. The flexible mixture-of-parts model proposed by Yang *et al.* [1] described human body parts as a mixture of multiple models. An exponential number of poses can be represented by using the mixed models. The strategy of using hierarchical model to represent the relationship among body parts has been presented in [30] and [31]. On this basis, [34] has given a variety of hybrid types to each hidden node, so that it has a more ''human-like'' expression for human pose estimation.

In the video domain, researches focus on the methods of establishing a reasonable time link between the neighbor frames and generating a continuous, smooth poses over the time. Most of them heading for two directions, respectively, articulated motion parsing [10], [36]–[38] and tracking-by-detections [15], [25], [26]. For the latter, the algorithms can be roughly divided into two steps: firstly, they estimated the human pose in each frame, such as proposed in [15] and [25] to generate a number of different candidates for each frame and smooth the pose space across the entire video sequences; secondly, to find the temporal information in the sequences, [26] and [27] addressed the model to obtain the correlation over the time. Ramakrishna *et al.* [26] modeled the symmetric structures of the body parts and proposed an effective approximate solution to the problem. In [15], many pose candidates have been generated in each frame, and the most consistent ones that possess high detection scores were selected. Zuffi *et al.* [38] proposed a new way of integrating information over the time to complete human pose estimation in video.

B. CONTRIBUTION

Two contributions are made in this paper. Firstly, a spatial model is designed to capture the high-order relationship between adjacent parts and meanwhile to generate a set of configurations in each frame under the temporal context constraintwhich is shown in Fig. 1(a); secondly, a competitive method is presented to select the best body parts among diverse pose configurations, which is shown in Fig. 1(b).

The remainder of this paper is organized as follows. The proposed approach is introduced in Section II: the mixture-of-parts model is described in Section II.A; spatial constraint and temporal context are introduced respectively in SectionII.B and II.C; Section II.D shows how to select the optimal pose from configurations. In Section III, we show experimental results and diagnostic experiments on the benchmark datasets. Section IV concludes this paper.

II. MODEL

With a reference to the major steps in Fig. 1, the main steps of our method are: a) for each frame, obtain a set of configurations of human body under the double restraints of temporal context and spatial condition (Fig. 1(a)); b) select an optimal configuration of human pose from the pose set using the strategy that we break the N-best configurations into body parts and then sample best ones successively (Fig. 1(b)).

A. MIXTURE-OF-PARTS MODEL

We present an improved model on the basis of the mixture-of-parts model proposed by Yang *et al.* [1] for its good performance and computation efficiency. This model was first defined in [39] In 2011 and then published in the IEEE Transactions on Pattern Analysis and Machine Intelligence in 2013. However, such an approach often makes errors in the video sequences with cluttered background, motion blur, occlusion or self-occlusion. To solve these problems in the video-based body pose estimation, we introduce a more effective model representation. Previous to the details of our approach, we briefly present the original model.

Supposing that the human body is divided into K parts, I represents an image and $z_i = (x_i, y_i)$ represents the location of part i, thus $z = \{z_1, z_2, \ldots, z_k\}$ represents the human pose in I. According to the tree-structured graphical model, the human body can be described as an undirected graph $G = (V, E)$, representing the structure of a human pose. Each vertex V corresponds to a body part and each edge E represents a relational constraints between pairs of parts. Moreover, the human pose estimation is actually a process of locating each component of the human body, the core of which is to estimate the rationality among the configurations

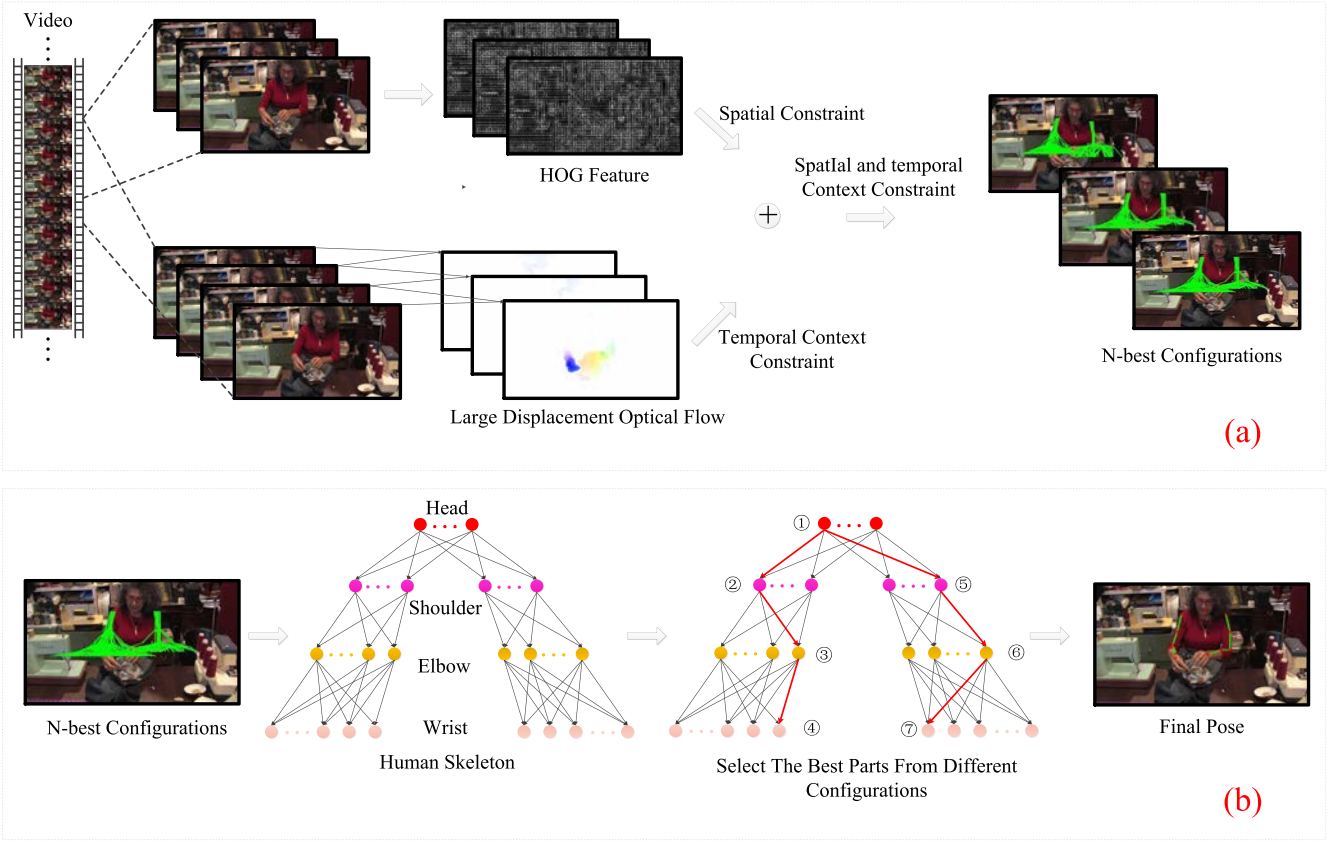


FIGURE 1. An outline of the proposed method. (a) Shows the process of generating a set of configurations based on the spatial and temporal context; in (b), the final human pose will be determined using the strategy that we break the N-best configurations into body parts and then reselect each optimal one.

of parts. Therefore, the above problem is indeed transferred into the calculation of scores for multiple configurations in an image. The single-image pose estimation problem is then formulated as the maximization of the following score function $score\,(I, z)$:

$$score(I, z) = \sum_{i \in V} \varphi_i(z_i) + \sum_{ij \in E} \psi_{ij}(z_i, z_j) \tag{1}$$

where, $\phi_i\,(z_i)$ is an appearance feature term for body part I at the position z_i in image I, which is described with a feature vector (e.g. HOG descriptor [40]), and $\psi_{ij}(z_i, z_j)$, compared to a ''spring'', is a pairwise deformation cost between parts (i, j). The first term is used to measure the degree of matching when part i is placed at location z_i in the image, whereas the second term depicts the rational distribution of human body parts in space, and it embeds particular kinematic constraints between pairwise parts to ensure that the parts satisfy the actual conditions. Both the part model and the kinematic constraints between pairwise of them can be learned using a structured SVM formulation.

B. SPATIAL CONSTRAINT IN POSES

Mixture-of-parts model proposed in [1] has efficient and tractable inference by restraining the spatial relationship between neighbor parts and maintaining a tree structure. However, small deformation in templates makes a ''human-unlike'' configuration. To solve this problem, high-order spatial constraint is utilized in our model. The multi-layer structure shown in Fig. 2 is composed of body parts and the combination of them. Each node except the root node (body) has a parent. All nodes without children are the original human body parts, and the overall model follows a tree structure.

In this work, we focus on human upper body only, as they convey the majority of information necessary to recognize the actions. Moreover, videos and films usually consist of close-up or medium-shots where legs stay outside the visible frame [7]. Besides our model can be easily extended to the entire body pose estimation. Lets take $l = latent\,(i)$ for the parents of the node i in our multi-layer structure (Note that the parents here are not the ones in the tree structure), and take type b_l, b_i for the mixture component of part l, i respectively. Such information is encoded in the spatial constraint term $\theta_{li}\,(b_l, b_i)$, which is a function between the parent type and the child type. For instance, a vertical arm cannot have a horizontal lower arm, so we set $\theta_{li}\,(b_l, b_i)$ in this case to be the negative infinity. The score function Eq.(1) is optimized into another form as follows.

$$Score_1\,(I, z) = Score\,(I, z) + \sum_{l=latent\,(i)} \sum_{i} \theta_{lj}\,(b_l, b_i) \tag{2}$$

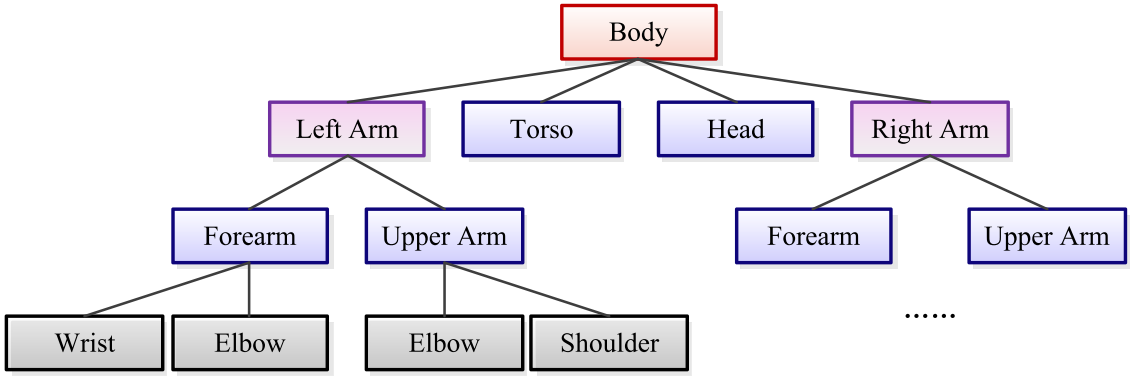


FIGURE 2. The multi-layer structure for pose estimation. All nodes in black are original human body parts, and the others are latent nodes.

High-order spatial constraint makes human pose more natural and "human-like". For further details, we refer the reader to [34].

C. MODEL BASED ON TEMPORAL CONTEXT

For pose estimation, dealing with video sequence is more valuable and significant than dealing with single image. Stable kinematical constraints and coincident geometric characteristics of body parts over the time implement a more effective estimation performance. We propose a model based on temporal context aiming at the human pose estimation in the videos, which has excellent performance on the difficult parts.

Given video sequence $Vs = \left(I^{(1)}, I^{(2)}, \ldots, I^{(T)}\right)$, in which $I^{(t)}$ and $I^{(t+1)}$ represent two neighbor frames, let us take $z^{(Vs)} = \left\{z^{(1)}, z^{(2)}, \ldots, z^{(T)}\right\}$ for a set of human poses in the video sequence Vs . A revised score function for a single frame $I^{(t)}$ based on Eq.(2) using temporal context is defined as:

$$\begin{aligned} &Score_2\left(I^{(t)}, z^{(t)}\right) \\ &\quad = Score_1\left(I^{(t)}, z^{(t)}\right) + Score'_1\left(I^{(t+1)}, \tilde{z}^{(t+1)}\right) \\ &\qquad + \varepsilon'\lambda\left(I^{(t)}, I^{(t+1)}, \tilde{z}^{(t)}, \tilde{z}^{(t+1)}\right) \end{aligned} \tag{3}$$

where, $Score_1\left(I^{(t)}, z^{(t)}\right)$ is defined in Section II.B, ε' is the normalized parameter (note that the transformations of parameter ε below are normalized parameters.), and the term of $Score'_1\left(I^{(t+1)}, \tilde{z}^{(t+1)}\right)$ contains the scores of several nodes only and is represented as:

$$\begin{aligned} Score'_1(I, \tilde{z}) = &\sum_i \phi_i(z_i) + \sum_{ij\in E} \psi_{ij}\left(z_i, z_j\right) \\ &+ \sum_{l=latent(i)} \sum_i \theta_{lj}\left(b_l, b_i\right) \\ &s.t.\ i \in \{elbow, wrist\} \end{aligned} \tag{4}$$

In Eq.(4), we only need to calculate the appearance feature term and the deformation cost of elbows and wrists, as the relation of human body in two neighbor frames is mainly reflected in these two parts.

Back to Eq.(3), $\lambda\left(I^{(t)}, I^{(t+1)}, \tilde{z}^{(t)}, \tilde{z}^{(t+1)}\right)$ is a term for describing a distance between $\tilde{z}^{(t)}$ in $I^{(t)}$ and $\tilde{z}^{(t+1)}$ in $I^{(t+1)}$. We compute it as follows:

$$\begin{aligned} &\lambda\left(I^{(t)}, I^{(t+1)}, \tilde{z}^{(t)}, \tilde{z}^{(t+1)}\right) \\ &\quad = \sum_{i\in\{elbow,wrist\}} dist\left(z_i^{(t+1)}, z_i^{(t)}, Ldof\left(z_i^{(t)}\right)\right) \end{aligned} \tag{5}$$

where, $dist\ (*)$ is any distance transform and Euclidean distance is specified here. $Ldof\left(z_i^{(t)}\right)$ is denoted as an optical flow term proposed by Thomas in [41]. It can be obviously seen in Fig. 3 that forearm gets a high flow term.

Eq.(3) with respect to Eq. (1) increases the temporal and spatial constraints of human body, and improves the accuracy of difficult parts detection significantly such as wrist in video sequences.

D. OPTIMAL SELECTION FROM CONFIGURATIONS

In order to obtain a pose set in each frame which consists of n maximally scoring and non overlapping configurations, we use the Eq. (3) and N-best algorithm [15]. Note that how to select an optimal pose from the set by maximizing the following function is the other contribution of our method.

$$\begin{aligned} &\max_{z^{(t)}\in Z^{(t)},\forall t} Score_3\left(Vs, z^{(Vs)}\right) \\ &= \max_{z^{(t)}\in Z^{(t)},\forall t} Score_1\left(I^{(T)}, z^{(T)}\right) \\ &\quad + \sum_{t=1}^{T-1} Score_1\left(I^{(t)}, z^{(t)}\right) + \varepsilon_1\lambda\left(I^{(t)}, I^{(t+1)}, z^{(t)}, z^{(t+1)}\right) \end{aligned} \tag{6}$$

We take $Score_3\left(Vs, z^{(Vs)}\right)$ for the score function of the human pose estimation in a video Vs , where $Z^{(t)}$ denotes the set of pose (configurations) in the frame $I^{(t)}$ which is calculated by Eq. (3). What we need is to select $z^{(t)}$ from $Z^{(t)}$ and it is the optimal matching for the frame $I^{(t)}$. It is obvious that the construction of the set $Z^{(t)}$ will directly affect the results of the pose estimation. We are not sure that an optimal pose $z^{(t)}$ can be found from the set $Z^{(t)}$ in every frame. Eq. (3)

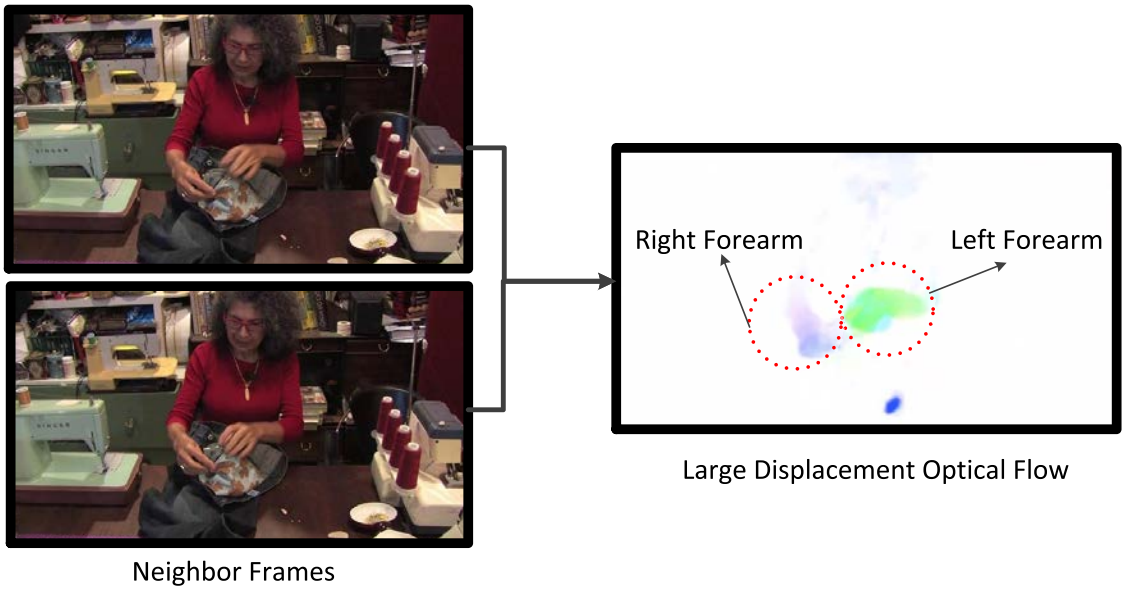


FIGURE 3. The multi-layer structure for pose estimation. All nodes in black are original human body parts, and the others are latent nodes.

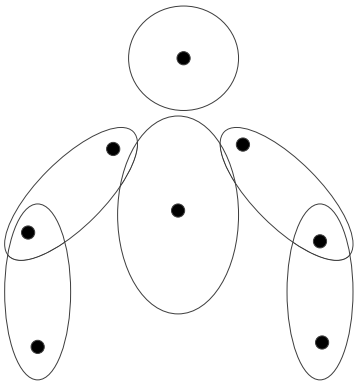


FIGURE 4. The structure of human body skeleton. The nodes are latent nodes painted blue in Figure 2; and the ellipses are the representation of human skeleton.

FIGURE 5. The examples of Movies dataset. (a) illumination varieties. (b) background clutter. (c) occlusion.

based on temporal context completes the task to generate an excellent pose set, but it does not work unless there is an optimal solution in the set.

Usually there are only sectional body parts fitting well in a certain configuration obtained by Eq. (3). In this work, we propose a new strategy to select the optimal parts from the different poses $z^{(t)}$ in set $Z^{(t)}$.

In this strategy, we represent the human body as a form of multiple latent nodes or their children, which are painted blue, showed in Fig. 2. More visually, the human body is represented as in Fig. 4. Our goal is to pick the best skeleton from the components which are split from the pose set $Z^{(t)}$, and then combine them into a human pose. Next, we will give more details and the visual description is given in Fig. 1(b).

Firstly, we take skeleton as $z_{i,j} = (z_i, z_j)$, where $(i, j) \in E$, and then calculate the score of all the skeletons across the whole video sequences as follows:

$$\begin{aligned} &Score'_3\left(Vs, z_{i,j}^{(Vs)}\right) \\ &= Score_1\left(I^{(T)}, z_{i,j}^{(t)}\right) + \sum_{t=1}^{T-1} Score_1\left(I^{(t)}, z_{i,j}^{(t)}\right) \\ &\quad + \varepsilon_1 \lambda\left(I^{(t)}, I^{(t+1)}, z_{i,j}^{(t)}, z_{i,j}^{(t+1)}\right) \end{aligned} \tag{7}$$

Compared with formula (6), we only keep the term $z_j^{(t)}$ and the relationship between $z_i^{(t)}$ and $z_j^{(t)}$ in this score function.

Secondly, we constrain the combination of the skeleton. Skeleton $z_{i,j} = (z_i, z_j)$ in pose z can stick with the skeleton $z_{j,n} = (z'_j, z'_n)$ in pose z' , if they both have a same part j. We define a distance to describe the difference between the recombined skeletons as follows, which can be understood as

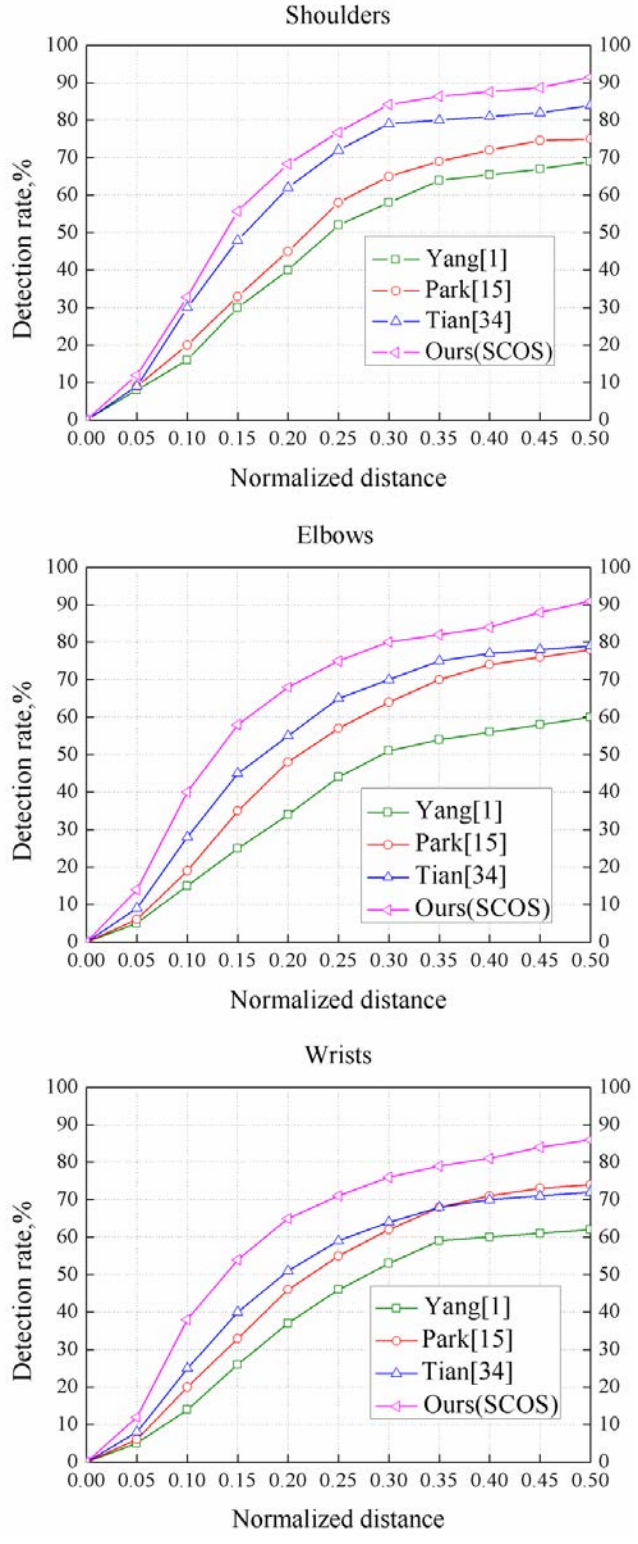


FIGURE 6. Comparison PCkh accuracy of different approaches on MPII dataset.

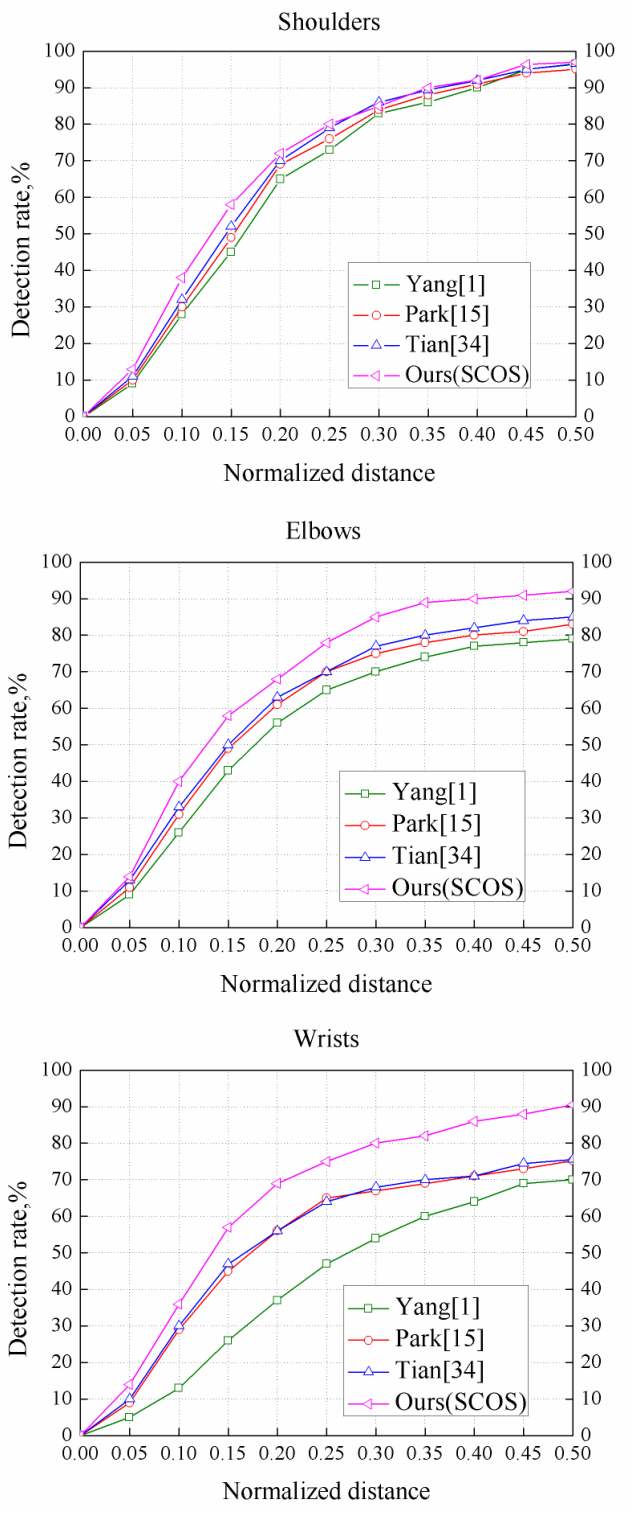


FIGURE 7. Comparison PCkh accuracy of different approaches on SHPED dataset.

a spring.

$$dist\left(z_{i,j}, z_{j,n}\right) = \varepsilon_2 \left\| z_i - z'_n \right\|_2^2 \tag{8}$$

Thirdly, we recombine the skeletons from different poses into a new body, and steps are as follows: human head is located first, because it is the easiest part to be accurately positioned among all the body parts. We find the head by calculating the maximum value of the score function $Score'_3\left(Vs, z_{h,h}^{(Vs)}\right)$, which is defined in Eq. (7). And then the skeleton linked with head is estimated. After that, we locate

TABLE 1. Comparison results with the similar methods and components of the proposed method, on three datasets (MPII, SHPED and Movies, respectively).

MPII dataset				
	Method	Head	Upper arm	Forearm
PCP	Yang et al. [1]	80.4	40.1	35.1
	Park et al. [15]	84.1	42.5	37.8
	Tian et al. [34]	86.4	50.4	50.2
	Ours (SC)	90.5	75.4	80.4
	Ours (SCOS)	**91.4**	**80.5**	**86.5**
SHPED dataset				
	Method	Head	Upper arm	Forearm
PCP	Yang et al. [1]	96.8	90.5	50.4
	Park et al. [15]	95.9	92.4	53.6
	Tian et al. [34]	96.0	91.4	60.5
	Ours (SC)	96.1	88.5	85.4
	Ours (SCOS)	**96.7**	**92.7**	**91.2**
Movies dataset				
	Method	Head	Upper arm	Forearm
PCP	Yang et al. [1]	62.9	50.1	41.8
	Park et al. [15]	63.5	52.6	42.0
	Tian et al. [34]	62.1	55.4	53.3
	Ours (SC)	68.5	62.4	61.2
	Ours (SCOS)	**71.1**	**68.7**	**64.6**

the other skeleton joined with the aforesaid one successively. In this process, the score function as follows needs to be maximized.

$$Score'_4\left(Vs, z_{i,j}^{(Vs)}\right) = Score'_3\left(Vs, z_{i,j}^{(Vs)}\right) + dist\left(z_{p(i),i}^{(Vs)}, z_{i,j}^{(Vs)}\right) \quad (9)$$

where $Score'_3(*)$ and $dist(*)$ are defined in Eq. (7) and (8) respectively, and $p(i)$ represents the parent node of the node i in the tree structure model. We complete the human pose estimation by selecting the optimal skeleton from head to arm successively. In this strategy, each one connected with the next skeleton can be obtained by maximizing the score function. In this way, it can be guaranteed that every skeleton selected from the pose set is the most reasonable one.

III. EXPERIMENTS

A. DATASET

In order to verify the effectiveness of our human pose estimation algorithm in this paper, we will first evaluate it on two public datasets, named MPII and SHPED datasets, and a new dataset introduced in this paper named Movies. Details of the three datasets are given in the following part.

MPII: This dataset [42] is a state-of-the-art benchmark for evaluation of articulated human pose estimation. Each image was extracted from a YouTube video and provided with preceding and following un-annotated frames. In this paper, we will take 30 video sequences from MPII dataset, which only contain a single person, and each sequence is composed of 41 frames.

SHPED: This dataset [43] contains 1260 images which are grouped into 42 video clips of 15 frames each. The clips have been extracted from 26 videos obtained from the popular video-sharing website YouTube3

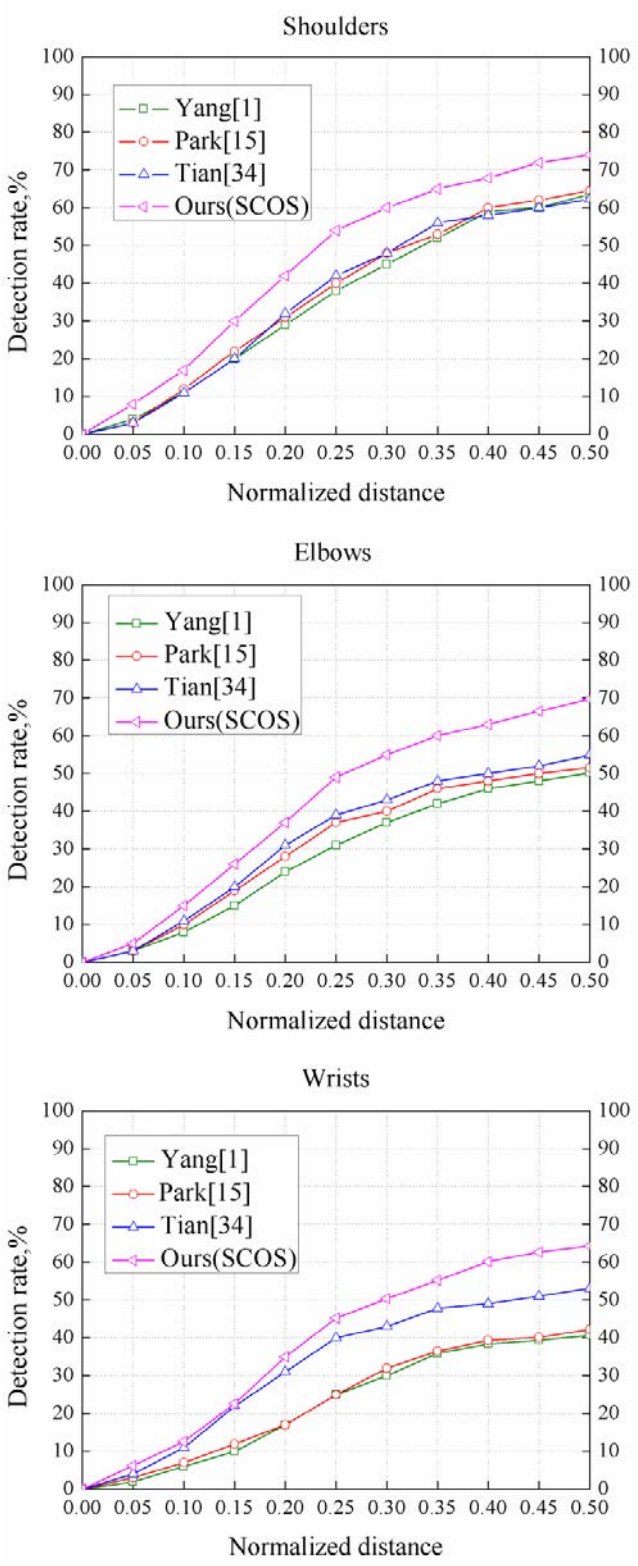


FIGURE 8. Comparison PCkh accuracy of different approaches on Movies dataset.

Movies: In this paper, we introduce a challenging dataset named Movies, which contains 40 video sequences of 15 frames each from four movies. In contrast to the datasets

FIGURE 9. Qualitative comparisons of different approaches on MPII dataset.

above, it contains more challenging scenes, which are shown in Fig. 5, with illumination varieties (Fig. 5(a)), background clutter(Fig. 5(b)), and occlusion(Fig. 5(c)).

B. EVALUATION METRIC

In this section, we describe two metrics for evaluating pose estimation.

PCP: Ferrari *et al.* [10] describe a broadly adopted evaluation metric (PCP), which measures the percentage of correctly localized body parts. A candidate body part is labeled as correct if its segment endpoints lie within 50 percent of the length of the ground-truth annotated endpoints. This metric was clearly crucial and influential in spurring quantitative evaluation, thus considerably moving the field forward.

PCKh: Similar to [42], we also use PCKh to evaluate all the results. PCKh is a standard evaluation metric, which measures the percentage of correctly localized body parts with a threshold. This metric is calculated by the following function:

$$\sqrt{\left(x_i - x_i^0\right)^2 + \left(y_i - y_i^0\right)^2} \leq ND * len_i \tag{10}$$

where, (x_i, y_i) is the estimated position coordinates of the body parts, $\left(x_i^0, y_i^0\right)$ is the ground truth, len_i is the head segment length, and normalized distance $ND \in [0, 0.5]$ is a parameter for controlling the threshold. If ND is equal to 0.5, $ND * len_i$ is called the standard threshold which is 50% of the head segment length. We choose to use head size because we would like to make the metric independent.

C. EVALUATION OF MODEL COMPONENTS

In Table 1, we show detailed results to further analyze the contributions of each step of our approach. Our (SC) is the proposed method which refines the human pose in a frame using the spatial and temporal context model; Our (SCOC) shows the result by using spatial and temporal context model and the strategy of optimal selection. As we can see from the PCP accuracy results in Table 1, SC does not necessarily improve the results, which is due to the fact that we are not sure each part in the 1-best configuration is optimal. In contrast, SCOC, as the final method, performs well in various datasets. For example, on the MPII dataset (Table 1.), SCOC improves upper arm localization by 5% PCP and forearm localization by 6% PCP compared with SC. On the more challenging Movies dataset (Table 1.), SCOC also shows about 5% PCP improvement on the upper arm and forearm compared with SC.

D. COMPARISON RESULTS WITH SIMILAR ALGORITHMS

We compare the proposed model with three similar human pose estimation algorithms for video sequences:

FIGURE 10. Qualitative comparisons of different approaches on SHPED dataset.

Yang *et al.* [1], Park *et al.* [15], and Tian *et al.* [34]. All these three algorithms and our method are based on the deformable part models. Yang *et al.* [1] first defined the original mixture-of-parts model for pose estimation in [39], which was late published in the IEEE Transactions on Pattern Analysis and Machine Intelligence in 2013. Park *et al.* [15] generated N-best configurations from mixture-of-parts model, and ensured that they do not overlap. More recently, Tian *et al.* [34] proposed a multi-layer structure, which made the combination of body parts more reasonable.

On the three available benchmark datasets, we compare our method with [1], [15], and [34] using PCP and PCKh metric. Quantitative results are given in Table 1, Fig. 6, Fig. 7 and Fig. 8, and qualitative results are shown in Fig. 9, Fig. 10 and Fig. 11.

In Table 1, it can be seen that all the methods can well detect the head, but perform poorly on the forearm. The method [34] has better detection accuracy than the other two algorithms [1], [15] for the forearms. This is due to the multi-layer structure which makes the detected human body more ‘‘human-like. Furthermore, The results show that our complete model (SCOS) outperforms the method [34] by nearly 30% PCP on the MPII dataset for estimating upper arm, and 36% PCP for forearm. On the SHPED dataset, SCOS shows 1.3% PCP (upper arm), 30% PCP (forearm) and 0.3% PCP (upper arm), 37% PCP (forearm) improvement over the methods in [15] and [34] respectively. In the more challenging case of the Movies dataset, the accuracy of all methods is reduced drastically. Even so, SCOC shows 12% PCP (upper arm), 13% PCP (forearm) improvement over the method [34]. In order to evaluate the accuracy of four methods more accurately, we also use the PCKh evaluation metric. In Fig. 6, Fig. 7 and Fig. 8, various methods are evaluated with our complete model on three datasets using PCKh metric. When is equal to 0.5, as we can see in Fig. 6 and Fig. 7: On the MPII dataset, SCOC surpasses the method [34] by almost 10% PCKh for estimating elbows, and 13% PCKh for wrists; on the SHPED dataset, SCOS shows 9% PCKh (elbows), 14% PCKh (wrists) improvement over [34]. In Fig. 8, we show the results of the Movies dataset, in which our complete model shows about 16% PCKh and 11% PCKh improvement on the elbows and wrists compared with the method in [34].

Sometimes the results are not fully reflected by numbers. Therefore, we show detailed results for some consecutive frames, and qualitative comparisons of our approach with [1], [15], and [34] for the three datasets are shown in Fig. 9, Fig. 10 and Fig. 11. It is obvious that our complete model performs much better (please note the forearm, in particular), and the poses are more consistent and all the body parts are located quite accurately. It should be noticed that such big improvement cannot be found in

FIGURE 11. Qualitative comparisons of different approaches on Movies dataset.

PCP accuracy, because the PCKh results are exactly the same.

We summarize the results of our complete model (SCOC), our partial model (SC), and three similar methods [1], [15], [34], based on PCP metric, for shoulders, upper arm and forearm in Table 1. Fig. 6, Fig. 7 and Fig. 8 show the comparison between SCOC with three similar methods [1], [15], [34] on three benchmark datasets using PCKh metric. Qualitative comparison between our method and the method [1], [15], [34] on the three datasets are shown Fig. 9, Fig. 10 and Fig. 11. As we can see, our complete model performs much better in localizing forearm (elbows and wrists). In addition, SCOC performs better when dealing with the more challenging Movies dataset.

E. COMPUTATION TIME

We conducted all the experiments on a laptop with Intel Core i7-4710HQ CPU at 2.5GHz and 8GB RAM. On average, to process one frame, the Matlab2012a implementation took about 3s to generate the set of body configurations and 0.5s to select the final optimal pose.

IV. CONCLUSION

In this paper we proposed a novel algorithm for human pose estimation in videos which achieved the state-of-the-art accuracy at an acceptable computational cost. The main steps of our approach are: 1) capture a set of plausible poses in each frame based on the spatial and temporal context; 2) select an optimal configuration of human pose from the pose set. Besides, we present a more challenging dataset, Movies, containing more complex scenes. Experimental results show that the proposed algorithm improved the accuracy compared with similar methods.

REFERENCES

[1] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2878–2890, Dec. 2013.

[2] M. A. Fischler and R. A. Elschlager, "The representation and matching of pictorial structures," *IEEE Trans. Comput.*, vol. 22, no. 1, pp. 67–92, Jan. 1973.

[3] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient matching of pictorial structures," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2. Jun. 2000, pp. 66–73.

[4] P. Buehler *et al.*, "Long term arm and hand tracking for continuous sign language TV broadcasts," in *Proc. 19th Brit. Mach. Vis. Conf.* 2008, pp. 1105–1114.

[5] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 1014–1021.

[6] A. O. Balan and M. J. Black, "An adaptive appearance model approach for model-based articulated object tracking," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1. Jun. 2006, pp. 758–765.

[7] M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari, "2D articulated human pose estimation and retrieval in (almost) unconstrained still images," *Int. J. Comput. Vis.*, vol. 99, no. 2, pp. 190–214, Sep. 2012.

[8] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *Int. J. Comput. Vis.*, vol. 61, no. 1, pp. 55–79, Jan. 2005.

[9] M. W. Lee and R. Nevatia, "Human pose tracking in monocular sequence using multilevel structured models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 27–38, 2009.

[10] V. Ferrari, M. Marin-Jimenez, and A. Zisserman, "Progressive search space reduction for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2008, pp. 1–8.

[11] B. Sapp, D. Weiss, and B. Taskar, "Parsing human motion with stretchable models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 1281–1288.

[12] C. Sminchisescu and B. Triggs, "Estimating articulated human motion with covariance scaled sampling," *Int. J. Robot. Res.*, vol. 22, no. 6, pp. 371–391, 2003.

[13] J. M. Wang, D. J. Fleet, and A. Hertzmann, "Gaussian process dynamical models for human motion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 283–298, Feb. 2008.

[14] D. Weiss, B. Sapp, and B. Taskar, "Sidestepping intractable inference with structured ensemble cascades," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 2415–2423.

[15] D. Park and D. Ramanan, "N-best maximal decoders for part models," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2627–2634.

[16] G. Gkioxari, P. Arbelaez, L. Bourdev, and J. Malik, "Articulated pose estimation using discriminative armlet classifiers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3342–3349.

[17] G. Mori and J. Malik, "Estimating human body configurations using shape context matching," in *Proc. Eur. Conf. Comput. Vis.*, 2002, pp. 666–680.

[18] D. Ramanan, "Learning to parse images of articulated bodies," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 1129–1136.

[19] K. Rohr, "Towards model-based recognition of human movements in image sequences," *CVGIP, Image Understand.*, vol. 59, no. 1, pp. 94–115, Jan. 1994.

[20] D. D. Morris and J. M. Rehg, "Singularity analysis for articulated object tracking," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 1998, pp. 289–296.

[21] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1653–1660.

[22] J. Shotton *et al.*, "Real-time human pose recognition in parts from single depth images," *Commun. ACM*, vol. 56, no. 1, pp. 116–124, Jan. 2013.

[23] M. Dantone, J. Gall, C. Leistner, and L. Van Gool, "Human pose estimation using body parts dependent joint regressors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3041–3048.

[24] F. Wang and Y. Li, "Beyond physical connections: Tree models in human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 596–603.

[25] D. Batra, P. Yadollahpour, A. Guzman-Rivera, and G. Shakhnarovich, "Diverse m-best solutions in Markov random fields," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2012, pp. 1–16.

[26] V. Ramakrishna, T. Kanade, and Y. Sheikh, "Tracking human pose by tracking symmetric parts," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3728–3735.

[27] R. Tokola, W. Choi, and S. Savarese, "Breaking the chain: Liberation from the temporal Markov assumption for tracking human poses," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2424–2431.

[28] M. Andriluka, S. Roth, and B. Schiele, "Monocular 3D pose estimation and tracking by detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 623–630.

[29] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation," in *Proc. BMVC*, 2010, vol. 2. no. 4.

[30] Y. Wang, D. Tran, and Z. Liao, "Learning hierarchical poselets for human parsing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 1705–1712.

[31] L. Zhu, Y. Chen, A. Yuille, and W. Freeman, "Latent hierarchical structural learning for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 1062–1069.

[32] H. Jiang and D. R. Martin, "Global pose estimation using non-tree models," in *Proc. IEEE Conf. Comput. Vis.Pattern Recognit. (CVPR)*, Jun. 2008, pp. 1–8.

[33] L. Sigal and M. J. Black, "Measure locally, reason globally: Occlusion-sensitive articulated pose estimation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2. Jun. 2006, pp. 2041–2048.

[34] Y. Tian, C. L. Zitnick, and S. G. Narasimhan, "Exploring the spatial hierarchy of mixture models for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2012, pp. 256–269.

[35] L. Karlinsky and S. Ullman, "Using linking features in learning non-parametric part models," in *Proc. Eur. Conf. Comput. Vis*, Oct. 2012, pp. 326–339.

[36] P. Buehler, M. Everingham, D. P. Huttenlocher, and A. Zisserman, "Upper body detection and tracking in extended signing sequences," *Int. J. Comput. Vis.*, vol. 95, no. 2,, pp. 180–197, Nov. 2011.

[37] K. Fragkiadaki, H. Hu, and J. Shi, "Pose from flow and flow from pose," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2059–2066.

[38] S. Zuffi, J. Romero, C. Schmid, and M. J. Black, "Estimating human pose with flowing puppets," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3312–3319.

[39] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 1385–1392.

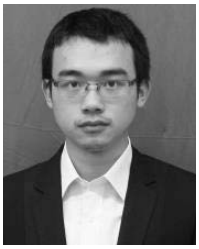
[40] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1. Jun. 2005, pp. 886–893.

[41] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3686–3693.

[42] T. Brox, C. Bregler, and J. Malik, "Large displacement optical flow," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 41–48.

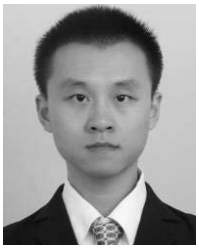
[43] M. I. López-Quintero *et al.*, "Stereo pictorial structure for 2D articulated human pose estimation," *Mach. Vis. Appl.*, vol. 27, no. 2, pp. 157–174, Feb. 2016.

QINGWU LI received the B.S. degree from Zhengzhou University, the M.S. degree from Xidian University, and the Ph.D. degree from Hohai University. He is currently a Professor and a Ph.D. Supervisor with Hohai University. His current research interests include digital image processing, information acquisition, and intelligent perception.



FEIJIA HE received the B.S. degree from Hohai University in 2015, where he is currently pursuing the master's degree. His main research interests focus on digital image processing, video analysis, and machine learning.

TIAN WANG received the B.S. degree from Hohai University in 2014, where she is currently pursuing the master's degree. Her main research interests focus on digital image processing and video analysis.



LIANGJI ZHOU received the B.S. and M.S. degrees from Hohai University, where he is currently pursuing the Ph.D. degree. His current research interests include digital image processing, machine learning, and pattern recognition.

SHUYA XI received the B.S. degree from Hohai University in 2015, where she is currently pursuing the master's degree. Her main research interests focus on digital image processing and video analysis.