IET Research Journals

Submission Template for IET Research Journal Papers

IET Journals
The Institution of Engineering and Technology

ISSN 1751-8644
doi: 0000000000
www.ietdl.org

A two-phase human pose estimation model for resolving double counting problem of tree-structured probabilistic graphical models

M.Moodi[1] *A.Nickabadi*[2]*

[1] *Computer Engineering and Information Technology, Amirkabir University of Technology, 424 Hafez Ave, Tehran, Iran*
[2] *Computer Engineering and Information Technology, Amirkabir University of Technology, 424 Hafez Ave, Tehran, Iran*
* *E-mail: nickabadi@aut.ac.ir*

Abstract: Human pose estimation(HPE) is a challenging problem in many computer vision tasks. Probabilistic graphical modeling is a common approach for HPE in which the human body parts are represented with a tree-based or loopy graph. Tree-based models are appealing due to their tractable exact inference while they suffer from the lack of constraints between symmetric parts. Double-counting, i.e. reporting one location for two symmetric parts of the body, is a common problem of tree-based graphical models. In this paper, we propose a two-phase human pose estimation (TPHPE) model which employs both tree-structured and loopy graphs in two subsequent phases. At the first phase, an initial pose is estimated using a tree-based model considering only kinematic relations. At the second phase, additional constraints are added to avoid the double-counted situations of the first phase. By fixing some parts of body, our model avoids loops in the second-phase's graph and generates new pose candidates with an exact inference. Generated candidates are evaluated using a new score function with further constraints to select the final proposed configuration. Experimental results and analyses on LSP dataset clearly show the efficiency of the proposed model in resolving the double-counting problem of tree-based models.

1 Introduction

Human pose estimation (HPE) in a single 2-D image is the problem of locating body parts in the input image. HPE has an important role in many computer vision tasks such as person search [1, 2], action recognition [3, 4] and human-object interaction [5]. Main challenges for human pose estimation are variation of body poses, complicated background and depth ambiguities.[6]

A large body of HPE approaches are based on the probabilistic graphical models in which the appearance information extracted from the input image is combined with the prior knowledge about the spatial constraints of the human body parts and the best pose is sought as the one that maximizes the posterior probability of the model. Regarding the local appearance information, Histogram of Oriented Gradients (HOG) [7], edge-based descriptors [8], shape context [9] and feature vectors extracted by deep convolutional neural networks [10, 11] are examples of features employed in probabilistic graphical models. The prior knowledge about kinematic and non-kinematic constraints of human body are usually modeled via a graph whose nodes represent the body parts and edges show the relations between parts or joints. Limiting the model to the kinematic relations results in a tree-structured graph while considering additional constraints, e.g. the relations between symmetric parts, makes the graph loopy. As the exact inference in loopy graphs is intractable [12, 13], many researchers have focused on tree-structured graphs.

The approaches based on tree-structured graphs are efficient as exact inference can be done in linear time. However, these models suffer from the double-counting problem, a situation in which one location in the image is detected as the position of two different non-overlapping parts, such as left and right ankles.

There have been studies to solve the double-counting problem in tree models [14–16]. One approach is to generate multiple pose candidates using tree-structured graphs and then rescore these candidates based on additional information that prefers double-count-free configurations [17, 18]. Considering the coverage rate of the extracted pose is another way of avoiding double-counted estimations [18].

In this paper, we propose a two-phase human pose estimation (TPHPE) model in which an initial pose is efficiently estimated in the first step of the algorithm using a tree-structured model. In the second phase, the connections between symmetric parts are added to the model, but to still have a tree, half parts of the human body is fixed to the positions found for these parts in the first phase of TPHPE. This way, two new pose candidates are generated. Finally, three generated candidates are re-evaluated using a more advanced score function and the best configuration is returned. The experimental results on the LSP dataset [19] show that TPHPE successfully eliminates most of the double-counted configurations.

The rest of this paper is organized as follows. In Section 2, we review the related literature on human pose estimation with a special focus on studies on double-counting problem. In Section 3, the proposed model is described. Section 4 evaluates and analyzes the proposed model and compares it with some state-of-the-art models. Finally, Section 5 concludes this paper and points out some possible improvements.

2 Related works

In this section , we first review different approaches and methods and then study solutions proposed for double-counting problem.

2.1 Human pose estimation methods

Most of the current HPE methods are based on the pictorial structure framework [20] which decomposes the human body into several parts (e.g. head, torso, upper and lower leg) and establishes connections between pairs of parts that are kinematically related. The visual features of each part are learned separately using a background/foreground model and counting the number of foreground

pixels within a rectangle area around the part and within a border region around that rectangle. The deformation of the model is statistically formulated by a Gaussian distribution whose parameters are learned with maximum likelihood estimation and assigns higher probabilities to configurations closer to the average pose of the training data. The final pose is estimated as the configuration that maximizes the posterior probability defined as the product of a set of unary (part appearance) and binary (kinematic connection) factors. An efficient inference algorithm is proposed for cases where a tree-structured graphical model is used.

Yang and Ramanan [7] improve the pictorial structure model by introducing a new representation of deformable part models to capture both the spatial relations and the appearance of the parts from the input image patches. To do so, a mixture of histogram of oriented gradient (HOG) descriptors is learned from image patches around each body joint. Each HOG descriptor represents a special type appearance of a part.

With the aim of reducing the search space of the inference step of the pictorial structure model, Eichner et al. [21] and Ferrari et al. [22] search the input image by an HOG descriptor to find the location of the human upper body and then extract the foreground of that area of the image using a grab cut algorithm.

As in other computer vision tasks, recent methods of HPE employ deep neural networks (DNNs) either as the feature extractor (part appearance model) [11] or both the feature extractor and the deformation learning model [23]. One of the first applications of deep neural networks to human pose estimation is DeepPose [24] which employs a sequence of DNNs. The goal of the first DNN is to estimate the overall location of the human in the input image. A number of refining DNNs are then applied to sub-images of the located area of the human body to determine the locations of all parts. In [11], the image evidences of existence and the type of the joints are both learned with a DNN, contributing as two factors of the proposed probabilistic model. In a different approach, a deep neural network and a probabilistic graphical model are trained simultaneously for human pose estimation [23]. In this model, the neural network (the part descriptor) is combined with a Markov Random Field in a single unified model. Pfirster et al. [25] propose deeper convolutional architectures by adding so called *fusion layers* to the previous deep networks. These fusion layers learn the spatial constraints of human body parts and refine the initial heat maps extracted from the input image by removing impossible configurations. Belagiannis and Zisserman use the idea of fusion layer in a recurrent module to iteratively improve the feature maps resulted from a feed forward module [26].

Each factor of a probabilistic model represents some kind of prior or evidential knowledge about the problem. The above unary appearance models of the parts and the binary kinematic relations between pairs of parts are two such factors. The addition of new constraints (factors) to HPE models enriches these models with more information but may create loops in their graphs [12, 13, 27]. The downside of such models is the computational complexity of the exact inference in them which exponentially grows with respect to the size of the largest clique in the graph. To deal with this problem, many approximate inference algorithms such as branch and bound [12], belief propagation [28] and dual decomposition [29] have been used.

In another HPE approach, the visual features of the input image are investigated in different granularities, from coarse-to-fine, to detect the human parts more precisely [27, 30, 31]. A new representation of human parts, called *hierarchical poselets*, is proposed in [30] where poselset ranges from basic rigid parts of the body (e.g. torso, head) to large pieces of human parts consisting of several smaller parts (e.g. upper body). A loopy graph is then used to model the relations of the poselets at the same or different levels. In another work [31], parts at different levels are connected through parent-child links to represent high-order spatial relationships. In this case, the final structure remains tree and the inference can be performed efficiently.

One of the most challenging problems of HPE is the occlusion of the human parts either by other parts of the human itself (self-occlusion) or by the other objects presented in the image (other-occlusion). To address the problem of occlusion, fu et al. consider a random variable for each body part that shows the status of the occlusion of the part as not occluded, self-occluded, or other-occluded [32]. The occlusion states of the parts are also estimated during the inference of this model. In most cases, other-occlusion occurs when a person is interacting with an object. In [33], occlusion is handled by incorporating additional information from these objects through simultaneous detection of actions (e.g. riding, walking), objects (e.g. bicycle, horse) and poses. An HOG representation of each human part and possible interacting (occluding) objects are learned and combined via a tree mixture of parts model.

2.2 Studies for Solving Double-counting problem

As stated before, the double-counting problem occurs when a body part in the input image is counted twice as two different parts of the human body. A straightforward but inefficient solution to the double-counting problem is to consider constraints between symmetric parts of body such as hands or legs[12, 13]. The addition of these new factors to the kinematic factors results a loopy graph with an intractable exact inference. To avoid the computational cost of the loopy models, Lan and Huttenlocher [14] add new latent variables to the kinematic models to capture the relations that are not reflected in the tree models. The final structure is still a tree and a Viterbi algorithm is used for inference. However, as mentioned in [7], such models that include stronger pose priors suffer from overfitting to particular datasets.

Wang and Mori use the idea of multiple trees to model human configurations [15]. Spatial constraints between symmetric parts that are not considered by the kinematic tree are modeled in a set of different trees. Final estimated pose is determined by combining information from different trees by a boosting procedure.

To remedy the problem of loops in the graphical model, [16] introduces a virtual part for each of the non-adjacent symmetric parts and replaces the relation between a pair of symmetric parts with two new connections each of which relating one real part to its virtual pair. The virtual parts are expected to be located close to the positions of their corresponding real parts and prevent pairs of real parts to be estimated at the same position of the input image.

Another approach to tackle the problem of double-counting is to first generate a number of candidate pose configurations using tree-structured models and then rank these configurations with a score function composed of a diverse set of constraints. As an example of these category of HPE models, [17] clusters the training data based on the relative positions of joints and learns a mixture model for each cluster. For each input test image, 20 different candidate positions are generated for each part based on the mixture model learned for the detected cluster of the input image. The final estimation of the pose is obtained by combining the initial candidates based on some constraints specially designed to avoid double-counting.

As a different candidate generation and rescoring method, Gong et al. [18] use additional information in their mixture of parts model to alleviate double-counting problem. They assume that the background can be learned from a sequence of the input video frames and the region occupied by the human can be detected in the form of human blobs by subtracting the background from the input image. The scoring scheme of this model considers the local appearance of the parts as well as the global coverage of the human blobs extracted in the previous step by the estimated pose.

3 The proposed model

In the two-phase human pose estimation (TPHPE) model of this paper, the human body is modeled with an unidirectional graph $G = (V, E)$, in which $V = \{1, \cdots, k\}$ represents human body parts and E models pairwise kinematic relationships between the parts.

The human pose in the input image is represented by $\{L, t\}$ where $L = \{l_i = (x_i, y_i) | i \in V\}$ are the locations of the human parts and $t = \{t_{ij} | (i, j) \in E, t_{ij} \in \{1, \cdots, T\}\}$ show the types of the connections between pairs of parts. t_{ij} models the spatial relation between part i and j and T is a constant representing the number of connection types.

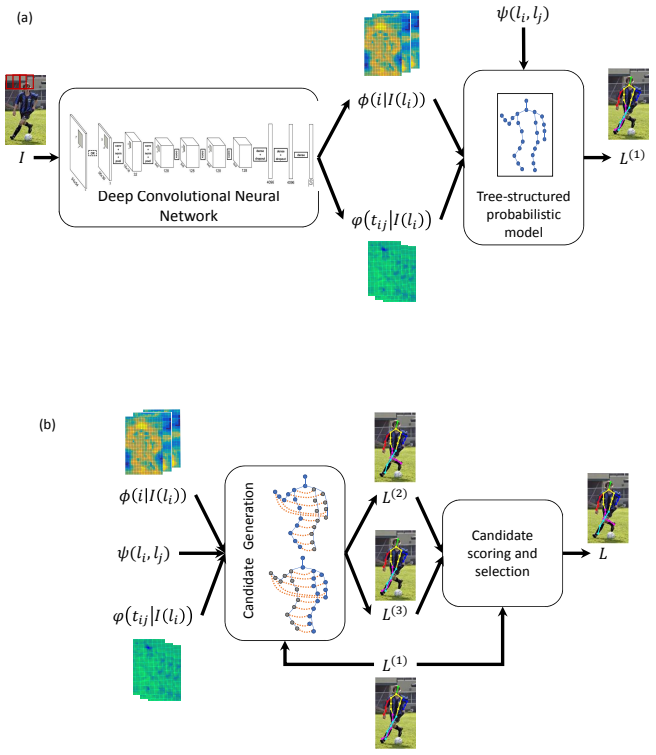


Fig. 1: The overal structure of the proposed Two-phase Human Pose Estimation (TPHPE) model: a) phase one: inital pose estimation, b) phase two: candidate generation and rescoring.

The outline of TPHPE is shown on Fig. 1. As depicted in this figure, TPHPE consists of two steps.The goal of the first step is to guess an initial configuration of the body parts of the human present in the input image. To do so, the human configuration is modeled as an undirected graphical model (a Markov network). A set of unary and binary factors are defined in this probabilistic model to represent the prior knowledge and evidences extracted from the input image. First a set of unary and binary factors are extracted from the input image by means of a deep convolutional neural network(DCNN) applied to the patches of the input image. These factors are then combined with spatial constraints representing the prior knowledge obtained from the training data to keep the extracted configuration consistent. The final configuration of the first step is then obtained via an inference algorithm described in the following subsections.

The graphical model used in the first step is a tree which enables us to perform the inference in a fast and efficient manner employing dynamic programming.

As stated before, tree-structured models suffer from the double-counting problem. The second phase of TPHPE aims at resolving the possible double-counted configurations extracted during the first phase. The main idea of this step of TPHPE is to add factors to the graphical model that enforce the spatial constraints between symmetric body parts. The addition of these new factors to the kinematic constraints converts the structure of the graphical model to a loopy graph and makes the exact inference intractable. To keep the computational advantages of the tree-structured models along with the stronger prior knowledge provided by the newly added constraints, in the candidate generation step of the second phase of TPHPE, the location of one side (e.g. the left side) of the initial configuration is fixed and the model searches for more fitted estimation for the parts of the other side. Repeating this process for the left and right sides of the body, two configurations are found using the same dynamic programming algorithm of the first step albeit with a new set of factors. Finally, a new score function is defined to score the three candidate configurations generated in the two phases of TPHPE.

The reminder of this section details the important parts of TPHPE.

3.1 *Initial pose estimation*

As stated before, the goal of this phase of TPHPE is to provide an initial guess of the target human pose in the form of the location of the body parts (L) and the types of the connections between pairs of parts (t). Here, we use the relational graph $G = (V, E)$ shown in Fig. 1-a In this graph, one node is considered for each of the 26 parts of the body and only kinematic constraints are considered. The types of the connections are defined as discrete values representing different relative positions of the pairs of parts as in [11].

The best configuration for an input image I is estimated as the one that maximizes the following energy function

$$\begin{aligned} F_1(L,t|I) = &\sum_{i\in V} w_i \phi(i|I(l_i);\theta) \\ &+ \sum_{(i,j)\in E} w_{ij}^{t_{ij}} \varphi(t_{ij}|I(l_i);\theta) \\ &+ \sum_{(i,j)\in E} \langle \mathbf{W}_{ij}^{t_{ij}}, \psi(l_i,l_j)\rangle \\ &+ w_0 \end{aligned} \tag{1}$$

where the two first terms, $\phi(i|I(l_i);\theta)$ and $\varphi(t_{ij}|I(l_i);\theta)$, are image dependent factors representing the probability of part presence and connection type, respectively. The third factor of Eq. 1, $\psi(l_i,l_j)$, is defined based on the relative positions of the parts which is learned from the training data. θ shows the parameters of the neural network and w_i, $w_{ij}^{t_{ij}}$ and $\mathbf{W}_{ij}^{t_{ij}}$ are weighting parameters of the model. The weighting parameters are learned using S-SVM approach as in [7].

The values of the two image dependent factors, $\phi(i|I(l_i);\theta)$ and $\varphi(t_{ij}|I(l_i);\theta)$, are obtained from a deep convolutional neural network (DCNN). To do so, fixed-size local patches of the input image ($I(l_i)$) are fed into a DCNN, similar to the one used in [11]. This network consists of five convolutional layers, tree fully connected layers and two max-pooling layers and at the output layer, it has a neuron for each part and connection type of a part. The output of this DCNN is the joint probability of part presence and types of connections in each image patch.

The parameters of the network (θ) are learned from a set of labeled training images. Local patches of the training images centered on the body parts are used as the positive samples labeled with the type of the connection at the position of that part and the random patches from the background of the images are used as negative samples.

The third factor of Eq. 1, $\psi(l_i,l_j)$, is a binary factor that measures the amount of deformation from the average relative positions of pairs of connected parts in the training images. For two adjacent parts i and j with the average relative position $r_{ij}^{t_{ij}} = (\Delta x_{ij}^{t_{ij}}, \Delta y_{ij}^{t_{ij}})$, $\psi(l_i,l_j)$ is defined as follows:

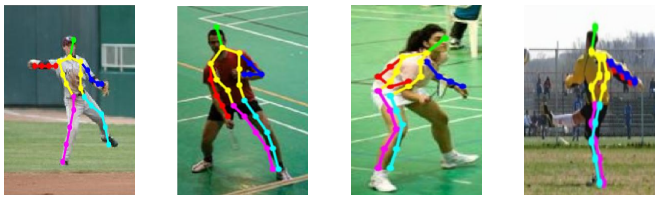
$$\begin{aligned} \psi(l_i,l_j) = &[(x_i - x_j - \Delta x_{ij}^{t_{ij}}), (x_i - x_j - \Delta x_{ij}^{t_{ij}})^2, \\ &(y_i - y_j - \Delta y_{ij}^{t_{ij}}), (y_i - y_j - \Delta y_{ij}^{t_{ij}})^2]. \end{aligned} \tag{2}$$

To learn $(r_{ij}^{t_{ij}})$, the relative positions of each part with respect to its parent are clustered and the centers of these clusters are used as representatives of different connection types of that part.

The final step of the first TPHPE is to find the pose configuration $((L,t))$ that maximizes $F_1(L,t|I)$. As the model used in this phase is a tree, the message passing of the inference algorithm can be done in a dynamic programming manner as in [7].

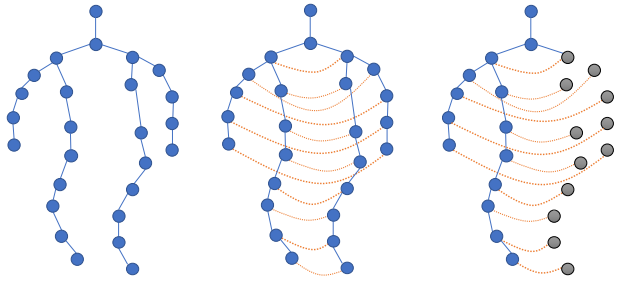
3.2 *Pose correction*

Fig. 2 reports some pose estimation results of the first phase of TPHPE. As expected, some extracted poses suffer from the double-counting problem. For example, in Fig. 2-c, the left leg is detected



(a) (b) (c) (d)

Fig. 2: Sample pose estimation results of the first phase of TPHPE



(a) (b) (c)

Fig. 3: Graph structures used in different phases of TPHPE a) the initial tree-structured graph, b) adding constraints between symmetric parts resulting a loopy graph, c) converting the loopy graph to a tree by fixing the gray nodes.

as the left and right legs or in Fig. 2-d, both hands and legs are overlaid on a single hand and leg of the input image. This is due to the fact that the score function used in the first phase of TPHPE is composed of unary image dependent and binary kinematic relations and no constraint on the positions of symmetric parts of the body is considered in it. The goal of the second phase of TPHPE is to diminish pose estimation errors of this type. To do so, a new factor is added to the energy function as follows:

$$\begin{aligned} F_2(L,t|I) = & \sum_{i\in V} w_i \phi(i|I(l_i);\theta) \\ & + \sum_{(i,j)\in E} w_{ij}^{t_{ij}} \varphi(t_{ij}|I(l_i);\theta) \\ & + \sum_{(i,j)\in E} \langle \mathbf{W}_{ij}^{t_{ij}}, \psi(l_i,l_j)\rangle \\ & - \sum_{(i,k)\in E'} w_{ik} \xi(l_i|l_k,\sigma_i) \\ & + w_0 \end{aligned} \tag{3}$$

where $\xi(l_i|l_k,\sigma_i)$ is a Gaussian function with mean l_k and standard deviation σ_i. This new energy term acts as a penalty for configurations in which the locations of two symmetric parts i and k are too close to each other. As before, w_{ik} is the weighting parameter of the factor defined over parts i and k.

However, the inclusion of these new relations (the dashed links of Fig. 3-b) results in a loopy graph with intractable exact inference. As mentioned earlier, the main focus of this phase of TPHPE is to eliminate the double-counting errors. So, we can have the tree structured graph back by fixing the location of one side of the body and searching for a better candidate for the other side. This way, in the candidate generation step of TPHPE, two new configurations are generated which may be the same as the initial configuration found at the first phase. Fig. 3-c shows the case in which the left side is fixed (the gray nodes). Here, the $\xi(l_i|l_k,\sigma)$ factors are unary factors in the remaining tree structure. The other parts of the model remain the same as the first phase.

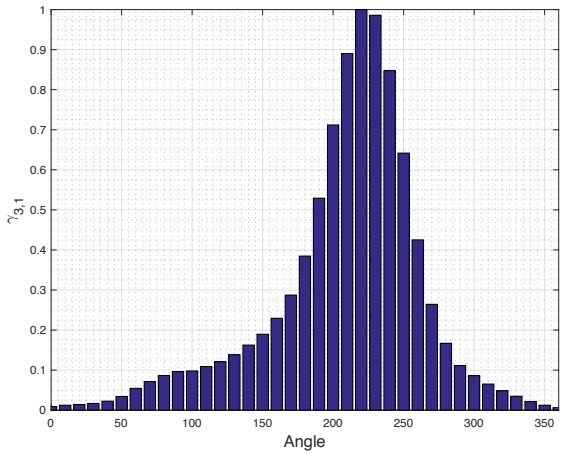


Fig. 4: Histogram of the angle between neck and shoulder learned from the training data.

At the last step of this phase, the three generated pose candidates are evaluated based on a new score function including much more prior knowledge about the human poses. This score function is defined as:

$$\begin{aligned} F_3(L,t|I) = & F_2(L,t|I) \\ & + w_\alpha \sum_{(i,j)\in E} \alpha_{ij} bin(x_i - x_j, y_i - y_j) \\ & + w_\beta \sum_{(i,k)\in E'} \beta_{ik} bin(l_i - l_k) \\ & + w_\gamma \sum_{i\in V} \sum_{c\in kids(i)} \gamma_{ic} bin(a_{ic}) \end{aligned} \tag{4}$$

where $F_2(L,t|I)$ is the score function previously defined in equation 3, $bin(.)$ is a vector of all zeros except for the occupied bin, α_{ij}, β_{ik}, and γ_{ic} are the parameters of the histograms of lengths of limbs, distances between symmetric pairs of parts, and the angles between links of body, respectively. w_α, w_β and w_γ are the weighting parameters and a_{ic} is the angle of the link connecting node i to its kid c. As an example of the above histogram functions, Fig. 4 shows the histogram of the angle between neck and shoulder learned from the training data. As shown in this figure, the possible range of 0 to 360 of this parameter is divided into 36 bins and γ_{ic} is higher for more frequent angles. Now, for an angle of 200 degree, the output of the bin function will be the vector of zeros except for the 20th entry which is one, which multiplied to γ_{ic} vector will result in the output of 0.71.

3.3 Computational complexity

The full-body pose estimation of TPHPE consists of the running of a deep neural network at the first phase for feature extraction, three inferences (one at the first phase and two at the second phase), and a final scoring of the three candidate poses. The time complexity of the scoring function is independent of the size of the input image and the time required for executing the deep neural network depends on the processing hardware and can be done in real time. As mentioned earlier, for the three inference steps of TPHPE performed on tree-structured graphical models, we use a dynamic programming algorithm. Tree structured models which use quadratic functions as pairwise term can benefit from GDT algorithm [20] to reduce computational complexity of dynamic programming inference from

$O(KL^2T^2)$ to $O(KLT^2)$ where K is number of parts, L is number of possible locations for each part in the image and T is the number of types for each part. By fixing half part of the body in the second phase of TPHPE, the graphical model structure remains tree and inference is done as before using dynamic programming in $O(KLT^2)$. Initial inference is done in $O(KLT^2)$, but for the other two candidates, we run inference for $\frac{K}{2}$ parts and the computational complexity would be $O(KLT^2)$. So, the overall complexity of the inference will be $O(KLT^2)$.

4 Experimental Results and Evaluations

In this section, we first describe the dataset and performance criteria used for evaluating the studied HPE methods as well as our proposed model. Then, guidelines are provided for setting the values of TPHPE's parameters. In the following, the results of TPHPE are compared with some state-of-the-art HPE methods and finally, the different scoring functions of TPHPE are studied.

4.1 Datasets and evaluation metrics

In this paper, the popular LSP [19] dataset of annotated human poses is used. This dataset contains 2000 images of people in sport environments. Each image is originally annotated with 14 joints which is then augmented to a 26-key points representation as described in [7]. The dataset is devided into 1000 train and 1000 test images. Left and right joints are labelled from a person-centric viewpoint which can be converted into observer-centric annotation using [34] in which the right and the left body parts are annotated according to the viewpoint of the observer.

The two prevalent criteria for evaluating the performance of HPE methods are the Percentage of Corrected Parts (PCP) [22] and Percentage of Detected Joints (PDJ) [35]. PCP is the most widely used criterion in human pose estimation literature and evaluates the localization accuracy of body parts. We use the strict version of PCP in which two estimated end points are required to be both within half of the parts length from their ground truth positions.

While PCP is still preferred in HPE evaluations, a reported drawback of this criterion is that the score of the estimation of the position of each part depends on the length of the related limb so that the shorter limbs are penalized further. PDJ, in contrast, calculates the error of estimation based on a reference length. The detection rate of the estimated pose is then measured using different thresholds of this reference length. For example, a joint is considered to be correctly detected if the distance between the true position of the joint and its estimated position is less than a fraction of the torso diameter.

4.2 Parameter setting

The main parameters of TPHPE that need to be adjusted before the use of the model are the weighting parameters of F_3 (Eq. 4),$(w_\alpha, w_\beta, w_\gamma)$. To tune the values of these parameters, a 10-fold cross validation is performed on the training part of the LSP dataset based on the PCP measure. Parameters' values are adjusted using a coarse-to-fine iterative approach and at each step the value of a single variable is changed. The best settings found for the LSP dataset is $w_\alpha = 0.01, w_\beta = 9.2, w_\gamma = 6.1$.

The three diagrams of Fig. 5 show the impact of TPHPE's parameters on the PCP performance of this model. In each diagram the values of the other parameters are set to their best values given above. As depicted in this figure, in the best configuration of parameters, w_α is close to zero indicating the small impact of the length factor in finding better pose estimations. On the other hand, w_β representing the weight of the pair distance factor has the largest value indicating the important role of the related factor in finding better pose configurations.

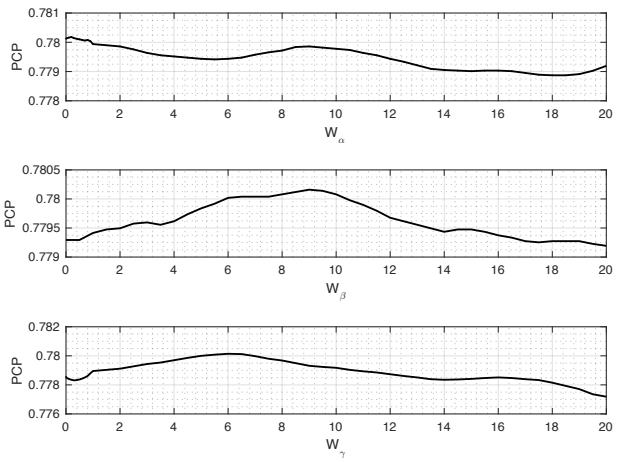


Fig. 5: Changes of PCP criteria with respect to the changes of TPHPE's parameters

4.3 Comparison with other methods

Table 1 gives the results of the proposed model (TPHPE) and some other human pose estimation methods applied to the LSP dataset based on the PCP criterion. As the results show, TPHPE provides better results in comparison to Chen & Yuille. [11] which used same feature extractor and graph-structure that has been used in first phase of TPHPE. A more detailed investigation of the results for different body parts shows that TPHPE's performance on the torso is worse than Chen model due to false double count detection of TPHPE in images containing a side view of human. Finally, as it is clear from the results of Table 1, the superiority of TPHPE over other methods is mainly because of the better estimation of the positions of the body parts that are prone exposed to double-counting problem such as ankles and knees.

To further investigate the ability of TPHPE in resolving double-counting problem,Fig. 6, Fig. 7 and Table 2 provide more detailed results. Fig. 6 compares TPHPE with [11],[38] and [37] in estimating the positions of two challenging parts, knee and ankle. The x-axis represents the PDJ threshold and the Y-axis represents the PDJ accuracy for part localization. It is clear that the TPHPE model outperforms other methods in most thresholds.

Fig. 7 shows examples of the output of the [11] that suffer from the double-counting problem and the results of TPHPE on these images. Table 2 shows the statistical results of double-counted configurations in the first and the second steps of TPHPE. As the results show, the number of pictures that suffer from double-counting problem is dropped from 167 to 77 by applying the second phase of TPHPE.

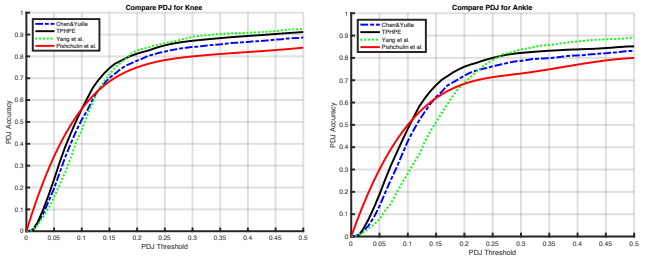
Finally, to investigate the effectiveness of the introduced terms of score function F_3, PDJ metric is calculated on the test data using the three score functions F_1,F_2 and F_3 for selecting the best configuration. The results for two challenging parts of wrist and ankle are shown in Fig. 8. The results show that F_3 score selects better candidates and improves PDJ accuracy in different precision thresholds. In other hand, F_2's results are worse than F_1 and F_3. This happens because F_2 is specially designed to penalize close parts that might cause double-counting to search for possibly better configurations.

5 Conclusion and Future Works

In this paper, we proposed an algorithm for enhancing the performance of human pose estimation by alleviating the double-counting problem of tree-based probabilistic models. To do so, multiple pose candidates are generated employing tree-structured and loopy graphs. These candidates are finally scored using a novel score function combining the prior knowledge about the human pose and evidences obtained from the input image. The proposed TPHPE

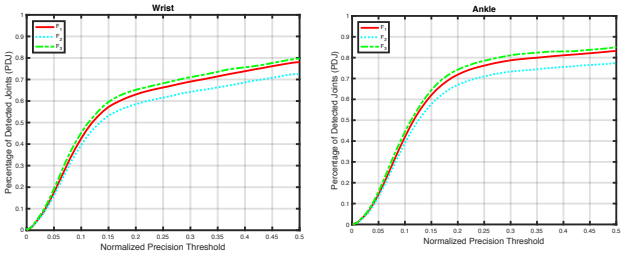
Table 1 Strict PCP of HPE models on LSP dataset. As it is clear, our method outperforms other tree-based HPE models specially in double counted parts(upper and lower legs)

Method	Torso	Head	U.arms	L.arms	U.Legs	L.Legs	Mean
Yang&Ramanan[7]	88.1	77.1	52.5	35.9	69.5	65.6	60.8
Eichner&Ferrari[34]	86.2	80.1	56.5	37.4	74.3	69.3	64.3
Pose Machines[36]	88.1	80.4	62.8	39.5	79.0	73.6	67.8
Pishchulin et al.[37]	88.7	85.1	61.8	45	78.9	73.2	69.2
DeepPose[24]	-	-	56	37	78	71	-
Chen&Yuille [11]	92.7	87.8	69.2	55.4	82.9	77	75
Yang et al.(ChenNet-T)[38]	**94.8**	82.4	**75.0**	**62.4**	85.3	79.2	78.1
ORGM-IDPR[32]	93.9	**89.8**	73	60.7	85.3	79.8	78.1
TPHPE	92.5	87.8	71.9	56.8	**85.5**	**80.7**	77.0



(a) (b)

Fig. 6: PDJ results of TPHPE, Yang et al. [38], Chen & Yuille [11] and Pishchulin et al. [37] on a) knee and b) ankle localization.



(a) (b)

Fig. 8: PDJ results of three different score functions F_1,F_2 and F_3 for lozalizing a) wrist and b) ankle.

Fig. 7: Sample pose estimation results of Chen et al. [11] (up) and TPHPE (down).

Table 2 The number of double-counted configurations for 1000 test images of LSP.

	Phase 1 Output			Phase 2 Output		
	Hands	Legs	Total	Hands	Legs	Total
Number of DCs	33	134	167	23	54	77

model can be added to any tree-based model and as the experimental results on LSP data show, it can effectively reduce the double-counting problem of these models. The inclusion of new features, e.g. part affinity fields [39], generating more candidates in the second step of the method, and considering more informative terms in the final score function are some possibilities for the improvement of TPHPE in the future.

6 References

1 M. Weber, M. Bauml, and R. Stiefelhagen, "Part-based clothing segmentation for person retrieval," in *Advanced Video and Signal-Based Surveillance (AVSS), 2011 8th IEEE International Conference on*. IEEE, 2011, pp. 361–366.

2 A. C. Gallagher and T. Chen, "Clothing cosegmentation for recognizing people," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

3 L. Liu, L. Shao, and P. Rockett, "Human action recognition based on boosted feature selection and naive bayes nearest-neighbor classification," *Signal Processing*, vol. 93, no. 6, pp. 1521–1530, 2013.

4 I. Theodorakopoulos, D. Kastaniotis, G. Economou, and S. Fotopoulos, "Pose-based human action recognition via sparse representation in dissimilarity space," *Journal of Visual Communication and Image Representation*, vol. 25, no. 1, pp. 12–23, 2014.

5 B. Yao and L. Fei-Fei, "Modeling mutual context of object and human pose in human-object interaction activities," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 17–24.

6 Z. Liu, J. Zhu, J. Bu, and C. Chen, "A survey of human pose estimation: the body parts parsing based methods," *Journal of Visual Communication and Image Representation*, vol. 32, pp. 10–19, 2015.

7 Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2878–2890, 2013.

8 J. Sullivan and S. Carlsson, "Recognizing and tracking human action," *Computer Vision-ECCV 2002*, pp. 629–644, 2002.

9 G. Mori and J. Malik, "Estimating human body configurations using shape context matching," *Computer Vision-ECCV 2002*, pp. 150–180, 2002.

10 V. Belagiannis, C. Rupprecht, G. Carneiro, and N. Navab, "Robust optimization for deep regression," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2830–2838.

11 X. Chen and A. L. Yuille, "Articulated pose estimation by a graphical model with image dependent pairwise relations," in *Advances in Neural Information Processing Systems*, 2014, pp. 1736–1744.

12 T.-P. Tian and S. Sclaroff, "Fast globally optimal 2d human detection with loopy graph models," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 81–88.

13 L. Sigal and M. J. Black, "Measure locally, reason globally: Occlusion-sensitive articulated pose estimation," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2. IEEE, 2006, pp. 2041–2048.

14 X. Lan and D. P. Huttenlocher, "Beyond trees: Common-factor models for 2d human pose recovery," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 1. IEEE, 2005, pp. 470–477.

15 Y. Wang and G. Mori, "Multiple tree models for occlusion and spatial constraints in human pose estimation," in *European Conference on Computer Vision*. Springer, 2008, pp. 710–724.

16 L. Fu, J. Zhang, and K. Huang, "Context aware model for articulated human pose estimation," in *Image Processing (ICIP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 991–995.

17 Y. Xiao, H. Lu, and S. Li, "Posterior constraints for double-counting problem in clustered pose estimation," in *Image Processing (ICIP), 2012 19th IEEE International Conference on*. IEEE, 2012, pp. 5–8.

18 W. Gong, Y. Huang, J. Gonzalez *et al.*, "Enhanced mixtures of part model for human pose estimation," *arXiv preprint arXiv:1501.05382*, 2015.

19 S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation," in *British Machine Vision Conference*, 2010.

20 P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *International journal of computer vision*, vol. 61, no. 1, pp. 55–79, 2005.

21 M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari, "2d articulated human pose estimation and retrieval in (almost) unconstrained still images," *International journal of computer vision*, vol. 99, no. 2, pp. 190–214, 2012.

22 V. Ferrari, M. Marin-Jimenez, and A. Zisserman, "Progressive search space reduction for human pose estimation," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

23 J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *Advances in neural information processing systems*, 2014, pp. 1799–1807.

24 A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1653–1660.

25 T. Pfister, J. Charles, and A. Zisserman, "Flowing convnets for human pose estimation in videos," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1913–1921.

26 V. Belagiannis and A. Zisserman, "Recurrent human pose estimation," in *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*. IEEE, 2017, pp. 468–475.

27 K. Duan, D. Batra, and D. J. Crandall, "A multi-layer composite model for human pose estimation." in *BMVC*, vol. 2, 2012, p. 5.

28 S. C. Tatikonda and M. I. Jordan, "Loopy belief propagation and gibbs measures," in *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 2002, pp. 493–500.

29 N. Komodakis, N. Paragios, and G. Tziritas, "MRF energy minimization and beyond via dual decomposition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 3, pp. 531–552, 2011.

30 Y. Wang, D. Tran, and Z. Liao, "Learning hierarchical poselets for human parsing," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1705–1712.

31 L. Zhao, X. Gao, D. Tao, and X. Li, "A deep structure for human pose estimation," *Signal Processing*, vol. 108, pp. 36–45, 2015.

32 L. Fu, J. Zhang, and K. Huang, "Orgm: Occlusion relational graphical model for human pose estimation," *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 927–941, 2017.

33 C. Desai and D. Ramanan, "Detecting actions, poses, and objects with relational phraselets," in *European Conference on Computer Vision*. Springer, 2012, pp. 158–172.

34 M. Eichner and V. Ferrari, "Appearance sharing for collective human pose estimation," in *Asian Conference on Computer Vision (ACCV)*, 2012.

35 B. Sapp and B. Taskar, "Modec: Multimodal decomposable models for human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3674–3681.

36 V. Ramakrishna, D. Munoz, M. Hebert, J. A. Bagnell, and Y. Sheikh, "Pose machines: Articulated pose estimation via inference machines," in *European Conference on Computer Vision*. Springer, 2014, pp. 33–47.

37 L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele, "Strong appearance and expressive spatial models for human pose estimation," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3487–3494.

38 W. Yang, W. Ouyang, H. Li, and X. Wang, "End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3073–3082.

39 Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," *arXiv preprint arXiv:1611.08050*, 2016.