Body Field: Structured Mean Field with Human Body Skeleton Model and Shifted Gaussian Edge Potentials

Sara Ershadi-Nasab
Faculty of Electrical Engineering, Sharif University of Technology, Tehran, Iran
ershadinasab@sharif.edu
Shohreh Kasaei*
Faculty of Computer Engineering, Sharif University of Technology, Tehran, Iran
kasaei@sharif.edu
Esmaeil Sanaei
Faculty of Electrical Engineering, Sharif University of Technology, Tehran, Iran
sanaei@sharif.edu
Erfan Noury
Faculty of Computer Engineering, Sharif University of Technology, Tehran, Iran
erfan.noury@gmail.com
Hassan Hafez-Kolahi
Faculty of Computer Engineering, Sharif University of Technology, Tehran, Iran
hafez@ce.sharif.edu

Received: 28/Aug /2019 Revised: 2/Aug/2020 Accepted: 15/Aug/2020

Abstract

An efficient method for simultaneous human body part segmentation and pose estimation is introduced. A conditional random field with a fully-connected graphical model is used. Possible node (image pixel) labels comprise of the human body parts and the background. In the human body skeleton model, the spatial dependencies among body parts are encoded in the definition of pairwise energy functions according to the conditional random fields. Proper pairwise edge potentials between image pixels are defined according to the presence or absence of human body parts that are near to each other. Various Gaussian kernels in position, color, and histogram of oriented gradients spaces are used for defining the pairwise energy terms. Shifted Gaussian kernels are defined between each two body parts that are connected to each other according to the human body skeleton model. As shifted Gaussian kernels impose a high computational cost to the inference, an efficient inference process is proposed by a mean field approximation method that uses high dimensional shifted Gaussian filtering. The experimental results evaluated on the challenging KTH Football, Leeds Sports Pose, HumanEva, and Penn-Fudan datasets show that the proposed method increases the per-pixel accuracy measure for human body part segmentation and also improves the probability of correct parts metric of human body joint locations.

Keywords: Human body parts; skeleton model; mean field approximation; pose estimation; segmentation; shifted Gaussian kernel**.**

1- Introduction

Human body part segmentation is the problem of segmenting a given image to *human body* (HB) parts and the background. The main difference between this process and the general object segmentation is that the HB has an articulated structure. Human pose estimation is defined as the problem of localization of human body joints in the 2D image or 3D space. Human body part segmentation and pose estimation are challenging tasks in computer vision. Their wide applications include surveillance, motion analysis, human-computer interaction, image understanding, augmented reality, and action recognition. As HB has an articulated structure, pose estimation methods aim to find that configuration in a given image. The articulation in HB is often realized by a skeleton model with 14 body joints as well as the corresponding connections among them [1], [2], [3], [4], [5]. The main challenges involved in HB part segmentation and pose estimation are the occluded body parts, the foreshortening effect on the length of some body parts (caused by projection from the 3D space to the 2D image plane), and the ambiguity in defective body parts (due to motion blur or self-occlusion).

In this paper, a new and efficient method for simultaneous HB part segmentation and pose estimation is introduced. The block diagram of the proposed method is shown in Figure 1. The method is based on a *conditional random field* (CRF) graphical model.

* Corresponding Author

The graphical model is a fully connected graph (shown in Figure 2). The graphical model for human skeleton in the proposed dual pose and segmentation method is shown in Figure 3. The label of each image pixel (graph node) is a random variable of this CRF, taking values from the set $p_0, \dots, p_{14}$, where labels $p_1, \dots, p_{14}$ are body part labels and p_0 is the background label (see Figure 4). In this work, HB joints are modeled in a graph with 14 nodes and the corresponding connections among graph nodes are determined according to the HB skeleton, as it is shown in Figure 3. In the proposed method, the HB skeleton is not restricted to tree; it can also have cycles. Only the unary and pairwise relations are considered in defining the energy function, and higher order relations (e.g. ternary, quadratic, etc.) are neglected.

The spatial dependency of HB joints in the skeleton model, the length of limbs, and the difference between the features of two joints are encoded in the pairwise terms of the CRF energy function. The main contributions of this paper are summarized as following.

- The semantic human body part segmentation and pose estimation problems are modeled, simultaneously, in a single graphical model. Then, an efficient inference method is proposed to minimize the energy function defined by the model.
- The body length constraint is modeled in the proposed fully connected graphical model by the shifted Gaussian kernels considered in the definition of pairwise energy terms.
- It is demonstrated that although the proposed graphical model is fully connected and Gaussian kernels are shifted, the message passing operation in the inner part of the mean field inference can be computed using the fast bilateral filtering approach. Therefore, the inference algorithm remains tractable.
- Experimental results on the popular and challenging pedestrian parsing benchmark Penn-Fudan dataset [6] for semantic human segmentation, and also on the HumanEva I [7], Extended Leeds Sports Pose [8], and KTH Football I [9] datasets show that the proposed method outperforms the method of Xia [10] that is the state-of-the-art in HB segmentation in terms of per-pixel accuracy measure. It also achieves substantial improvement in finding the locations of corresponding joints according to the *probability of correct pose* (PCP) and *probability of correct key points* (PCK) measures in comparison with Chu *et al.[2]* that is state-of-the-art in 2D pose estimation.

The rest of this paper is organized as follows. In Section 2-, related literature and previous research is reviewed. In Section 3-, the method of Kraehenbuehl *et al.*[11] is reviewed that is necessary for explaining the proposed method. In Section 4-, the proposed method is explained. Next, in Section 5-, experimental results are given. Finally, Section 6- concludes the paper.

2- Related Work

The problem of HB part segmentation and pose estimation can be approached simultaneously. The best graphical model for solving this problem would have to take into account the relations among all image pixels. However, when considering image pixels as the nodes of a fully connected graphical model, the computational cost of the inference step will be very high. Kraehenbuehl *et al.* [11] showed that the inference in dense CRF can successfully be performed by mean field approximation using efficient high dimensional Gaussian filtering operations [12]. The method is specifically designed for the general segmentation problem without any constraint on articulation of HB part.

The kernels are Gaussian functions on the position or color space. No other image features, such as *histogram of oriented gradients* (HOG) [13] are used. Other researchers tried to use this efficient inference and filtering in pose estimation tasks. Vineet *et al.*[14] used this efficient inference in the joint HB pose estimation, segmentation, and depth estimation in a method called *PoseField*.

However, the energy function defined by them is not specialized for HB and does not reflect the HB skeleton model. Kiefel *et al.*[15] tried to extend the inference method introduced in [11] to pose estimation problem. They introduced the *field of parts* method to detect HB joints in 2D images. In their method, the local appearance and joint spatial configuration of HB are modeled. Recently, models based on *deep convolutional neural networks* (DCNN) have been studied extensively in 2D human pose estimation [1], [2], [3], [16].

The *convolutional pose machines* (CPM) architecture proposed by Wei *et al.*[16] is a sequential convolutional neural network that enforces intermediate supervision at the end of each stage to prevent vanishing gradients. DeeperCut [1] is a multi-person pose estimation approach that adapts the deep residual network for human body part detection and uses integer linear programming to jointly detect multiple persons and estimate their body part configurations. Chu *et al.*[2] incorporated the DCNN with a multi-context attention mechanism into an end-to-end framework for human pose estimation. They adapt stacked hourglass networks to generate attention maps from features at multiple resolutions with various semantics. Bulat *et al.*[3] designed a DCNN cascaded architecture specifically for learning part relationships and spatial context. The first part of their cascade outputs part detection heat maps and the second part performs regression on these heat maps to estimate the 2D body pose. Kazemi *et al.*[9] tried to learn the body shape in a discriminative approach using *random forest* (RF) classifier to capture the variations in appearances of HB parts in 2D images. Semantic segmentation and human parsing based on shape-based methods has been studied in [17]. They generate region proposals, rank them using shape and appearance features, and assemble the proposals with simple geometric constraints. A Bayesian framework for jointly estimating articulated body pose and pixel-level segmentation of each body part is proposed in [18].

Wang *et al*.[19] proposed a joint solution that tackles the semantic object and part segmentation, simultaneously. In that method [19], the higher object-level context is provided to guide the part segmentation process. Also, more detailed part-level localization is utilized to refine the object segmentation process.

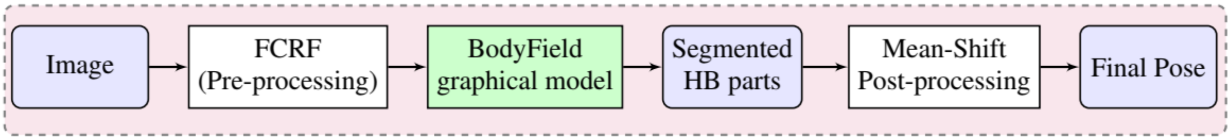


Figure 1: Schematic view of proposed method.
Input: Image block. Outputs: Segmented HB Parts and Final Pose.

A *deep decompositional network* (DDN) for parsing pedestrian images into semantic regions is proposed in [20]. This method tries to directly map low-level visual features to the label maps of body parts. Top-down pose cues as well as deep-learned features are used in an *and-or graph* (AOG) for semantic part assembling [10]. This method tries to refine the semantic parts of objects by using the pose cues. DeepLab framework [21] augments fully convolutional network with dilated convolutions, atrous spatial pyramid pooling, and CRF. DeepLab obtains state-of-the-art performance in general problem of semantic segmentation. Guler *et al*. [22] proposed a surfaced based framework for dense human pose estimation and body part segmentation. It is based on finding dense correspondence between image and a surface of human body. Since there is not a large-scale dataset containing correspondence between image and human body surface, this method has some challenges with general and natural images. An occlusion aware framework for human pose estimation is proposed in [23]. It is based on adversarial training of a *Convolutional Neural Network* (CNN). They designed discriminators to distinguish the real poses from the fake ones (such as biologically implausible ones) to avoid fake estimated poses. Peng *et al*. [24] used data augmentation method in training phase of an adversarial learning framework. They proposed to optimize data augmentation and network training jointly to avoid overfitting for the task of human pose estimation. *Yang et al.* [25] tried to learn 3D human pose structure from a dataset with only 2D pose annotation as the ground-truth. Their method is based on an adversarial learning framework using multi-source discriminators to distinguish the predicted 3D poses from the ground-truth one. In fact, they tried to enforce the pose estimator to generate anthropometrically valid poses even with images from natural scenes. Chen *et al.* [26] proposed a method for multi-person pose estimation in challenging scenes that contain occluded or invisible keypoints and complex backgrounds. They used cascaded networks of GlobalNet and RefineNet. Simple key points like eyes and hands are localized with the GlobalNet. Hard keypoints such as occluded or invisible key points are addressed with the RefineNet network. Also, this method handles only the pose estimation problem and does not handle the body part segmentation problem. PoseTrack is a large-scale benchmark for video-based human pose estimation and articulated tracking [27]. It is a more suitable dataset for multiple human tracking task in video sequences rather than body part segmentation since it does not have any ground-truth information for human body segmented regions. It is worth mentioning that the proposed method is different from the Kraehenbuehl *et al*.'s work[11], in that in the proposed method, the CRF formulation is specifically defined according to the HB configuration such that HB segments are naturally considered to appear in a set of constrained positions relative to each other.

Also, the definitions of pairwise energy terms are different from that work. Since the graphical model used in the proposed method is a fully-connected graph constrained to image pixels, it is similar to the work of Kiefel *et al.* [15], albeit that method does not produce the HB part segmentation and they only report the PCP values on the Leeds Sports Pose [8] dataset.

3- Efficient Mean Field in Object Segmentation

Kraehenbuehl *et al.* [11] proposed an efficient inference in mean field approximation for general segmentation problem. Their method is not designed for articulated objects such as human body and is only evaluated in PASCAL dataset for general object segmentation problem.

In this Section a brief description of Kraehenbuehl *et al*.'s [11] method is reviewed that is needed for introducing the proposed method in the next Section. They defined a conditional random field over a set of random variables $X = \{x_1, \dots, x_N\}$, where N is the total number of pixels in image I. Each variable has a set of possible labels $P = \{p_0, \dots, p_k\}$, where p_0 corresponds to the background and $p_1, \dots, p_k$ are possible pixel labeling.

The conditional random field is characterized by the Gibbs energy function defined on this graph by

$$E(x) = \sum_i \psi_{unary}(x_i = p) + \sum_{i<j} \psi_{pairwise}^{(1)}(x_i = p, x_j = p') \qquad (1)$$

where i, j range from 1 to N.

The Gibbs energy function is a summation of pairwise and unary terms. The $\psi_{unary}(x_i = p)$ is the cost of assigning label p to random variable x_i . The second term, $\psi_{pairwise}^{(1)}(x_i = p, x_j = p')$, measures the cost of assigning label p and p' to two neighboring pixels i and j , respectively. The pairwise term is the cost of assigning two different labels to two arbitrary pixels, given by

$$\psi_{pairwise}^{(1)}(x_i = p, x_j = p') = \mu_1(p, p') \sum_{m=1}^{M} w_1^{(m)}\, k^{(m)}(f_i, f_j) \qquad (2)$$

where $k^{(m)}(f_i, f_j)$ is a Gaussian kernel and is defined as

$$k^{(m)}(f_i, f_j) = \exp\left\{-\frac{1}{2}\left((f_i - f_j)^T (\Sigma^{(m)})^{-1} (f_i - f_j)\right)\right\} \qquad (3)$$

in which vectors f_i and f_j are feature vectors of pixels i and j in an arbitrary feature space, respectively, $w_1^{(m)}$ is the weight of the kernel, m is the index of the kernel, and M is the number of kernels. $\Sigma^{(m)}$ is matrix of variances between f_i and f_j of m −th Gaussian kernel. The energy function defined in the CRF formulation is minimized during the inference phase. The mean field approximation is an iterative process that instead of computing the exact distribution P, computes the approximated $Q(x)$ such that minimizes the KL -divergence $D(Q||P)$ among all distributions, where Q can be expressed as the product of independent marginal $Q(x) = \prod_i Q_i(x_i))$. According to the energy function defined in Equation(1), the closed-form solution of the mean field approximation can be written as

$$Q(x_i = p) = \frac{1}{Z_i} \exp\{-\psi_{unary}(x_i = p) - \hat{Q}_1(x_i = p)\} \qquad (4)$$

where $Q(x_i = p)$ is the belief of pixel i about having the label p and is updated in iterative steps. Z_i is defined as $\mathrm{Z}_i = \sum_{p=1}^{P} Q(x_i = p)$ and is the normalization term. Also, $\psi_{unary}(x_i = p)$ is the initial belief about pixel i having the label p. The belief of all other pixels about pixel i having the part label p is defined as

$$\hat{Q}_1(x_i = p) = \sum_{p' \in P} \mu_1(p, p') \sum_{m=1}^{M} \omega_1^{(m)} \tilde{Q}_1^{(m)}(x_i = p') \qquad (5)$$

in which, $\mu_1(p, p')$ is the label compatibility function between two possible labels p and p' for each pixel.

A simple label compatibility function is the *Potts model*, in which

$$\mu_1(p, p') = 1(p \neq p') \qquad (6)$$

where $1(p \neq p')$ denotes the indicator function.

$\omega_1^{(m)}$ is the weight of m −th Gaussian kernel, M is the total number of kernels, and

$$\tilde{Q}_1^{(m)}(x_i = p') = \sum_{i \neq j} k^{(m)}(f_i, f_j) Q(x_j = p') \qquad (7)$$

in which $k^{(m)}(f_i, f_j)$ is a Gaussian kernel as is defined in Equation (3). It is worth mentioning that Equation (7) is performed once for all pixels by using the Permutohedral lattice filtering. Every channel p' of matrix Q is blurred by Gaussian kernel of $k^{(m)}(f_i, f_j)$ as in Equation (7) that are applied on all image pixels. By substituting Equations (5), (6), and (7) in Equation (4) the message passing is performed as

$$Q(x_i = p) = \frac{1}{Z_i} \times \exp\{-\psi_{unary}(x_i = p) - \sum_{p' \in P} \mu(p, p') \sum_{m=1}^{M} \omega_1^{(m)} \sum_{j \neq i} k^{(m)}(f_i, f_j) Q(x_j = p')\}. \qquad (8)$$

Since the graphical model is a fully-connected graph, the message passing step is the bottleneck of the mean field approximation. Its run-time is quadratic in the number of pixels N.

4- Proposed Method

The block diagram of the proposed method is illustrated in Figure 1. The Image block is input to the method, the FRCF block is a pre-processing step that computes the initial pose that is needed in the next block. The details of this pre-processing step are explained in Subsection 4-1-. The BodyField Graphical model is the proposed method that is explained in detail in Subsection 4-2-. The Segmented HB parts are the output of the method. The Mean-Shift block is a post-processing step that is applied to the distribution of the segmented body parts for computing the final estimated pose. The Final Pose block is the final estimated pose and output of the method.

4-1- Pre-processing: Computing the Initial Pose by a Fully Connected Pairwise CRF

A fully connected pairwise CRF is proposed for computing the initial pose that is needed in the proposed dual pose and segmentation method. The graphical model of human body according to this CRF is shown in Figure

2. The nodes of this graph are human body joints that all of them are connected to each other.

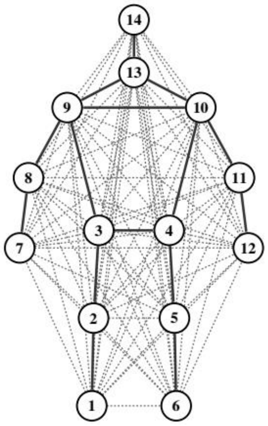


Figure 2. Proposed fully connected model of human body. This model is used in pre-processing step to find the initialy estimated pose.

Images are initially processed with the DeeperCut 2D part detector [1] and the score map of body joints in the images are obtained. The score map, S, is an array of size $W \times H \times 15$ where W and H are the width and height of the image, respectively, and 15 is the number of body joints (14) plus a special class for the background. The unary term of the energy function is computed by the first output of body part detector, S, as

$$\psi_{unary}(X_{(i,j)} = p) = S(x^{(i,j)}, y^{(i,j)}, p). \tag{9}$$

Another output of 2D part detector is R, that is an array of size $W \times H \times 14 \times 13 \times 2$, where 14×13 indicates the number of permutations of length two of 14 distinct variables, and 2 is for two dimensions x and y.
According to the output R of the part detector [1]

$$\left(\delta x^{(i,j)}, \delta y^{(i,j)}\right) = R\left(x^{(i,j)}, y^{(i,j)}, \upsilon(p,p')\right) \tag{10}$$

which implies that if a pixel in location (i,j) has the joint label p, it is expected that the joint p' will occur with an offset $\left(\delta x^{(i,j)}, \delta y^{(i,j)}\right)$ from it. Also, $\upsilon(p,p')$ is an index between 1 and 182 which indicates one of the possible permutations 14×13 belonging to joints p and p', according to [1]. Therefore, if joint i is in location $(x^{(i,j)}, y^{(i,j)})$, then the model expects that joint j to be in location

$$\left(\tilde{x}^{(i,j)}, \tilde{y}^{(i,j)}\right) = \left(x^{(i,j)} + \delta x^{(i,j)}, y^{(i,j)} + \delta y^{(i,j)}\right). \tag{11}$$

In the same way, if a pixel in location (i',j') has the joint label p', acording to the output of the part detector[1], it expects that the joint p be in the offset

$$\left(\delta x^{(i',j')}, \delta y^{(i',j')}\right) = R\left(x^{(i',j')}, y^{(i',j')}, \upsilon(p,p')\right), \tag{12}$$

from it. Therefore the expected location of joint p from the point of view of pixel (i',j') that has joint label p' is

$$\left(\tilde{x}^{(i',j')}, \tilde{y}^{(i',j')}\right) = \left(x^{(i',j')} + \delta x^{(i',j')}, y^{(i',j')} + \delta y^{(i',j')}\right). \tag{13}$$

The difference vector between the expected location of joint p' from the point of view of pixel (i,j) that has joint label p and pixel (i',j') that has joint label p' is

$$\Delta^{(1)} = \| \left(\tilde{x}^{(i',j')}, \tilde{y}^{(i',j')}\right) - \left(x^{(i',j')}, y^{(i',j')}\right) \|. \tag{14}$$

Also, the difference vector between the expected location of joint p from the point of view of pixel (i',j') that has joint label p' and pixel (i,j) that has joint label p is

$$\Delta^{(2)} = \| \left(\tilde{x}^{(i,j)}, \tilde{y}^{(i,j)}\right) - \left(x^{(i,j)}, y^{(i,j)}\right) \|. \tag{15}$$

The pairwise term as the cost of assigning label p to pixel (i,j) and label p' to pixel (i',j') is defined as

$$\psi_{pairwise}(X_{(i,j)} = p, X_{(i',j')} = p') = exp\left\{-\frac{1}{2} \| \Delta^{(1)} + \Delta^{(2)} \|_2^2\right\}. \tag{16}$$

The inference in the proposed fully connected CRF is computed by the loopy belief propagation method [28]. Using this pre-processing step improves the estimated pose of the DeeperCut method. The comparison between the estimated pose in this pre-processing step, (FCRF), and DeeperCut method is provided in experimental results Section 5-. The initial pose obtained by the pre-processing, (FCRF), is used in computation of the amount of needed shift values in the proposed method in the next Section.

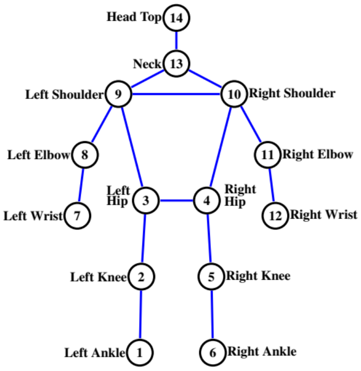


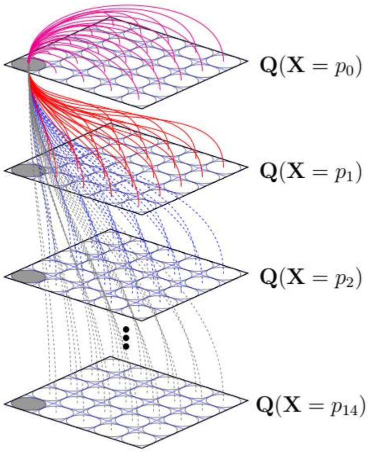
Figure 3. Proposed graph for human skeleton model in proposed dual pose and segmentation method.

4-2- BodyField Graphical Model Definition

According to Figure 3, human skeleton model is considered to contain 14 joints and their connections are set according to the HB configuration. Furthermore, as the graph is not restricted to be a tree, the model can easily be extended to arbitrary number of HB parts and there is no hard constraint on the number of joints in the model. Figure 4 illustrates the proposed fully-connected graphical model. The nodes in the proposed graphical model are image pixels, and pairwise terms are weights of any connection between two arbitrary pixels. A pixel i is shown to be connected to all other pixels with labels in $p_0, \dots, p_{14}$. It is also true for all other pixels (due to visualization restrictions, other connections are not shown). Also, it is important to note that there are no connections, and thus pairwise terms, between a pixel and itself. Since the pairwise terms between two pixels are constrained to the label compatibility, for visualization purposes, image labels are separated to $L = 15$ channels.

These 15 channels should be added to create a fully connected graphical model. Therefore, there are $W \times H$ nodes in the graph, in which W and H are the width and height of the image, respectively. Also, $Q(X = p)$ is the probability of assigning label p to a set of image pixels X.

Energy function should be defined such that a true configuration of HB corresponds to the minimum value of the energy function, otherwise, finding the minimum value of the energy function will not lead to a good configuration. Note that the sum of probability values of parts for each pixel is one. The label assigned to each pixel is the HB part with the highest probability value among all HB parts and the background. The pairwise energy terms are defined such that the pairwise terms have lower values when the two paired pixels have corrected HB part labels. In the general segmentation problem, it is assumed that pixels that are close to each other (in the feature space) lie in the same segment. It can be met in general segmentation problems, but it does not always hold in HB part segmentation.



$\mathbf{Q}(\mathbf{X} = p_0)$
$\mathbf{Q}(\mathbf{X} = p_1)$
$\mathbf{Q}(\mathbf{X} = p_2)$
$\mathbf{Q}(\mathbf{X} = p_{14})$

Figure 4. Fully connected graph for human skeleton model.

Some body parts should occur in pre-defined distances to each other in accordance to the existence of a connection among related joints in the HB skeleton model. The inference process tries to find the minimum of the energy function; in the final solution, all nearby (generally, in the feature space) pixels will have similar labels. But, in pose estimation problems, this is not always true. The reason is simply that there might be nearby and similar pixels in the image of HB that do not belong to the same part. In images of HB, there are three common categories of relationships among pixels.

- Pixels that are close in the feature space and belong to the same HB part.
- Pixels that are close in the feature space but do not belong to the same HB part, however their corresponding parts are connected in the HB skeleton model.
- Pixels that may or may not be close in the feature space and do not belong to the same HB part, but their corresponding parts are not connected in the HB skeleton model.

Pixels belonging to the third type can move and eventually appear close to each other; e.g. the wrist can appear near the other parts of HB. When defining the energy function and pairwise terms, all of these situations should be considered and the suitable kernel and compatibility functions should be assigned to any two labels. For any two arbitrary pixels, according to their labels, two different pairwise terms are defined. One for resolving the first and the third type and the other for resolving the second type. In the proposed method the energy function is defined as

$$E(X|I,\theta) = \sum_i \psi_{unary}(x_i = p|I,\theta) + \sum_{i<j} \psi^{(2)}_{pairwise}(x_i = p,\, x_j = p'|I,\theta) + \sum_{i<j} \psi^{(2)}_{pairwise}(x_i = p,\, x_j = p'|I,\theta) \tag{17}$$

where i, j range from 1 to N. Variable θ denotes the set of parameters of HB. It is computed by the initial pose that is estimated in the Subsection 4-1-.

For the sake of conciseness, in the remainder of the paper, I and θ are omitted in equations. If these two pixels are close to each other in the feature space, the energy cost for assigning different labels to these two pixels is high. When minimizing the energy function during the inference process, this configuration of labeling (two nearby pixels with two different labels) will be avoided. Therefore, in the best configuration, neighboring pixels approach towards getting identical labels. This is generally true in articulated HB shapes and therefore these pairwise terms are defined between any two arbitrary pixels by using a simple Potts model. The second type of pairwise energy terms is specifically defined to encode HB joints' constraints in the proposed CRF formulation, given by

$$\psi^{(2)}_{pairwise}(x_i = p, x_j = p') = \mu_2(p,p') \sum_{m=1}^{M} \omega_2^{(m)} \kappa_{p,p'}^{(m)}(f_i, f_j) \tag{18}$$

where $\omega_2^{(m)}$ is the weight of the shifted kernel function. The label compatibility function $\mu_2(p,p')$ is defined as

$$\mu_2(p,p') = -1(p\ is\ connected\ to\ p') \tag{19}$$

according to the existence of a connection between body part p and body part p' in the HB skeleton model as it is shown in Figure 3. The value of $\kappa^{(m)}_{p,p'}(f_i,f_j)$ is defined as

$$\kappa^{(m)}_{p,p'}(f_i,f_j) = \exp\left\{-\frac{1}{2}\left((f_i - f_j - \eta^m_{p,p'})^T(\Sigma^{(m)}_{p,p'})^{-1}(f_i - f_j - \eta^m_{p,p'})\right)\right\} \tag{20}$$

in which $(\Sigma^{(m)}_{p,p'})^{-1}$ is the variance matrix between feature vector of joint p and p' of m −th shifted Gaussian kernel. $\eta^m_{p,p'}$ is the mean expected difference vector between the features f_i and f_j. When the features are simply the positions of points, the value of $\eta^m_{p,p'}$ is a difference vector that is computed from the initial pose that is estimated by the preprocessing step of Subsection 4-1-.

Let us consider two arbitrary pixels which have two different labels and are connected according to their labels in the HB skeleton model. The pairwise term that is defined for these two pixels, takes the minimum value when these pixels are placed at a predefined distance from each other. By this definition, the Gaussian term is shifted by $\eta^m_{p,p'}$, such that the mean of Gaussian lies on pixels for which the difference between their features and the feature of pixel i is $\eta^m_{p,p'}$. The pairwise energy is the weight of edges in the fully connected model between pixels and it is constrained on labels of pixels. There will be $15\ \times 14$ pairwise energy terms between any two pixels. There are some constraints on HB skeleton model according to the skeleton graph. The goal of the proposed method is enforcing all constraints presented in the HB skeleton model in the mean field approximation process. Note that, up to here, body part lengths and nearby joints that are connected in the skeleton graph are successfully encoded in the energy function definition of the fully connected conditional random field model that is defined on image pixels. In defining the pairwise energy terms between two arbitrary pixels, $XYRGBHOG$ kernel is used that is a 36-D vector $f_i = (x_i, y_i, r_i, g_i, b_i, f_i^{(1)}, f_i^{(2)}, \ldots, f_i^{(31)})$ where x_i and y_i are coordinates of pixel i, r_i, g_i, b_i, the RGB values of the pixel, and $f_i^{(l)}$ is the l^{th} element of HOG feature vector of the cell containing that pixel. f_i and f_j in Equation (20) are defined as above.

4-3- Efficient Inference Via High Dimensional Gaussian Filtering

According to the energy function defined in Equation (17), the closed-form solution of the mean field approximation can be written as

$$Q(x_i = p) = \frac{1}{Z_i}\exp\{-\psi_{unary}(x_i = p) - \hat{Q}_1(x_i = p) - \hat{Q}_2(x_i = p)\} \tag{21}$$

where $Q(x_i = p)$ is the belief of pixel i about having the label p and is updated in iterative steps. Z_i is defined as $Z_i = \sum_{p=1}^{P} Q(x_i = p)$ and is the normalization term.

Also, $\psi_{unary}(x_i = p)$ is the initial belief about pixel i having the label p. $\hat{Q}_1(x_i = p)$ and $\hat{Q}_2(x_i = p)$ are the belief of all other pixels about pixel i having the part label p.

The value of $\hat{Q}_2(x_i = p)$ in Equation (21) is defined as

$$\hat{Q}_2(x_i = p) = \sum_{p'\in P}\mu_2(p,p')\sum_{m=1}^{M}\omega_2^{(m)}\tilde{Q}_2^{(m)}(x_i = p') \tag{22}$$

in which $\mu_2(p,p')$ is the label compatibility function and is defined in Equation (19), $\omega_2^{(m)}$ is the weight of shifted Gaussian kernel, and

$$\tilde{Q}_2^{(m)}(x_i = p') = \sum_{j\neq i}\kappa^{(m)}_{p,p'}(f_i,f_j)Q(x_j = p') \tag{23}$$

in which $\kappa^{(m)}_{p,p'}(f_i,f_j)$ is a shifted Gaussian kernel, is defined in Equation (20).

It is worth mentioning that Equations (7) and (23) are performed once for all pixels by using the Permutohedral lattice filtering.

Every channel p' of matrix Q is blurred by Gaussian kernel of $k^{(m)}(f_i,f_j)$ as in Equation (7) and by shifted Gaussian kernel of $\kappa^{(m)}_{p,p'}(f_i,f_j)$ as in Equation (23) that are applied on all image pixels. By substituting Equations (5), (7), (22), and (23) in Equation (21) the message passing is performed as

$$\begin{aligned}Q(x_i = p) = \frac{1}{Z_i} \times \exp\{&-\psi_{unary}(x_i = p) - \\ &\sum_{p'\in P}\mu_1(p,p')\sum_{m=1}^{M}\omega_1^{(m)}\sum_{j\neq i}k^{(m)}(f_i,f_j)Q(x_j \\ &= p') \\ &- \sum_{p'\in P}\mu_2(p,p')\sum_{m=1}^{M}\omega_2^{(m)}\sum_{j\neq i}\kappa^{(m)}_{p,p'}(f_i,f_j)Q(x_j \\ &= p')\}.\end{aligned} \tag{24}$$

Since the graphical model is a fully-connected graph, the message passing step is the bottleneck of the mean field approximation. Its run-time is quadratic in the number of pixels N. As another contribution of the proposed method, shifted Gaussian kernels are used in the pairwise terms in addition to the non-shifted Gaussian kernels, while keeping the inference step computationally tractable.

4-4- Implementation Details of Shifted Gaussian Kernels

Permutohedral lattice high dimensional Gaussian filtering, performs the filtering task in three steps [12]:

- Splatting the points to the lattice space.
- Performing the blurring process in lattice space.
- Slicing the lattice to find the final values of blurred points.

Splatting is the initial phase of lattice construction in high dimensional space according to the definition in [12]. Since we want to blur the value of $Q(x_j = p')$ with position vectors that are shifted by $\eta_{p,p'}^{(m)}$, it implies that at first the position vector shifts by $\eta_{p,p'}^{(m)}$ before performing the blurring task. But, in lattice space the operation of shifting and then blurring is equivalent to blurring and then slicing at the shifted positions. Substituting Equation (20) in Equation (23) will result in

$$\tilde{Q}_2^{(m)}(x_i = p') = \sum_{i \neq j} exp\left\{-\frac{1}{2}\left(f_i - f_j - \eta_{p,p'}^{(m)}\right)^T \left(\Sigma_{p,p'}^{(m)}\right)^{-1} \left(f_i - f_j - \eta_{p,p'}^{(m)}\right)\right\} \times Q\left(x_j = p'\right). \tag{25}$$

The *permutohedral lattice filter* [12] is implemented in the ImageStack library [29], which is a toolbox for high dimensional Gaussian filtering. It is used for performing the high dimensional blurring in the inference step of the proposed method. In implementation process, according to Equation (25), Q is a matrix of size $(W \times H \times L)$, given that the input image is of size $(\mathrm{W} \times \mathrm{H})$, where L is the total number of body parts and background labels ($L = 15$). $Q(x_i = p)$ is the probability of part p for each arbitrary x_i pixel of the image. It is necessary that $\sum_{p \in P} Q(x_i = p) = 1$, in which $i \in N$ and N is the set of all image pixels. Taking Equations (7) and (23) into account, it is apparent that both equations are similar, except that in the former, Gaussian weights are shifted. Baek *et al*.[30] proved that to use shifted Gaussian kernels, it is sufficient to slice the lattice at shifted positions. Using the ImageStack library, the lattice points are ordinary position vectors without shifting and the values of $Q(x_j = p')$ are blurred in the lattice space by using position vectors in Gaussian weights. Afterwards, the lattice should be sliced to find the final values of $Q(x_j = p')$ in the initial space. In Equation (7), for all channels of matrix Q, these operations are performed once using only a single lattice. Permutohedral lattice filter reduces the time complexity of Gaussian operation to $\mathrm{O}(nd^2)$, where n is the number of points to be blurred and d is the dimension of the position space (despite the fact that its three required steps of splatting, blurring, and slicing are time consuming; specifically in high dimensional spaces like HOG feature space). In the proposed method, the shifted Gaussian filtering is performed for several times, which is time consuming. It is worth mentioning that to further speed-up the process, one can force all $\Sigma_{p,p'}^{(m)}$ to be the same for all p and p', and therefore some steps need only be performed once for updating the belief about each label, as done in Equation (24). Note that for all shifted Gaussian kernels that use the same feature space and covariance matrix, constructing and blurring the lattice in the feature space is only performed once. On the contrary, due to different values of $\eta_{p,p'}^{(m)}$, the lattice is sliced in different shifted positions.

5- Experimental Results

The proposed method is evaluated on (i) KTH Football I dataset with 3900 training images and 2007 test images, (ii) Extended Leeds Sports Pose dataset with 11000 training and 1000 test images, and (iii) Sequences 1 and 2 of the HumanEva I dataset for jogging, walking, and balance actions. Since the proposed method has the 2D pose output in addition to the body part segmentation output, it is also evaluated on those datasets that have 2D pose annotations. The obtained results are then compared with that of the 2D pose estimation methods. It should be noted that as the KTH Football I, Extended Leeds Sports Pose, and HumanEva I datasets do not have any ground-truth data annotations for HB part segmentation, the results can only be evaluated qualitatively. For evaluating the proposed method in human body segmentation, the Penn-Fudan dataset is used. It contains 170 test images and the ground-truth of the body part segmentation. Some typical results of the proposed HB part segmentation are shown in Figure 5, Figure 6, Figure 7 and Figure 8 in which the first column (a) shows the original images. BodyField segmentation result and estimated pose are shown in column (b) of these Figures. In column (c), the HB part segments of the BodyField method is visualized by similar colors as in the ground-truth of Penn-Fudan dataset using the estimated pose of the BodyField method. As it can be seen from Figure 5, Figure 6, Figure 7, and Figure 8, the locations of joints have been estimated accurately, due to the refined HB part segmentation obtained by the proposed method. In KTH Football the PCP metric is used to evaluate the accuracy of pose estimation methods. According to the definition in [31], a part is considered correctly localized if the average distance between its endpoints (joints) and the ground-truth data is less than α times of the length of annotated endpoints in the ground-truth data.

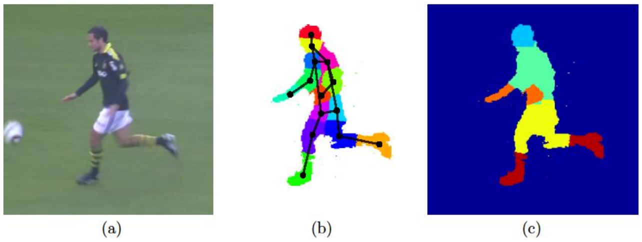


Figure 5. (a) Original image from KTH Football I dataset. (b) Pose and segmentation result of proposed method. (c) Different visualization of proposed method.

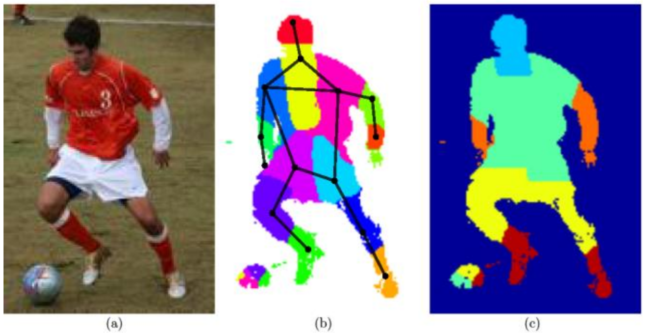


Figure 6. (a) Original image from Leeds Sports Pose dataset. (b) Pose and segmentation result of proposed method. (c) Different visualization of proposed method.

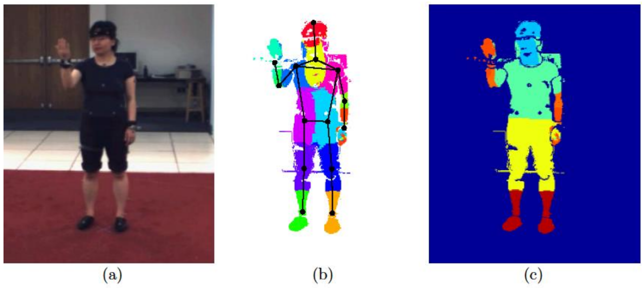


Figure 7. (a) Original image from HumanEva I dataset. (b) Pose and segmentation result of proposed method. (c) Different visualization of proposed method.

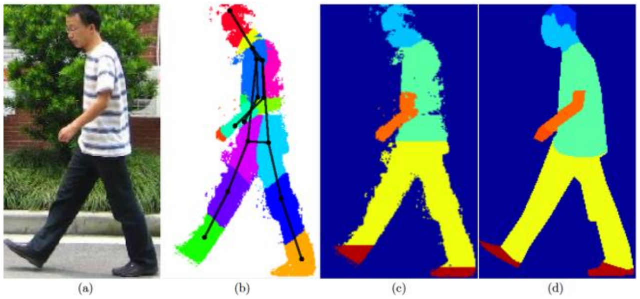


Figure 8. (a) Original image from Penn-Fudan dataset. (b) Pose and segmentation result of proposed method. (c) Different visualization of the proposed method. (d) Ground-truth.

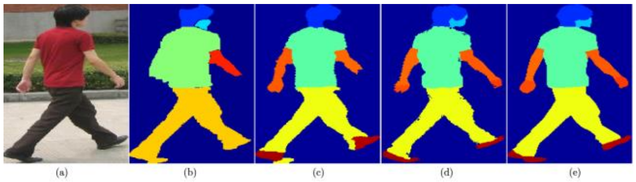


Figure 9. (a) Original image from Penn-Fudan dataset. (b) Output of Bo et al. [17] method. (c) Output of Xia [10] method. (d) Segmentation result of proposed method. (e) Ground-truth.

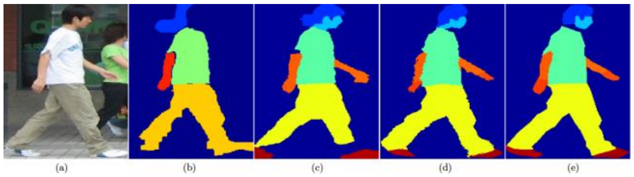


Figure 10. a) Original image from Penn-Fudan dataset. (b) Output of Bo [17] et al. method. (c) Output of Xia et al. [10] method. (d) Segmentation result of proposed method. (e) Ground-truth.

For the Penn-Fudan dataset there is a ground-truth segmentation for body part segmentation. Since the face and hair are segmented in two different classes in this dataset we used these data in training phase of the Body Field method. In fact the training phase is performed with one more extra class for this dataset. It shows the generalizability of the proposed method to the datasets that have more fine body parts segmented regions. We could define extra classes for each of the fine segmented regions and compute the mean expected difference vector between fine regions and other classes to use in training phase of the BodyField method. Quantitative results of the proposed method on the challenging KTH Football I datasets are summarized in Table 1. The pre-processing step, (FCRF), improves the results obtained by the DeeperCut method by up to 6%. Also DeeperCut [1] method is evaluated on this dataset and it has 85% total PCP. DeeperCut is a powerful body part detector. It uses integer linear programming for estimating the pose from the probability map. However, it sometimes fails to estimate the correct pose of player because of high degree of motion blur in images of KTH Football I dataset. The proposed BodyField method has 99% PCP and improves the results of the original DeeperCut method by 14%. Also the proposed BodyField method improves the results obtained by Kazemi *et al.*[9] in terms of PCP measure by up to 10% due to its better and refined HB part segments. In the Extended Leeds Sports Pose dataset, the standard *probability of correct key points* (PCK) evaluation metric is used [1], [2]. According to the definition in [31], a candidate key point is considered to be correct if it falls within $\beta \times max\,(h, w)$ pixels of the ground-truth key point, where h and w are the height and width of the bounding box of human respectively, and β controls the relative threshold for considering correctness. Results in Table 2 is based on Person-Centric ground-truth with $\beta = 0.2$. In Table 2, comparison results of the proposed method with the method of Chu *et al.*[2], Bulat *et al.*[3], Wei *et al.*[16], and Insafutdinov *et al.*[1] are presented. As it can be seen from Table 2, the original method of Insafutdinov *et al.*[1] has 90.1% PCK. The pre-processing step, FCRF, improves the PCK to 91.8%. The proposed BodyField method has 96.2% efficiency in terms of PCK measure. It outperforms the original method of Insafutdinov *et al.*[1] by 6.1%, and also the method of Chu *et al.*[2] by 3.6% in terms of PCK measure. For the HumanEva dataset, the standard method for computing the accuracy of pose estimation methods is the average 2D error [7]. The proposed method is evaluated on sequences 1 and 2 in walking, jogging and balance actions. As it can be seen in Table 3, the 2D error between the estimated pose and ground-truth location of joints is decreased by using the proposed BodyField method.

The overall average 2D error of Sigal *et al.*[7] is $10.7 \pm 1\,pix$, while it decreases to $5.4 \pm 1.2\,pix$ in pre-processing step, FCRF, and to $2.9 \pm 1\,pix$ in the proposed BodyField methods. Since the official evaluation server of the HumanEva dataset, http://humaneva.is.tue.mpg.de/, is currently out of service, we used the validation set for reporting the average values of 2D error.

Table 1. PCP values for KTH Football dataset (Observer-Centric,$\alpha = 0.5$*)*

Method	Torso	Upper Leg	Lowe Leg	Upper Arm	ForeArm	Head	Total
BodyField	**100**	**98.6**	**98.7**	**98.9**	**98.8**	**100**	**99**
FCRF(Pre-processing)	95	98	90	92	82	91	91
Kazemi *et al*. (RF) [9]	96	94	84	90	69	94	87
Kazemi *et al*.(RF+PosePrior) [9]	98	97	88	93	71	96	89
DeeperCut [1]	91	91	87	89	72	82	85
Belagiannis *et al*. [32]	98	92	80	88	57	86	84
Yang & Ramanan [31]	98	89	73	86	55	84	80

Table 2. PCK values for Extended Leeds Sports Pose dataset (Person-Centric,$\beta = 0.2$*)*

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
BodyField	98.2	96.8	97.5	93.7	85.9	96.3	95.6	96.2
Chu *et al*. [2]	98.1	93.7	89.3	86.9	93.4	94.0	92.5	92.6
FCRF (Pre-processing)	90.9	90.8	92.2	92.9	90.7	91.9	90.8	91.8
Bulat *et al*. [3]	97.2	92.1	88.1	85.2	92.2	91.4	88.7	90.7
Wei *et al*. [16]	97.8	92.5	87.0	83.9	91.5	90.8	89.9	90.5
DeeperCut [1]	97.4	92.7	87.5	84.4	91.5	89.9	87.2	90.1

Table 3.Average 2D error for HumanEva I dataset

Method	Sequence 1			Sequence 2			Overall
	Walk	Jog	Balance	Walk	Jog	Balance	
BodyField	3.9± 0.3 pix	2.9±0.6	2.2±1.4	2.8±1 pix	2.4±1.7pix	3.2±0.9pix	2.9±1pix
FCRF(Pre-processing)	4.1±0.4pix	5.6±0.8pix	5.4±1.6pix	3.9±1.2pix	7.2±2.1pix	5.9±1.2pix	5.4±1.2pix
Sigal[7]	10.1±0.9pix	11.3±0.7pix	11.3±2.3pix	7.9±0.6pix	12.4±2.3pix	10.9±2.8pix	10.7±1pix

Table 4. Comparison of our approach with other state-of-the-art methods on the Penn-Fudan benchmark dataset in terms of per-pixel accuracy (%). The Avg∗ means the average without shoes class since it was not reported in other methods.

Method	hair	face	u-cloth	arms	l-cloth	legs	shoes	Avg*
BodyField	63.4	61.7	79.8	58.4	82.3	65.0	47.2	65.4
AOG [10]	63.2	56.2	78.1	40.1	80.0	45.5	35.0	60.5
DDN [20]	43.2	57.1	77.5	27.4	75.3	52.3	------	56.2
SBP [17]	44.9	60.8	74.8	26.2	71.2	42.0	------	53.3
P&S [18]	40.0	42.8	75.2	24.7	73.0	46.6	------	50.4
Wang *et al*.[19]	48.7	49.1	70.2	33.9	69.6	29.9	36.1	50.2

The proposed method is evaluated on the popular Penn-Fudan benchmark [6], which consists of pedestrians in outdoor scenes with much pose variations. Labels of the dataset include 7 body parts namely hair, face, upper-clothes, lower-clothes, arms (arm skin), legs (leg skin), and shoes.

Also, since the proposed method segments the human body parts into 14 classes, we used the mapping process to convert the corresponding classes to those used in the Penn-Fudan dataset. For this conversion, the estimated pose is used as auxiliary information. The typical part segmentation results in these datasets are illustrated in Figure 8, Figure 9 and Figure 10. This dataset does not have the ground-truth of joints and therefore the results in pose estimation cannot be compared in this dataset in terms of PCP or PCK measures. But since for this dataset the ground truth for segmentation is available in pixel by pixel, the standard evaluation metric is used as per-pixel accuracy [10]. In Figure 9 and Figure 10 the comparison between the proposed method and the method of Xia *et al.* [10] and Bo *et al*. [17] are provided. For the method of Xia *et al*. [10] and Bo *et al*. [17] we used the source image provided by the authors. As shown in Table 4, the proposed method is compared with state-of-the-art methods, namely, AOG [10], DDN[20], P&S[18], SBP [17], and Wang *et al.* [19] on the Penn-Fudan dataset. The proposed BodyField method outperforms the DDN [20] method by over 9% and it has 4.9% improvement in comparison with the state-of-the-art method of Xia *et al*.[10] (AOG method). The improvement in the proposed method is due to the fact that estimated pose and corresponding body part segments are refined simultaneously. In other words, use of pose information in semantic human body part segmentation has increased the per-pixel accuracy. More visual output results of the inner steps of the BodyField method are available in http://ipl.ce.sharif.edu/bodyfield.html.

6- Conclusion

A new and efficient method for simultaneous single-view human body part segmentation and pose estimation is introduced that opens a new approach to the problem of structured semantic segmentation. A new energy function is introduced that encodes the spatial dependency between human body parts, in addition to the available segmentation constraints. In the proposed method, despite the fact that shifted Gaussian kernels are used, it is shown that finding the minimum of the proposed energy function is possible by applying an efficient mean field approximation process. Due to challenges such as occlusion and self-occlusion effects that occur frequently in human body pose data, the previous learning methods that only use the appearance model cannot converge to a proper pose estimation. That is because there are not enough evidences about the occluded and self-occluded parts available. The proposed BodyField method uses the probability map of the DeeperCut method to define a proper energy function with shifted Gaussian kernels between connected body parts. During the inference step, the evidence for occluded parts is refined by using the information of other parts that are connected in the human body skeleton model. Although shifted Gaussian kernels (in pairwise terms of the proposed energy function) add huge computational cost to the inference process, the problem is solved by proposing an efficient mean field approximation algorithm that speeds up message-passing steps, despite the fact that kernels are shifted.

For demonstrating the effectiveness of the proposed fully connected model in comparison with the state-of-the-art pose estimation methods, the probability maps of the DeeperCut method are used in the training phase and it is shown that results improve significantly in KTH Football I, LSP, HumanEva I in comparison with the original DeeperCut method. Also it is shown that the BodyField method has substantial improvement in HB segmentation in Penn-Fudan dataset in per pixel segmentation measure.

The experimental results on the challenging KTH Football I, Extended Leeds Sports Pose, HumanEva I, and Penn-Fudan datasets show the superiority of the proposed method over other existing methods in terms of PCP, PCK and per pixel segmentation accuracy.

References

[1] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "Deepercut: A deeper, stronger, and faster multi-person pose estimation model," in European Conference on Computer Vision, 2016: Springer, pp. 34-50.

[2] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1831-1840.

[3] A. Bulat and G. Tzimiropoulos, "Human pose estimation via convolutional part heatmap regression," in European Conference on Computer Vision, 2016: Springer, pp. 717-732.

[4] S. Ershadi-Nasab, S. Kasaei, and E. Sanaei, "Regression-based convolutional 3D pose estimation from single image," Electronics Letters, vol. 54, no. 5, pp. 292-293, 2018.

[5] S. E. Nasab, S. Kasaei, E. Sanaei, A. Ossia, and M. Mobini, "Multiview 3D reconstruction and human point cloud classification," in 2014 22nd Iranian Conference on Electrical Engineering (ICEE), 2014: IEEE, pp. 1119-1124.

[6] L. Wang, J. Shi, G. Song, and I.-f. Shen, "Object detection combining recognition and segmentation," in Asian conference on computer vision, 2007: Springer, pp. 189-199.

[7] L. Sigal, A. O. Balan, and M. J. Black, "Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion," International journal of computer vision, vol. 87, no. 1-2, p. 4, 2010.

[8] S. Johnson and M. Everingham, "Learning effective human pose estimation from inaccurate annotation," in CVPR 2011, 2011: IEEE, pp. 1465-1472.

[9] V. Kazemi, M. Burenius, H. Azizpour, and J. Sullivan, "Multi-view body part recognition with random forests," in 2013 24th British Machine Vision Conference, BMVC 2013; Bristol; United Kingdom; 9 September 2013 through 13 September 2013, 2013: British Machine Vision Association.

[10] F. Xia, J. Zhu, P. Wang, and A. L. Yuille, "Pose-guided human parsing by an and/or graph using pose-context features," in Thirtieth AAAI Conference on Artificial Intelligence, 2016.

[11] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in Advances in neural information processing systems, 2011, pp. 109-117.

[12] A. Adams, J. Baek, and M. A. Davis, "Fast high‐dimensional filtering using the permutohedral lattice," in Computer Graphics Forum, 2010, vol. 29, no. 2: Wiley Online Library, pp. 753-762.

[13] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," 2005.

[14] V. Vineet, G. Sheasby, J. Warrell, and P. H. Torr, "Posefield: An efficient mean-field based method for joint estimation of human pose, segmentation, and depth," in International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition, 2013: Springer, pp. 180-194.

[15] M. Kiefel and P. V. Gehler, "Human pose estimation with fields of parts," in European Conference on Computer Vision, 2014: Springer, pp. 331-346.

[16] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4724-4732.

[17] Y. Bo and C. C. Fowlkes, "Shape-based pedestrian parsing," in CVPR 2011, 2011: IEEE, pp. 2265-2272.

[18] I. Rauschert and R. T. Collins, "A generative model for simultaneous estimation of human body shape and pixel-level segmentation," in European Conference on Computer Vision, 2012: Springer, pp. 704-717.

[19] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. L. Yuille, "Joint object and part segmentation using deep learned potentials," in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1573-1581.

[20] P. Luo, X. Wang, and X. Tang, "Pedestrian parsing via deep decompositional network," in Proceedings of the IEEE international conference on computer vision, 2013, pp. 2648-2655.

[21] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 4, pp. 834-848, 2017.

[22] R. Alp Güler, N. Neverova, and I. Kokkinos, "Densepose: Dense human pose estimation in the wild," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7297-7306.

[23] Y. Chen, C. Shen, X.-S. Wei, L. Liu, and J. Yang, "Adversarial posenet: A structure-aware convolutional network for human pose estimation," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1212-1221.

[24] X. Peng, Z. Tang, F. Yang, R. S. Feris, and D. Metaxas, "Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2226-2234.

[25] W. Yang, W. Ouyang, X. Wang, J. Ren, H. Li, and X. Wang, "3d human pose estimation in the wild by adversarial learning," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5255-5264.

[26] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7103-7112.

[27] M. Andriluka et al., "Posetrack: A benchmark for human pose estimation and tracking," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5167-5176.

[28] J. M. Mooij, "libDAI: A free and open source C++ library for discrete approximate inference in graphical models," Journal of Machine Learning Research, vol. 11, no. Aug, pp. 2169-2173, 2010.

[29] A. Adams and J. Dolson, "ImageStack Library," https://github.com/abadams/ImageStack.

[30] J. Baek, A. Adams, and J. Dolson, "Lattice-based high-dimensional gaussian filtering and the permutohedral lattice," Journal of mathematical imaging and vision, vol. 46, no. 2, pp. 211-237, 2013.

[31] 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, 2012.

[32] V. Belagiannis, C. Amann, N. Navab, and S. Ilic, "Holistic human pose estimation with regression forests," in International Conference on Articulated Motion and Deformable Objects, 2014: Springer, pp. 20-30.

Sara Ershadi-Nasab was born in Mashhad, Iran, in 1986. She received the B.Sc. degree in Electrical Engineering from Ferdowsi University of Mashhad in 2008, her M.Sc. degree in 2010 from Amir Kabir University of Technology and her Ph.D. degree in 2017 at Sharif University of Technology. Her research interests are in 3D pose estimation, computer vision, and human activity recognition.

Shohreh Kasaei received the B.Sc. degree from the Department of Electrical and Computer Engineering, Isfahan University of Technology, Iran, in 1986, the M.Sc. degree from the Department of Electrical and Electronics Engineering, University of the Ryukyus, Japan, in 1994, and the Ph.D. degree from Signal Processing Research Centre, School of Electrical Engineering and Computer Science, Queensland University of Technology, Australia, in 1998. She joined Sharif University of Technology since 1999, where she is currently a full professor and the director of image processing laboratory (IPL). Her research interests include image and video processing as well as 3D computer vision with primary emphasis on 4D reconstruction and graphical element addition in dynamic sports scenes, human activity recognition, pose estimation, 4D object tracking, virtual reality, semantic scene understanding, 3D SLAM, image/video mosaicing, multi-resolution texture analysis, scalable video coding, image retrieval, video indexing, face recognition, hyperspectral change detection, video restoration, fingerprint authentication, and watermarking.

Esmaeil Sanaei received the B.Sc. degree in Electronics Engineering from the Department of Electrical Engineering, Amir Kabir University of Technology (Tehran Polytechnic), Iran, in 1979, the M.Sc. degree in Control Systems from the University of Technology of Compi`egne (Universit`e de Technologie de Compi`egne), Paris, France, in 1981, the M.Sc. degree in Information Technology Systems from the ` Ecole sup`erieure d'`electricit`e (Sup`elec), Paris, France, in 1982, and the Ph.D. degree in Information Technology Systems from University of Paris, Paris, France in 1984. He joined Sharif University of Technology since 1986, where he is currently an assistant professor. His research interests include computer vision and machine learning.

Erfan Noury was born in Urmia, Iran, in 1993. He received a B.Sc. degree at Computer department at Sharif University of Technology in 2017. His research interests include deep learning and computer vision.

Hassan Hafez-Kolahi was born in Mashhad, Iran, in 1989. He received the B.Sc. degree in Computer Engineering from Ferdowsi University of Mashhad in 2011, her M.Sc. degree in 2013 from Sharif University of Technology and he is currently a Ph.D. candidate at Sharif University of Technology. His research interests are in deep learning, computer vision.