PAPER

Occluded Appearance Modeling with Sample Weighting for Human Pose Estimation

Yuki KAWANA[†], ***Nonmember and*** **Norimichi UKITA**[†*a)], ***Senior Member***

SUMMARY This paper proposes a method for human pose estimation in still images. The proposed method achieves occlusion-aware appearance modeling. Appearance modeling with less accurate appearance data is problematic because it adversely affects the entire training process. The proposed method evaluates the effectiveness of mitigating the influence of occluded body parts in training sample images. In order to improve occlusion evaluation by a discriminatively-trained model, occlusion images are synthesized and employed with non-occlusion images for discriminative modeling. The score of this discriminative model is used for weighting each sample in the training process. Experimental results demonstrate that our approach improves the performance of human pose estimation in contrast to base models.

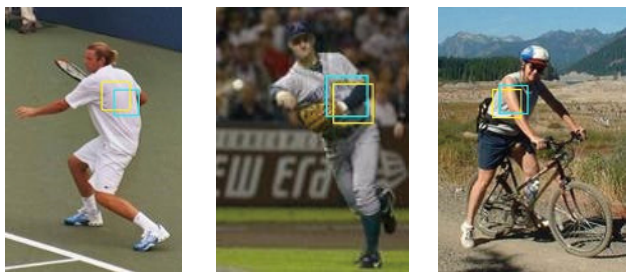
key words: *human pose estimation, pictorial structure models, occlusion*

1. Introduction

Human pose estimation is a task to infer the configuration of a person's body parts in an image. The task is a highly challenging problem due to a wide variety of appearance resulting from nonrigid deformation of human body, occlusion, and a variety of clothing.

We base our approach on the pictorial structure model (PSM) [1]–[3]. The PSM represents a human body configuration as a graphical tree model capturing inter-part spatial relationships such as relative position and orientation, and decomposes appearance of a human body into local part templates. It is important to model appearance of each body part robustly against inter-person difference. For this purpose, robust feature descriptors such as HOG [4] and PHOG [5] have been proposed. The feature representation is much improved by recent advance in convolutional neural networks [6].

However, it is difficult to represent the all appearance variation of body parts occluded by other body parts, other people, or background objects. Since a human body is highly articulated and a single image only represents a unidirectional view of the body, many body parts are often occluded in an image, as shown in Fig. 1 (a) (b) (c). This kind of occlusion is not handled by appearance modeling in the PSM, while the appearance features of body parts are not distinctive (e.g., left/right lower and upper arms could be represented as the similar figures of two parallel lines).



(a) (b) (c)

Fig. 1 Examples of occlusion. A torso covers an arm in (a). An arm covers a torso in (b) and (c).

To solve the problem above, this paper proposes a method for suppressing the bad effect of occluded parts in appearance learning by measuring the degree of occlusion. These issues have been researched in previous works such as [7], [8], but they require annotation data of occluded body parts for prior learning of an occlusion detector. However, usually-available datasets have no annotation about whether or not each part is occluded. Manual annotation of occluded parts for all training images is a highly expensive option. In addition, those previous methods [7], [8] do not use a sample image if this image includes at least one occluded part. This means that useful training data of other non-occluded body parts in this image are also not used.

Our approach enhances appearance modeling of a local part template by automatically weighting sample images based on occlusion. The overview of the proposed method is shown in Fig. 2. The proposed method employs conventional pose-annotated training data with no annotation of occluded parts (e.g., "Sample images with parts annotation" in Fig. 2). For sample image weighting based on occlusion, we introduce the occlusion confidence model which detects possible occlusion on a body part and measures its degree of occlusion. For robustly measuring the degree of occlusion, various sorts of sample images for non-occluded and occluded body parts, which are respectively indicated by "Body part images without occlusion" and "Body part images with occlusion" in Fig. 2, are useful. In general datasets, however, the variation of occluded parts is limited in contrast to that of non-occluded parts. To resolve this problem, a large number of occluded body part images are synthesized from non-occluded ones; "Occlusion image

Manuscript received March 13, 2017.
Manuscript revised May 22, 2017.
Manuscript publicized July 6, 2017.

[†]The authors are with Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma-shi, 630–0192 Japan.

[*]Presently, with Graduate School of Engineering, Toyota Technological Institute.

a) E-mail: ukita@toyota-ti.ac.jp

DOI: 10.1587/transinf.2017EDP7088

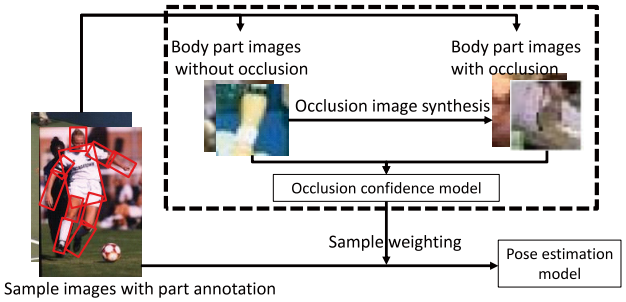


Fig. 2 Illustration of our proposed approach. From sample images with part annotation, part windows are cropped and divided into non-occluded and occluded parts. These part windows are employed for learning the occlusion confidence model. This model weights each sample image in learning the pose estimation model. Processes enclosed by a dashed rectangle are executed for each body part.

synthesis" in Fig. 2.

We examine our approach on benchmark datasets, the Image Parse [9] and Leeds Sports Pose [10] datasets. Experimental results demonstrate that our approach outperforms the base methods [1], [11]. While our proposed method improves robustness against occlusion in a learning stage of each base method, its pose inference process is performed with no change in our experiments.

In contrast to our earlier work [12], [13], this paper presents additional experimental results with a larger dataset (i.e., Leeds Sports Pose [10]) and a pose estimation method using deep neural networks [11].

2. Related Work

For pose estimation in still images, a graphical model has been used to learn the distribution of human poses in recent works [14]–[16]. In the graphical model, each node and link corresponds to a part and a physical connection between parts. Among all possible models, tree-based models [17], [18], including pictorial structure models (PSMs) [2], [19], are widely used because of their abilities to represent a variety of shape structures and to obtain globally-optimal part locations. Indeed, a variety of objects, including human bodies, cars, bicycles, horses, etc., are expressed by the tree-based models [20].

While the PSM consists of appearance and body-structure components as described in Sect. 3, this paper focuses on the appearance model. To maintain discriminativity in appearance, recent advances have proven that discriminative training of part appearance can improve part distinguishability [17], [20], [21]. For appearance representation, a variety of features have been proposed such as image gradient features [22] and segmentation features [23]. In addition to the appearance of each body part, that of connectivity of neighboring body parts can be also discriminatively trained [11], [24]. For utilizing the advantage of discriminative training, training samples must be correctly divided into positive and negative samples. The positive and negative samples are each body part and all other images, respectively, in the PSM. For PSM training, selection of the positive and negative samples can be basically done by using the body part annotations provided in each training image. However, if a body part is occluded by other body parts or background objects, the region of this body part should be removed from the positive samples. The sample selection is critical because incorrect selection may lead to bad learning results as suggested in [20]. The research on effect of occlusion in appearance modeling has not been conducted, whereas a problem relating to occlusion at a pose inference stage has been researched in previous methods [7], [8].

Another approach regarding occlusion is proposed by Johnson and Everingham in [25]. They use the dataset which has (incomplete) occlusion annotations where occluded body parts are not annotated. Their approach discards a sample in appearance modeling if one or more body parts are not annotated in a sample due to occlusion.

In this paper, based on the approach in [1], we explicitly examine the effect of occlusion in appearance modeling. In our model we weight a sample containing an occluded body part in order to mitigate the adverse effect of occlusion. The adverse effect of occlusion is partly suppressed in appearance modeling proposed in [25]. In [25], training data including occluded body parts is not used even if other body parts are visible and can be used for appearance modeling. The proposed approach is more efficient than [25] in a sense that we fully utilize all training data, which are made with high annotation cost.

3. Pictorial Structure Model

This section describes the basis of the PSM [2] used in the base models [1], [11] of our implementation. A tree-based model is defined by a set of body parts V and a set of links E connecting two of the body parts. Given N body parts, hypothesis $z = (p_1, \ldots, p_N)$ specifies the locations of all parts, where p_i represents the pixel location, orientation, and scale of part i.

The score of hypothesis z in image I is given by a sum of two scores as follows:

$$S_a(z) + S_d(z) \tag{1}$$

$$= \sum_{i \in V} w_i \cdot \phi(I, p_i) + \sum_{i,j \in E} w_{ij} \cdot \phi_d(p_i - p_j) \tag{2}$$

$$= \beta \cdot \psi(I, z) \tag{3}$$

$$\beta = (w_1, \ldots, w_N, w_{11}, \ldots, w_{NN}) \tag{4}$$

$$\psi(I, z) = (\phi(I, p_1), \ldots, \phi(I, p_N), \phi_d(p_1 - p_1), \ldots, \phi_d(p_N - p_N)) . \tag{5}$$

where the first term in score (1) represents the appearance score and the second term for the deformation score. In the appearance score, w_i represents a filter for body part i and $\phi(I, p_i)$ is a feature vector (e.g., HOG descriptor [4]) extracted from pixel location p_i in I. In a typical example of the deformation score, w_{ij} represents a four dimensional vector specifying coefficients of quadratic function [1]. $\phi_d(p_i - p_j)$ defines the deformation between body

parts i and j. By integrating the first and second terms in (2), the score of hypothesis z can be expressed by the inner product of β and $\psi(I, z)$ as shown in (3), (4), and (5).

4. Appearance Modeling with Occlusion Confidence

This section introduces our approach of appearance modeling based on measuring the degree of occlusion of each body part. Our method consists of three tasks below:

Section 4.1 Occlusion confidence modeling, which is achieved with a set of occluded body part images and a set of non-occluded ones. This modeling aims to measure the degree of occlusion of each body part.

Section 4.2 Data synthesis for occluded body part images, where the goal is to synthesize occluded body part images from non-occluded ones. The synthesized images allow us to train the occlusion confidence model without manually annotating occlusion for thousands of sample images.

Section 4.3 Weighting samples for appearance modeling, where a sample image which has more degree of occlusion less effects the learning process.

4.1 Occlusion Confidence Model

We aim to evaluate the probability of body part i being not occluded. If the local part filter of body part i is applied to an image region in which body part i is observed with no occlusion, the response value tends to be larger than the value in the region of occluded body part i. This tendency is exploited for representing an occlusion confidence model. The occlusion confidence model derives the degree of occlusion of a body part based on its appearance. Remember that the appearance score is denoted by $w_i \cdot \phi(I, p_i)$ in Eq. (2). Let $P_i(\overline{o}|w_i \cdot \phi(I, p_i))$ be a conditional probability that body part i is not occluded in $p_i \in z$ of image I when the appearance score is $w_i \cdot \phi(I, p_i)$. $P_i(\overline{o}|w_i \cdot \phi(I, p_i))$ is expressed as follows:

$$P_i(\overline{o}|w_i \cdot \phi(I, p_i)) = \begin{cases} f(a), & (O(i) \neq \emptyset) \\ 1, & \text{otherwise} \end{cases} \tag{6}$$

$$f(a) = 1 - \frac{1}{1 + \exp((a - 0.75)^{20})} \tag{7}$$

$$j \in O(i) \quad \text{if} \quad \frac{|D(i) \cap D(j)|}{|D(i)|} > \gamma \tag{8}$$

$$a = \frac{G(w_i \cdot \phi(I, p_i)|\mu_i^{pos}, \Sigma_i^{pos}) P_i(\overline{o})}{G(w_i \cdot \phi(I, p_i)|\mu_{i,O(i)}^{neg}, \Sigma_{i,O(i)}^{neg}) P_i(o)} \tag{9}$$

- o and $\overline{o}$ denote that a body part is occluded and is not occluded, respectively.
- a becomes greater if the probability that body part i is not occluded is larger.
- $f(a)$ is an arbitrary function to adjust the reliability of occlusion confidence modeling using appearance cues (i.e., $\phi(I, p_i)$ in Eq. (6)). In our experiments, $f(a)$ was designed as Eq. (7) so that $f(a)$ is almost fixed around 0.5, if a is not sufficiently high. $f(a) \simeq 0.5$ means that it is difficult or impossible to decide whether i is visible or occluded. So, $f(a)$ is just a middle point between 0 (i.e., occluded) and 1 (i.e., visible). When a is high, $f(a) \simeq 1$. The curve of $f(a)$ is shown in Fig. 3. $f(a)$ increases sharply from 0.5 to 1.0 because the appearance score (i.e., $w_i \cdot \phi(I, p_i)$ in (2)) of an occluded body part increases sharply when the visible region of part i is above a certain ratio. Any similar-shaped curves work well, and our $f(a)$ is robust to a change in parameters (i.e., 0.75 and 20 in Eq. (2)). While the parameters were determined based on cross-validation trials using the training dataset of the Image Parse dataset, they can be determined depending the dataset.
- $O(i)$ denotes a set of the indices of body parts that overlap with body part i. If the overlap between the regions (i.e., pixels) of body parts i and j is greater than a threshold, γ, body part j is included in $O(i)$. More specifically, body part j is included in $O(i)$ if inequality (8) is satisfied. In (8), $D(i)$ and $|D(i)|$ denote the set of pixels included in the region of body part i and the number of pixels in $D(i)$, respectively. The graphical idea of (8) is illustrated in Fig. 4. $\gamma = 0.25$ in our experiments.
- $G(w_i \cdot \phi(I, p_i)|\mu_i^{pos}, \Sigma_i^{pos})$ is the Gaussian distribution of the appearance score of non-occluded body part i. The mean and variance of the Gaussian distribution are denoted by μ_i^{pos} and Σ_i^{pos}, respectively. The Gaussian distribution for occluded body part i is expressed as $G(w_i \cdot \phi(I, p_i)|\mu_{i,O(i)}^{neg}, \Sigma_{i,O(i)}^{neg})$.
- Let $P_i(\overline{o})$ and $P_i(o) = 1 - P_i(\overline{o})$ be the probability values of body part i being not occluded and being occluded, respectively. We assume that body part i is likely to

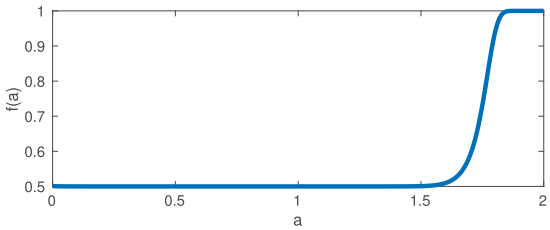


Fig. 3 Function $f(a)$. The horizontal and vertical axes indicate a and $f(a)$.

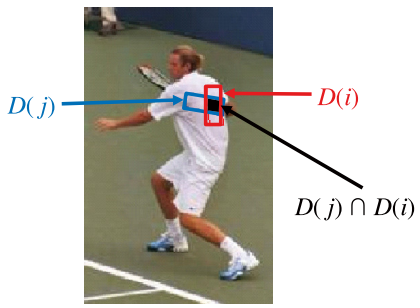


Fig. 4 Illustration of $D(i)$, which indicates the region of body part i.

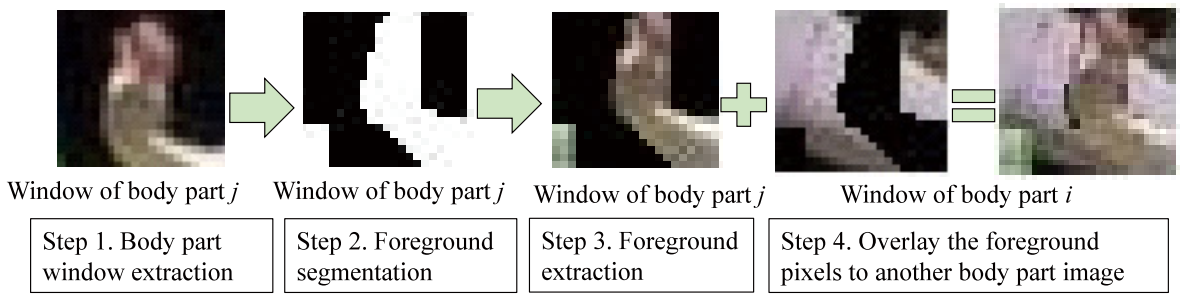


Fig. 5 Negative image synthesis for scoring the degree of occlusion. Here, positive and negative images are those with non-occluded and occluded body parts, respectively. Negative samples are synthesized by overlaying a body part (i.e., body part j in the figure) onto the window of another body part (i.e., body part i in the figure).

be not occluded if body parts overlapped with i (i.e., $j \in O(i)$) have overlaps with less number of other body parts. We formulate this idea as Eq. (10) that is required to compute Eq. (9).

$$P_i(\overline{o}) = \begin{cases} \frac{|O(i)|}{|O(i)|+\sum_{j \in O(i)} |O(j)|} & |O(i)| \neq 0 \\ 1, & \text{otherwise} \end{cases} \quad (10)$$

4.2 Occluded Sample Data Synthesis

The occlusion confidence model utilizes several stochastic elements (e.g., prior probability, $P_i(o)$ and $P_i(\overline{o})$, and the distribution of the appearance score, $G(w_i \cdot \phi(I, p_i)|\mu_i^{pos}, \Sigma_i^{pos})$). Among all, the appearance score, $w_i \cdot \phi(I, p_i)$, is the most important and primitive information for reliability of the occlusion confidence model. In order to make this score more reliable, we need more sample images of body parts being not occluded as positive samples and those being occluded as negative samples. In contrast to the positive samples, the negative samples are short on numbers for robust discriminative learning due to the following reasons:

- In general, the number of occluded body parts is fewer than that of non-occluded body parts as seen in Figs. 1 and 4.
- In contrast to a body part observed with no occlusion, the appearance of occluded body parts is versatile because of the variation of occluding body parts.

To augment the negative samples, images of body part i being occluded by body part j are synthesized. The negative samples are synthesized so that the region of body part j in the sample image is cropped and overlaid onto the region of body part i. Note that the synthesized negative samples are used only for training w_i used in Eq. (9). All other stochastic elements used in occlusion confidence modeling (e.g., $P_i(o)$, $P_i(\overline{o})$, and $G(w_i \cdot \phi(I, p_i)|\mu_i^{pos}, \Sigma_i^{pos})$) are derived only from real training samples.

The procedure of synthesizing negative samples is summarized as follow (Fig. 5):

Step 1 Crop the image window of body part j.
Step 2 Superpixelize the image window of body part j. In our experiments, superpixelization is achieved by an off-the-shelf method [26]. Then label each superpixel either the foreground (i.e., a body part) or the background. This labeling is achieved by employing the body part annotation as described below in detail.
Step 3 Crop foreground superpixels as a set of RGB pixels.
Step 4 Overlay them to the region of body part i. The position and orientation of overlaid body part j are determined randomly in order to represent various kinds of occlusions by another person as well as self occlusion. Note that body parts i and j can be extracted from different images.

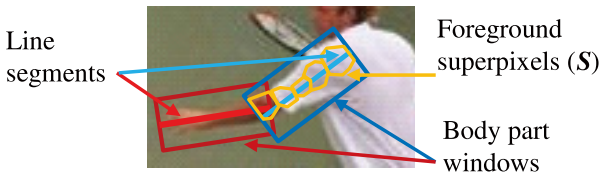


Fig. 6 Line segments of body parts in human pose annotation and superpixels on the line segments.

The foreground/background labeling process in step 2 is performed based on a probability that each pixel is included in the foreground region (denoted by F in what follows). With the annotation of body parts, a conditional probability that a pixel is included in F is expressed in the Bayesian manner as follows:

$$p(F|x, d) = \frac{p(x, d|F)p(F)}{p(x, d)} = \frac{p(x|F)p(d|F)p(F)}{p(x, d)}, \quad (11)$$

where x and d are the properties of the pixel. x denotes the RGB value of the pixel. Let $\boldsymbol{d}$ be a set of distance values from the pixel to the line segments of all body parts; the line segments of body parts are illustrated in Fig. 6. d is the shortest distance included in $\boldsymbol{d}$. d is normalized by the size of the image window of each body part, which is indicated by a rectangle in Fig. 6. More specifically, d is normalized so that $d = \frac{\hat{d}}{s_w}$ where $\hat{d}$ and s_w denote the pre-normalized value of d and the window size, respectively. Let $\boldsymbol{S}$ denote a set of superpixels each of which is on the line segment of each body part (i.e., superpixels in the foreground), as

Table 1 Average number of occluded body parts on each body region. All occluded parts in all training images (i.e., 100 images in the Image Parse dataset) were counted manually.

	Torso	Head	U.arms	L.arms	U.legs	L.legs
Number of occlusion per each body part	10.75	2	11.75	10.75	9.25	6

shown in Fig. 6. That is, all superpixels in $\boldsymbol{S}$ are assumed to be in F. $p(x|F)$ and $p(d|F)$ are computed from a mixture Gaussian distribution over x and d of all pixels included in $\boldsymbol{S}$, respectively. We regard $p(x, d|F)$ to be conditionally independent over F. This is because, on the same body part of different images, pixels tend to share relatively similar color distribution, and the distribution of normalized width of a body part is also consistent. We apply Eq. (11) to all pixels in a sample image to derive the heat map of per pixel likelihood of being foreground. Then we calculate the mean likelihood per superpixel and select superpixels each of whose mean likelihood is above a threshold as the foreground. In actual calculation, we use $p(F|x, d) \propto p(x|F)p(d|F)$ for calculation efficiency by assuming $\frac{p(F)}{p(x,d)}$ is constant compared to $p(x|F)p(d|F)$. This assumption is satisfied when (i) the size of a human body is almost fixed in human-cropped images and (ii) a color variation in a human body is smaller than that in background regions.

4.3 Sample Weighting with Occlusion Confidence

Our base models [1] and [11] learn the parameters of the PSM by a discriminative manner, namely by using the Latent SVM [20] and the Structured SVM [27], respectively. In both methods, the following cost function is minimized with labeled samples $(< I_1, y_1 >, \cdots, < I_B, y_B >)$, where $y_s \in \{-1, 1\}$ for optimizing a set of parameters, β in Eq. (4):

$$\arg\min_{\beta} \frac{1}{2}\|\beta\| + \sum_{s} C(I_s) \max\left(0, 1 - y_s f_\beta(I_s)\right), \quad (12)$$

where $f_\beta(I_s)$ denotes the score of the SVM with sample image I_s. $C(I_s)$ is a weighting function for I_s. In the base models [1], [11], $C(I_s)$ is constant over all sample images.

In our proposed method, the influence of each sample image, I_s, is adjusted by weighting $C(I_s)$. $C(I_s)$ is determined by the occlusion confidence values, Eq. (6), of all body parts of a target person in I_s. In our experiments, the following weighting function was used:

$$C(I_s) = \frac{1}{|V|} \sum_{i \in V} P_i\left(\overline{o}|w_i \cdot \phi(I_s, p_i)\right) \quad (13)$$

In Eq. (13), the weight is a simple average over the occlusion confidence values of all body parts. In our sample weighting strategy, a sample having more occluded body parts influences less in learning. In this strategy, not only the appearance score but also the deformation score in Eq. (1) are affected in the learning process. This is justifiable because occluded body parts are mostly invisible thus the deformation information from the annotation can be arguably noisy and less preferable to be used for learning.

Note that our proposed method is identical to the base methods [1], [11] if $C(I_s)$ is constant over all sample images. $C(I_s)$ is constant if $\gamma = 1$ as defined in Eqs. (6) and (8).

5. Experimental Results

This section reports the results of our experiments on the proposed method using the Image Parse dataset [9] and the Leeds Sports Pose dataset [10]. The Image Parse dataset contains 305 images with pose annotation in total. First 100 images are for training and the rest of 205 images are for testing. The Leeds Sports Pose dataset contains 2000 images with pose annotation in total. First 1000 images are for training and the rest of 1000 images are for testing.

The effect of our proposed confidence modeling is validated with two base models [1], [11]† for human pose estimation.

First of all, our proposed model is validated with a base model proposed in [1]. In accordance with this base model [1], a full-body skeleton model with 26 body ***parts*** were used for covering 10 body ***regions***, which are the torso, head, two upper arms, two lower arms, two upper legs, and two lower legs; 2 body parts for the head, 8 for the torso, and 2 for each of the upper and lower arms and legs. In this approach, the model parameter β in Eq. (4) was learned in coordinate descent manner alternating between selection of the body part locations $z = (p_0, \ldots, p_n)$ which maximizes $\beta \cdot \psi(I, z)$ in Eq. (3) and optimizing β given z.

For validating the ability of detecting occlusion, we manually gave the annotation of occlusion to all body parts in the Image Parse dataset. The averaged number of occlusion in each body part is shown in Table 1. The number is averaged over the number of body parts in each body region††.

The ability of detecting occlusion is evaluated with two criteria, namely precision and recall, explained in what follows. In the proposed method, the occlusion confidence model is required to remove occluded parts as many as possible for reducing their negative effect in appearance modeling. Since occluded parts are regarded as negative samples in occlusion confidence modeling, the precision rate should be higher.

$$\text{Precision} = \frac{TP}{TP + FP}, \quad (14)$$

where TP and FP denotes the number of correctly detected non-occluded body regions and falsely detected occluded

†The codes of these two base models are publicly available.

††For example, in four body parts of the upper arms, 12, 12, 12, and 11 parts are occluded among all training images, respectively. So $\frac{12+12+12+11}{4} = 11.75$ is the average number of occluded parts.

Table 2 Accuracy in occlusion confidence modeling.

	Torso	Head	U.arms	L.arms	U.legs	L.legs	mean
Precision	0.79	1.0	0.78	0.70	0.78	0.81	0.79
Recall	0.66	1.0	0.77	0.49	0.70	0.54	0.66

Table 3 Comparison of PCP in the Image Parse dataset. (a) Johnson [25], (b) Base method [1], and (c) Ours.

	Torso	Head	U.arms	L.arms	U.legs	L.legs	mean
(a) Johnson [25]	87.6	76.8	74.7	67.1	67.4	45.9	67.4
(b) Base method [1]	85.4	84.4	70.8	47.8	77.8	71.2	70.5
(c) Ours	87.3	86.3	73.2	55.1	79.0	71.2	73.1

Table 4 Comparison of PCP in Leeds Sports Pose dataset. (a) Base method [1], and (b) Our proposed method using [1].

	Torso	Head	U.arms	L.arms	U.legs	L.legs	mean
(a) Base method [1]	79.3	77.7	60.6	51.2	48.6	32.8	54.3
(b) Ours	79.6	77.8	61.1	51.2	49.4	34.0	54.9

body regions, respectively.

On the other hand, the occlusion confidence model is also required not to falsely remove non-occluded body parts. To evaluate whether or not this requirement is fulfilled, the recall rate of occlusion detection is an important criterion:

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{15}$$

where FN denotes the number of falsely removed non-occluded body regions.

Here we say an occluded body part is detected if the probability of the body part being not occluded as in Eq. (6) is less than one. In the dataset mentioned above, the precision and recall rates with Eq. (6) were shown in Table 2. While the recall rate is not high enough yet, the precision rate is relatively higher. In the trade-off between the precision and recall rates, the precision should have higher priority for suppressing the negative impact of occluded parts on learning part appearance. While efficiency of model learning is damaged if the recall rate is lower, this problem can be resolved by augmenting the number of sample images.

As described at the end of Sect. 4.3, γ in Eq. (8) changes the behavior of the proposed method. The effect of γ on pose estimation accuracy was tested on the Image Parse dataset. The accuracy is evaluated with a standard criterion, namely probability of a correct pose (PCP) criterion with Buffy implementation [28][†] using the person-centric annotation. Figure 9 shows the average PCP of all body regions obtained by the different values of γ. While the accuracy is changed smoothly, $\gamma = 0.25$ was selected based on Fig. 9 and used for further experiments including those with the Leeds Sports Pose dataset.

Next, the results of human pose estimation were evaluated with PCP and compared between different methods. The results tested in the Image Parse dataset and the Leeds Sports Pose dataset are shown in Table 3 and Table 4, respectively. In PCP tested with the Image Parse dataset, our approach gives superior performance to the base method in most of body parts. This is likely because our approach successfully learns appearance model which are less distracted by the appearance of occluded body parts appearance. This is especially true in lower and upper arms whose scores have significantly increased by 7.3% and 2.4% respectively in our approach. Performance improvement on the arms is likely because lower and upper arms have a number of occluded parts, thus our method is especially effective reducing the negative effect of occlusion. The typical examples of human poses estimated by our model from the Image Parse dataset are shown in Fig. 7. In the figure, we can see our model performs better to estimate the locations of body parts than the base model.

In the Leeds Sports Pose dataset also, our model performs better than the base method in most body parts as shown in Table 4. The graphical estimated results in the Leeds Sports Pose dataset are illustrated in Fig. 8.

Comparative experiments were conducted also with another base method using convolutional neural networks for appearance learning [11]. The implementation details were equal to the base model [11] as follows. A full-body graph structure is based on a human body annotation, in which each body consists of 10 body regions and 14 end points of the body parts, in the LSP dataset. These end points are connected by 12 midway points. In total, 26 nodes corresponding to these points compose a graph that represents the full body. As described in Sect. 4.3, the structured SVM [27] was used for optimizing model parameters.

The results of PCP evaluation are shown in Table 5. While the performance of this base method [11] is significantly higher than [1], the proposed method outperforms the base method [11] again, in particular in the lower limbs. This property is same with those observed in Tables 3 and 4. Several estimated poses are visualized in Fig. 10.

[†]The implemented code is distributed at the author's website [29].

Fig. 7 Typical results of our model compared to the base model [1] with the Image Parse dataset. The results of our model are shown on the left and the base model on the right in the each pair of images. A caption below each image indicates how many body parts were successfully localized out of all 10 body regions.

Fig. 8 Typical results of our model compared to the base model [1] with the Leeds Sports Pose dataset.

Table 5 Comparison of PCP in Leeds Sports Pose dataset. (a) Base method using CNN for appearance learning [11] and (b) Our proposed method using [11].

	Torso	Head	U.arms	L.arms	U.legs	L.legs	mean
(a) Base method [11]	92.7	87.8	69.2	55.4	82.9	77.0	75.0
(b) Ours	93.5	87.8	70.7	61.2	84.0	79.5	77.2

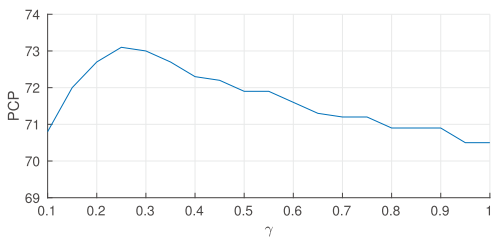


Fig. 9 Effect of γ on pose estimation accuracy.

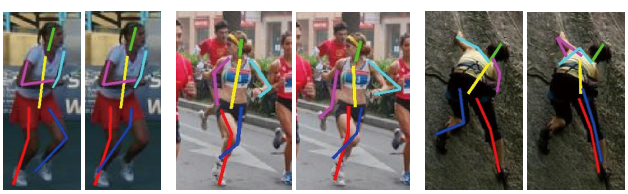


Fig. 10 Typical results of our model compared to the base method using CNN for appearance learning [11] with the Leeds Sports Pose dataset. In each sample image, the lefthand and righthand images show the results of our model and the based model [11].

6. Conclusion

We have introduced the model that improves appearance modeling for the PSM. We show that reducing the effect of occluded body parts can provide better appearance modeling to improve the performance in human pose estimation. For appearance modeling, our approach is more efficient than the previous work [25] that discards sample images including occluded body parts. We have shown that our method leads to superior results than the base methods [1], [11].

In future work, we would like to drop the false detection rate in sample image weighting. For example, while simple averaging in Eq. (13) is used in the current implementation, another approach that weights body parts depending on their occlusion probability may be useful. The occlusion confidence model can be also improved, for example, by a theoretically-valid function $f(a)$, while $f(a)$ was determined empirically. While the proposed method was evaluated with human pose estimation methods using the PSM, pose-regression and part-heatmap based methods are also prospective options as their performance is improved by convolutional neural networks (e.g., [30]–[33]). For the pose-regression based method also, the proposed method is applicable for reducing the bad effect of occluded body parts in appearance learning.

This work was partly supported by JSPS KAKENHI Grant Number 15H01583.

References

[1] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," CVPR, pp.1385–1392, 2011.
[2] 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.
[3] M.A. Fischler and R.A. Elschlager, "The representation and matching of pictorial structures," IEEE Transactions on computers, vol.C-22, no.1, pp.67–92, 1973.
[4] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," CVPR, pp.886–893, IEEE, 2005.
[5] A. Bosch, A. Zisserman, and X. Munoz, "Representing shape with a spatial pyramid kernel," Proceedings of the 6th ACM international conference on Image and video retrieval, pp.401–408, ACM, 2007.
[6] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "Imagenet classification with deep convolutional neural networks," NIPS, vol.60, no.6, pp.84–90, 2017.
[7] I. Radwan, A. Dhall, J. Joshi, and R. Goecke, "Regression based pose estimation with automatic occlusion detection and rectification," ICME, pp.121–127, 2012.
[8] L. Sigal and M.J. Black, "Measure locally, reason globally: Occlusion-sensitive articulated pose estimation," CVPR, pp.2041–2048, 2006.
[9] D. Ramanan, "Learning to parse images of articulated bodies," NIPS, pp.1129–1136, 2006.
[10] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation," BMVC, pp.1–11, 2010.
[11] X. Chen and A. Yuille, "Articulated pose estimation by a graphical model with image dependent pairwise relations," NIPS, pp.1736–1744, 2014.
[12] Y. Kawana, N. Ukita, and N. Hagita, "Occlusion-free appearance modeling of body parts for human pose estimation," International Conference on Machine Vision Applications, pp.321–324, 2015.
[13] Y. Kawana and N. Ukita, "Occlusion-robust model learning for human pose estimation," Asian Conerence on Pattern Recognition, pp.494–498, 2015.
[14] V. Ramakrishna, D. Munoz, M. Hebert, J.A. Bagnell, and Y. Sheikh, "Pose machines: Articulated pose estimation via inference machines," ECCV, vol.8690, pp.33–47, 2014.
[15] L.D. Bourdev, S. Maji, T. Brox, and J. Malik, "Detecting people using mutually consistent poselet activations," ECCV, vol.6316, pp.168–181, 2010.
[16] J. Puwein, L. Ballan, R. Ziegler, and M. Pollefeys, "Foreground consistent human pose estimation using branch and bound," ECCV, vol.8693, pp.315–330, 2014.
[17] R. Ronfard, C. Schmid, and B. Triggs, "Learning to parse pictures of people," ECCV, vol.2353, pp.700–714, 2002.
[18] G. Hua, M.-H. Yang, and Y. Wu, "Learning to estimate human pose with data driven belief propagation," CVPR, pp.747–754, 2005.
[19] P.F. Felzenszwalb and D.P. Huttenlocher, "Efficient matching of pictorial structures," CVPR, pp.66–73, 2000.
[20] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol.32, no.9, pp.1627–1645, 2010.
[21] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," CVPR, pp.1014–1021, 2009.
[22] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," CVPR, pp.886–893, 2005.
[23] N. Ukita, "Part-segment features with optimized shape priors for articulated pose estimation," IEICE Transactions, vol.E99-D, no.1, pp.248–256, 2016.
[24] N. Ukita, "Articulated pose estimation with parts connectivity using discriminative local oriented contours," CVPR, pp.3154–3161, 2012.
[25] S. Johnson and M. Everingham, "Learning effective human pose estimation from inaccurate annotation," CVPR, pp.1465–1472, 2011.
[26] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol.34, no.11, pp.2274–2282, 2012.
[27] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun, "Support vector machine learning for interdependent and structured output spaces," ICML, p.104, 2004.
[28] V. Ferrari, M. Marín-Jiménez, and A. Zisserman, "Progressive search space reduction for human pose estimation," CVPR, pp.1–8, 2008.
[29] V. Ferrari, M. Marin-Jimenez, and A. Zisserman, "Buffy stickmen v3.01 annotated data and evaluation routines for 2d human pose estimation." http://www.robots.ox.ac.uk/~vgg/data/stickmen/
[30] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," CVPR, pp.1653–1660, 2014.
[31] T. Pfister, J. Charles, and A. Zisserman, "Flowing convnets for human pose estimation in videos," ICCV, pp.1913–1921, 2015.
[32] S.E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," CVPR, pp.4724–4732, 2016.
[33] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," ECCV, vol.9912, pp.483–499, 2016.

Yuki Kawana He got the M.E. degree from Nara Institute of Science and Technology. His research themes were human pose estimation and deep neural networks.

Norimichi Ukita He received the B.E. and M.E. degrees in information engineering from Okayama University, Japan, in 1996 and 1998, respectively, and the Ph.D degree in Informatics from Kyoto University, Japan, in 2001. After working as an assistant professor at Nara Institute of Science and Technology (NAIST), he became an associate professor in 2007. He is a professor at Toyota Technological Institute, Japan (TTI-J), since 2016. He was a research scientist of PRESTO, Japan Science and Technology Agency (JST) from 2002 to 2006, and a visiting research scientist at the Robotics Institute, Carnegie Mellon University from 2007 to 2009. His main research interests are multi-object tracking and human pose estimation. He has received the best paper award from the IEICE in 1999.