Occlusion-Robust Model Learning for Human Pose Estimation

Yuki Kawana
Nara Institute of Science and Technology
kawana.yuki.kt7@is.naist.jp

Norimichi Ukita
Nara Institute of Science and Technology
ukita@is.naist.jp

Abstract

In this paper we examine the efficacy of self-occlusion-aware appearance learning for the part based model. Appearance modeling with less accurate appearance data is problematic because it adversely affects entire learning process. We evaluate the effectiveness of mitigating the influence of self-occluded body parts to be modeled for better appearance modeling process. To meet this end, We introduce an effective method for scoring degree of self-occlusion and we employ an approach learning a sample proportionally weighted to the score. We present our approach improves the performance of human pose estimation.

1. Introduction

Human pose estimation is a task to infer the configuration of a person's body parts in an image. The leftmost picture of Fig. 1 shows the inferred body configuration. We base our approach on the improved pictorial structured model (PSM) [19, 7, 11]. The PSM represents human body configuration as a graphical tree model capturing inter-part spatial relationships such as relative position and orientation, and decomposes appearance of human body into local part templates.

For appearance modeling, it is not feasible to accurately learn appearance of each body part with including occluded body parts altogether. Many body parts often result in self-occlusion in a natural image such as one body part covers another one, as shown in of Fig. 1 (b) (c). Since the feature of each body part would not have strong distinctive characteristic (e.g. feature of lower and upper arms could be both represented as similar figures of two parallel lines), appearance modeling for a local part template should be done carefully.

To solve the problem above, we need to mitigate the influence of occluded body part for appearance learning. Therefore detection and measuring degree of self-occlusion are necessary. Detection and measuring degree of self-occlusion have been researched in previous works such as [15, 18] but they require manual annotation for ground truth of occlusion for prior learning of occlusion detector. Usually available datasets consist of hundreds or thousands of images, therefore manually annotating ground truth of occlusion for each image is a highly expensive option.

Our approach enhances appearance modeling of a local part template by weighting a sample proportionate to the degree of occlusion of its body parts. For self-occlusion modeling using conventional pose-annotated training data with no annotation of occluded body parts, we propose our method in two folds. First, we introduce the occlusion con-

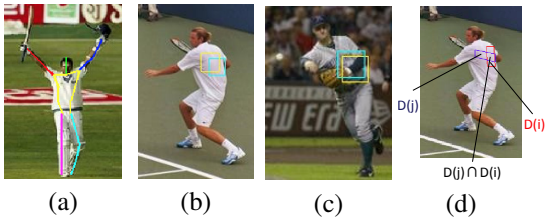

Figure 1. (a) Result of human pose estimation. Inferred locations of body parts are shown as colored lines. Different colored line indicates different body part region. (b) (c) An example of self-occlusion. A torso covers an arm in (b). An arm covers a torso in (c). (d) Illustration of $D(i)$, which is a approximate region surrounding a body part i.

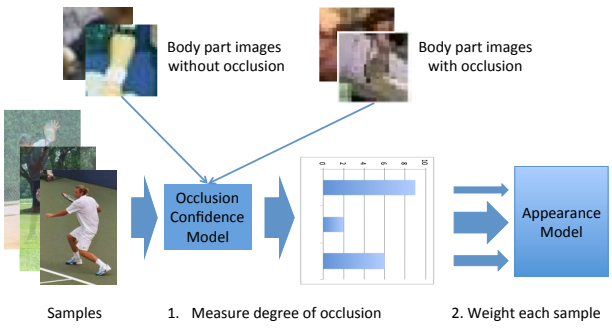

Figure 2. Illustration of our proposed approach. We aim to improve appearance modeling by reducing the effect of occluded appearance used for learning a local part filter.

fidence model which detects possible occlusion on a body part and measures its degree of occlusion. In the occlusion confidence model, firstly we extract candidates of occluded body parts based on pose annotation and then calculate possibility of being occluded for each extracted body part based on its appearance. For measuring self-occlusion given a body part appearance score in a sample, we need large number of various sorts of samples of occluded body part images to evaluate self-occlusion with various kinds of appearance. To meet this end, we artificially synthesize large number of occluded body part images from non-occluded ones. Note that we are proposing a method to improve robustness against self-occlusion in model learning stage, inference of human pose is done conventionally. The overview of the proposed method is shown in Fig. 2.

2. Related Work

In an area of pose estimation from a still image, a graphical model has been used to learn the distribution of human poses in recent works [16, 3, 14]. Especially the PSM approach of Felzenszwalb and Huttenlocher [6] has been widely adopted for the efficient globally optimized inference in number of previous works [2, 10]. For farther improvement of pose estimation, Felzenszwalb *et al.* [5] employs the iterative framework alternating between model learning and refinement of the annotation in training image.

Relating research on self-occlusion such as the foreshortening problem (i.e. A hand of an arm stretching forward looks occluding an elbow) is conducted by Yang and Ramanan [19]. Their approach mitigates the problem from foreshortening by dividing each rigid part (e.g. limb) into several smaller parts.

Another approach regarding self-occlusion is proposed by Johnson and Everingham in [12]. They use the dataset which has incomplete annotation where an occluded body part is not annotated in comparison to widely used the datasets such as the Image Parse dataset [17] and Leeds Sports Pose (LSP) dataset [13] which have annotation for a self-occluded body part as well. Their approach is characterized to discard a sample in appearance modeling if one or more body parts are not annotated in a sample due to self-occlusion. In this paper, based on the approach in [19], we explicitly examine effect of self-occlusion to appearance modeling. In our model we weight a sample containing an occluded body part to be used in appearance modeling to mitigate adverse effect of self-occlusion. Our approach is more efficient in a sense that we fully utilize all training data with high annotation cost compared to [12] in which training data including an occluded body part is not used even if the other body parts are visible and can be used for appearance modeling.

3. Pictorial Structure Model

A tree-based model is defined by a set of body parts V containing a root part and a set of links E connecting two of the body parts. We denote I for an image. A hypothesis $z = (p_0, ..., p_n)$ specifies the location of each part in the model, where p_i represents the pixel location, orientation and scale of part i.

Score of a hypothesis z is given by sum of the score of the filter response of each body part at its location plus deformation cost that depends on the relative position of each body part i with respect to body part j which forms a link of tree-based model,

$$\begin{aligned} score(z) &= \sum_{i \in V} w_i \cdot \phi(I, p_i) + \sum_{i,j \in E} w_{ij} \cdot \phi_d(p_i - p_j) \\ &= \beta \cdot \psi(I, z) \end{aligned} \quad (1)$$

where first term in Eq. (1) represents appearance score and second term for deformation cost. In appearance score term w_i represents a filter for body part i and we write $\phi(x, p_i)$ as a feature vector (e.g. HOG descriptor [4]) extracted from the pixel location p_i in image I. In a typical example of definition for deformation cost, term w_{ij} represents a four dimensional vector specifying coefficients of quadratic function defining the deformation cost and $\phi_d(p_i - p_j)$ defines the deformation cost between body part i and body part j. We define β as a concatenated vector of the filters w_i for each body part and dimensional vectors w_{ij} for the each pair of body parts, and $\psi(I, z)$ shows a concatenated vector of feature vectors $\phi(x, p_i)$ for each body part and the deformation cost $\phi_d(p_i - p_j)$ for the each pair of body parts.

4. Appearance Modeling based on Occlusion Confidence

Here we introduce our approach of learning appearance model based on degree of self-occlusion of each body part. Our method consists of three types of task below.

1. Occlusion confidence modeling, which is trained with a set of occluded body part images as positives and a set of non-occluded ones as negatives and it aims to measure the degree of occlusion of each body part.

2. Data synthesis for occluded body part image, where the goal is to generate occluded body part images from non-occluded ones so that we can train the occlusion confidence model in the first task without manually annotating ground truth of occlusion for thousands of images in a dataset.

3. Weighting sample for appearance modeling, where the aim is that a sample which has more degree of occlusion on its body parts less effects the learning process.

4.1. Occlusion Confidence Model

We aim to evaluate the probability of body part i being not occluded given the filter response value of the body part i to its local part filter. When we apply the local part filter of the body part i to the image of the body part i without self-occlusion, the response value tend to be larger than the case with self-occlusion. We try to exploit this tendency by formulating the probabilistic distribution of the linear generative model as the occlusion confidence model to derive the degree of occlusion of a body part based on its appearance.

The occlusion confidence model is two class softmax function between a class of being occluded and a class of being not occluded, which scores the degree of occlusion as probability given the appearance score of a hypothesis. We write probability of a body part i of an image I being not occluded by the other body parts given a hypothesis of body part locations $z = (p_0, ..., p_n)$ as

$$P_i(\overline{o}|w_i \cdot \phi(I, p_i)) = \begin{cases} f\left(\frac{1}{1+exp(-a)}\right), & (O(i) \neq \emptyset) \\ 1, & \text{otherwise} \end{cases} \quad (2)$$

$$\frac{|D(i) \cap D(j)|}{|D(i)|} > \gamma \quad (3)$$

$$a = \frac{N(w_i \cdot \phi(I, p_i)|\mu_i^{pos}, \Sigma_i^{pos})P_i(\overline{o})}{N(w_i \cdot \phi(I, p_i)|\mu_{i,O(i)}^{neg}, \Sigma_{i,O(i)}^{neg})P_i(o)} \quad (4)$$

$$P_i(\overline{o}) = \frac{\sum_k |O_k(i)|}{\sum_k (|O_k(i)| + \sum_{j \in O_k(i)} |O_k(j)|)} \quad (5)$$

In Eq. (2),$P_i(\overline{o}|w_i \cdot \phi(I, p_i))$ represent possibility of the body part i is being not occluded given the filter response of the local part template of the body part i to the location p_i of an image I. We denote the state of being occluded as o and being not occluded as $\overline{o}$ to be concise. $f(\cdot)$ is an arbitrary function to adjust the range of the output value. $N(w_i \cdot \phi(I, p_i)|\mu_i^{pos}, \Sigma_i^{pos})$ is Gaussian distribution of the appearance score of the body part i with μ_i^{pos} and Σ_i^{pos} which are mean and variance of the appearance score of the non-occluded body part i respectively. To express a set of the body parts which overlap with the body part i, we denote a set of the body part index j as $O(i)$ that satisfies the ratio of the size (pixels) of the intersecting region between the region $D(i)$ and the region $D(j)$ to the size of the region $D(i)$ is more than a threshold γ as show in Eq. (3). $D(i)$ is a region surrounding the body part i and $|D(i)|$ represents the number of pixels contained in $D(i)$. The graphical idea of Eq. (3) is illustrated in Fig. 1 (d). $N(w_i \cdot \phi(I, p_i)|\mu_{i,O(i)}^{neg}, \Sigma_{i,O(i)}^{neg})$ represents Gaussian distribution of the appearance score of the body part i occluded by a set of other body parts $O(i)$ with $\mu_{i,O(i)}^{neg}$ and $\Sigma_{i,O(i)}^{neg}$ which are mean and variance of the appearance score of the body part i occluded by the set of the other body parts $O(i)$. We denote the probability of the body part i being not occluded as $P_i(\overline{o})$ over all samples. $P_i(o)$ denotes the probability of the body part i being occluded over all samples, which is calculated as $1 - P_i(\overline{o})$. We assume there is always one body part occluding the other body parts among a set of the body parts in $O(i)$ thus a body part having an intersecting region with more number of the other body parts has more possibility of being not occluded. We formulate this idea as in Eq. (5).

4.2. Occluded Sample Data Synthesis

In order to train the occlusion confidence model, we need a sample of an image being not occluded as a positive sample and being occluded as a negative sample for each body part. We derive the positive sample set for the body part i as $\{\forall k|O_k(i) = \emptyset, \forall i \in V\}$. We generate the negative sample of the body part i being occluded by the body part j from the positive sample of the body part i and the body part j by cropping the region of the body part j in the sample image and overlay the cropped region onto a sample image of the body part i. The procedure is summarized as follow: a) Superpixelize an image of the body part j. b) Label each pixel cluster either foreground (a body part region) or background. c) Crop foreground and overlay it to the image of the body part i. Above procedure is illustrated in Fig. 3. We employ the off-the-shelf method of [1] to superpixelize. For cropping a region of a body part from background, we utilize annotation to extract the region from background in Bayesian manner as

$$p(F|x, d) = \frac{p(x, d|F)p(F)}{p(x, d)} = \frac{p(x|F)p(d|F)p(F)}{p(x, d)} \quad (6)$$

In Eq. (6) $p(x|F)$ represents distribution over an RGB value x of pixel belonging to the clusters which locates on the line segments connected from a parent of the body part to a child(ren) of the body part passing through the body part. We formulate $p(x|F)$ as mixture Gaussian distribution over with mean and variance over RGB pixel values of each cluster. $p(d|F)$ shows Gaussian distribution given mean and variance over the approximated body part width d. d is calculated as the shortest distance to the line segments as previously described and it is normalized by the size of the annotated image box for each body part. We regard $p(x, d|F)$ is conditionally independent over F because on the same body part region, pixels tend to share relatively similar color distribution and distribution of d is also arguably consistent. We apply Eq. (6) to all pixels in a sample image to derive per pixel likelihood of being foreground. Then we calcu-

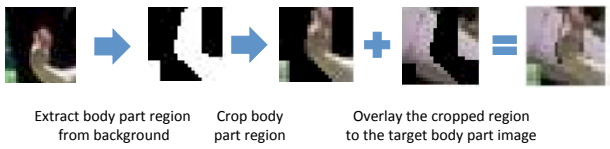

Figure 3. Illustration of negative image synthesis. We generate negative samples from positive samples.

late mean likelihood per cluster and select clusters whose likelihood is above a threshold as foreground. In actual calculation we use $p(F|x,d) \propto p(x|F)p(d|F)$ by assuming $\frac{p(F)}{p(x,d)}$ are constant compared to $p(x|F)p(d|F)$ for calculation efficiency.

4.3. Sample Weighting with Occlusion Confidence

In our base method [19], Latent SVM is used to learn the PSM. We control a sample I influencing learning a positive sample by weighting the SVM's cost function $L(\beta)$ according to the sample's occlusion confidence as

$$L(\beta) = \frac{1}{2}\|\beta\| + C(I) \cdot B \cdot \max(0, 1 - f_\beta(I)) \qquad (7)$$

In Eq. (7) $f_\beta(I)$ denotes score of the SVM with a sample image I and the parameter β over the best hypothesis z. B is the batch size or the number of total positive sample. $C(I)$ shows an arbitrary weighting function for a sample I regarding the occlusion confidence for all body part derived from Eq. (4). In our experiment we use following formulation:

$$C(I) = \frac{1}{|V|}\sum_{i \in V} p_i(\overline{o}|w_i \cdot \phi(I, p_i)) \qquad (8)$$

In Eq. (8) the cost is simple average over the occlusion confidence of all body parts. In our model of weighting sample, the aim is that sample having more occluded body parts influences less in learning. In this approach, not only appearance part but deformation part in Eq. (1) will be affected in learning process but we believe 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.

5. Experimental Results

In this section we report the results of our experiments on the proposed method using the Image Parse dataset [17]. The Image Parse dataset contains 305 images with pose-annotation in total with the standard train/test split. First 100 images are for training and the rest of 205 images are for testing.

In accordance with this base model [5], we use full-body skeleton model. 26 body parts were used in our implementation for covering each body region; 2 for head, 8 for the torso, 4 for each limb.

For validating the ability of detecting occlusion, we manually give annotation of occlusion to all body parts in the dataset to derive the average number of occlusion for each body region: torso, head, upper arms, lower arms, upper legs and lower legs. The number of occlusion in each body part is shown in Table. 1.

In Eq. (2), we experimentally decide to use $f(x) = \frac{1}{1+exp(-x+0.75)^{20}}$. In negative data synthesis for the occlusion confidence model, we use $\gamma = 0.25$ for thresholding foreground and background in Eq. (3).

We tested our approach in standard criteria, probability of a correct pose (PCP) Buffy implementation [9], using the person-centric annotation. The implemented code is distributed at the author's website [8]. The result is shown in Table. 2. In PCP 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 other occluding body parts appearance, and this leads to more accurate inference of the body parts configuration. This is especially true in lower and upper arms whose scores have significantly increased by 7.3% and 2.4% respectively in our approach. The typical examples of estimated human pose of our model is shown in Fig. 4. In the figure, we can see our model (an image shown on the right of the image pair) performs better to predict the locations of lower and upper arms than the base model (on the left of the image pair). Performance improvement on arms is likely because lower and upper arms have the most number of occlusion, thus our method is especially effective reducing the negative effect of occlusion.

In the proposed method, the occlusion confidence model is required to detect the occluded parts for reducing their negative effect of occlusion in appearance modeling. On the other hand, the occlusion confidence model is also required not to falsely detect the non-occluded body parts as occluded ones for efficiently using as many samples in a dataset as possible. To evaluate whether or not this requirement is fulfilled, the recall rate of occlusion detection is an important criterion. Here we say an occluded body part is detected if the probability of the body part being not occluded as in Eq. (2) is less than one. In the dataset mentioned above, the recall rate and the false detection rate with Eq. (2) are 66.3 % and % 21.3, respectively. While the recall rate is not high enough yet, almost all the occluded body parts, which give the huge negative impact on learning part appearance, could be detected by our method. In other words, the occluded body parts that were not detected by our method are expected to have less harmful impact on appearance model learning.

Table 1. Average number of self-occlusion on each body part region. The number is averaged over number of consisting body parts of each body part region (i.e. the number of occlusion for lower arms is derived by averaging the number of occlusion in their eight body parts).

Torso	Head	U.arms	L.arms	U.legs	L.legs
10.75	2	11.75	10.75	9.25	6

Table 2. Comparison of PCP. We list the other methods' on the Parse Image dataset. (a) Johnson [12], (b) Base appoach [19], and (c) Ours.

(a)	87.6	76.8	74.7	67.1	67.4	45.9	67.4
(b)	85.4	84.4	70.8	47.8	77.8	71.2	70.5
(c)	87.3	86.3	73.2	55.1	79.0	71.2	73.1

6. Conclusion

We have introduced the model that improves appearance modeling for the PSM. We show that reducing the effect of occluded body parts to be modeled can provide better appearance modeling to improve estimation performance. Our approach is more efficient on fully using all samples in a dataset for mode learning compared to the previous work [12] discarding whole images including occluded body parts for appearance model. We have shown that our method leads to superior result than the base approach. In future work we would like to drop the false detection rate to more efficiently learn samples and to investigate the other weighting approaches of a sample given its degree of occlusion of each body part other than simple averaging.

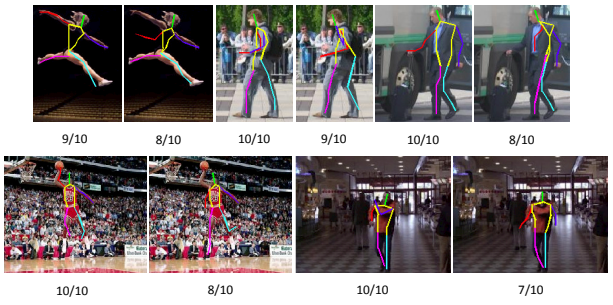

Figure 4. Typical results of our model compared to the base model. Result of our model is shown on the right and the base model on the left in the each pair of images. Caption below an image indicates how many body part regions are successfully localized out of all ten body part regions.

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(11):2274–2282, 2012. 3

[2] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR*, 2009. 2

[3] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. In *ECCV*. Springer, 2010. 2

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 2

[5] 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*, 32(9):1627–1645, 2010. 2, 4

[6] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient matching of pictorial structures. In *CVPR*, 2000. 2

[7] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, 2005. 1

[8] 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/. 4

[9] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR*, 2008. 4

[10] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Pose search: retrieving people using their pose. In *CVPR*, 2009. 2

[11] M. A. Fischler and R. A. Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on computers*, (1):67–92, 1973. 1

[12] S. Johnson and M. Everingham. Learning effective human pose estimation from inaccurate annotation. In *CVPR*, 2011. 2, 5

[13] X. Lan and D. P. Huttenlocher. Beyond trees: Common-factor models for 2d human pose recovery. In *ICCV*, 2005. 2

[14] J. Puwein, L. Ballan, R. Ziegler, and M. Pollefeys. Foreground consistent human pose estimation using branch and bound. In *ECCV*. Springer, 2014. 2

[15] I. Radwan, A. Dhall, J. Joshi, and R. Goecke. Regression based pose estimation with automatic occlusion detection and rectification. In *ICME*, 2012. 1

[16] V. Ramakrishna, D. Munoz, M. Hebert, J. A. Bagnell, and Y. Sheikh. Pose machines: Articulated pose estimation via inference machines. In *ECCV*. Springer, 2014. 2

[17] D. Ramanan. Learning to parse images of articulated bodies. In *NIPS*, 2006. 2, 4

[18] L. Sigal and M. J. Black. Measure locally, reason globally: Occlusion-sensitive articulated pose estimation. In *CVPR*, 2006. 1

[19] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, 2011. 1, 2, 4, 5