Occlusion-free Appearance Modeling of Body Parts for Human Pose Estimation

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

Norimichi Ukita
NAIST
ukita@is.naist.jp

Norihiro Hagita
NAIST
hagita@is.naist.jp

Abstract

In this paper we examine efficacy of occlusion-free appearance learning for part based model. Appearance modeling with less accurate appearance data is problematic because it adversely affects entire learning process. We evaluate the effectiveness of excluding occluded body parts to be modeled for better appearance modeling process. To meet this end, We employ a simple but effective occlusion detection method. We present our approach contributes to improve 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 left picture of Fig. 1 shows the boxes representing each body part as the inferred body configuration. We work on human pose estimation from unconstrained natural still images where we do not have prior information of a depicted activity, anatomical difference between individuals, and the background of the scene. Pose estimation from a single still image has a direct impact on scene understanding for both images and videos. Applications of such methods include automatic surveillance, image retrieval and motion capture.

The task is a highly challenging problem due to a wide variety of appearance resulting from nonrigid deformation of human body, variations of clothings, and inter-personal difference of anatomy. Difficulty also arises from confusion between greatly varying backgrounds of the scenes and a target person.

We base our approach on an improved pictorial structured model (PSM) [3]. PSM is an effective dominant framework [1, 2] for human pose estimation from a single still image. PSM represents a human body configuration as a graphical tree model capturing inter-part spatial relationships such as relative position and orientation and decomposes the appearance of human body into local part templates. PSM enables globally optimized search through distance transform [9] and dynamic programming to tackle the problem of highly articulated nonrigid deformation of the human body. It is important not only to represent human body with set of local parts but also to model appearance of each part robustly against inter-person difference, noise and background. For this purpose, robust feature descriptor such as HOG [16] and PHOG [17] is proposed.

However, regardless of robustness of feature descriptor, it is not feasible to accurately learn appearance of each part with including occluded body parts altogether. Since a human body is highly articulated objects with large degrees of freedom and a single image only represents unidirectional view of the body, many body parts often result in self-occlusion in a natural image such as one body part covers another one, as shown in (Right) of Fig. 1. These kinds of occlusion are not handled by appearance modeling in PSM learning. This means an inaccurate appearance model of the body part is learned under occlusion. Since the features of each body part would not have strong distinctive characteristic (e.g. features of lower and upper arms could be both represented as similar figures of two simple parallel lines), appearance modeling for local part template should be done carefully. Otherwise this problem leads to less accurate pose estimation and detection due to less reliable appearance modeling.

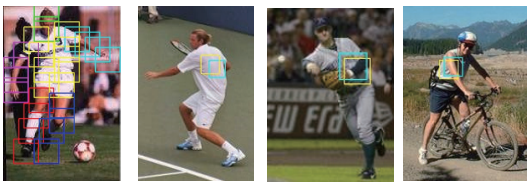
To solve the problem above, we need to exclude occluded body parts. Therefore self-occlusion detection is necessary. Detection of self-occlusion has been researched in previous works such as [12, 13] but it either requires prior learning of occlusion detector with manual annotation for ground truth of occlusion or computationally expensive image processing. Usually available dataset consist of hundreds or thousands of images with annotation even for occluded body parts therefore manual annotation of ground truth occlusion for each image is a highly expensive option.

Our approach enhances appearance modeling of local part templates by excluding occluded body parts to be used in initial appearance learning. We evaluate the result of excluding occluded body parts in learning appearance under idealistic condition where we model body part appearance with using annotation to detect occlusion. To assess effectiveness of the proposed method with dataset which only has pose annotation but does not have prior information on occlusion, we propose a method to automatically detect occluded body parts.

The overview of the proposed method is shown in Fig. 2. We examine our approach on standard benchmark dataset the Parse Image dataset [4] and empirically demonstrate that our approach outperforms the base method [3].

2 Related Work

In an area of pose estimation from still image, graphical model has been used to learn the distribution of human poses in recent works [5, 6, 7]. Especially the PSM approach of Felzenszwalb and Huttenlocher [8] has been widely adopted for efficient globally optimized inference in number of previous works [18, 19, 20]. For farther improvement of pose estimation, Felzenszwalb *et al.* [10] employs iterative framework alternating between model learning and refinement of the annotation in training image which treats annotated location of each body part as latent variable. This approach uses a Latent SVM which shares equivalent formulation with MI-SVM for multiple instance learning in



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

Figure 1. (a) Result of human pose estimation. Inferred body parts location are shown as colored boxes. (b) (c) (d) An example of self-occlusion. Torso covers an arm in (b). An arm covers torso in (c) and (d).

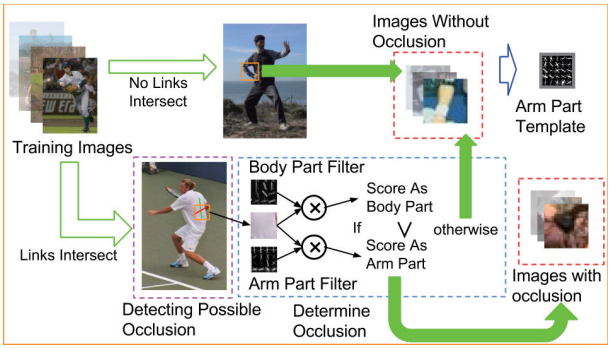


Figure 2. Illustration of our proposed approach. We aim to improve appearance modeling by excluding occluded appearance from being used for local part filter in appearance modeling. This figure shows the case of arm appearance modeling.

[11] for optimizing over latent part location through a coordinate-descent style algorithm for training. Latent SVM improves discriminative inference of pose.

In addition to the problem mentioned above (i.e. globally optimized solution, how to improve discriminative parameters), as mentioned in introduction, occlusion is also a severe problem. The problem relating to occlusion in pose estimation at pose inference stage has been researched in previous works [12, 13] producing framework for articulated pose estimation which is robust to self-occlusion. But the research on explicit effect of self-occlusion in appearance modeling has not been conducted. Relating research on self-occlusion such as foreshortening problem (i.e. A hand of a arm stretching forward looks occluding an elbow) is conducted by Yang and Ramanan [3]. Their approach mitigates problems from foreshortening by introducing representation for appearance modeling such as a family of affinely-warped templates dividing each rigid part (e.g. limb) into several smaller parts.

Approach proposed by Johnson and Everingham in [15] uses a dataset which has incomplete annotation where occluded body parts are not annotated in comparison to widely used datasets such as the Image Parse dataset [4] and Leeds Sports Pose (LSP) dataset [14]. It is important to note that their approach discards images in appearance modeling if one or more body parts are not annotated due to occlusion. In contrast to this, our approach In order to effectively utilize pose annotated teaching data, we only ignore occluded body parts and the rest of parts in a same image is still used to learn appearance model.

In this paper, based on approach in [3], we explicitly examine effect of self-occlusion to initialization in appearance modeling by comparing results from the proposed approach and the base method, as opposed to [15], with same dataset. In our model we exclude occluded body parts to be used in initial appearance modeling to mitigate adverse effect of self-occlusion to appearance learning. Our approach is more efficient in a sense that we fully utilize all training data with high annotation cost compared to [15] in which a set of training images including occluded parts are not used even if other parts are visible. Initialization of appearance modeling is especially core issue for the model using a latent SVM where careful initialization of the model is necessary since unreasonable selection of the initial model can lead to bad local optimum as suggested in [10].

3 Pictorial Structure Model

In this section we describe the basis of PSM. A tree-based model is defined by a set of parts V containing a root part and a set of links E connecting two of the 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 and scale of part i.

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

$$
\begin{aligned}
& score(z) \\
&= Sa(z) + Sd(z) && (1) \\
&= \sum_{i \in V} w_i \cdot \phi(I, p_i) + \sum_{i,j \in E} w_{ij} \cdot \phi_d(p_i - p_j) && (2) \\
&= \beta \cdot \psi(I, z) && (3)
\end{aligned}
$$

where first term in equation (1) represents appearance score and second term for deformation cost. In appearance score term w_i represents a filter for part i and we write $\phi(x, p_i)$ as a feature vector (e.g. HOG descriptor [16]) extracted from 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 deformation cost between part i and part j. In this example of definition for deformation cost, the score of a hypothesis z can be expressed as $\beta \cdot \psi(I, z)$ where β and $\psi(I, z)$ are expressed as,

$$
\begin{aligned}
& \beta = (w_0, ..., w_i, ..., w_n, w_{11}, ..., w_{ij}, ..., w_{nn}) && (4) \\
& \psi(I, z) = (\phi(I, p_0), ..., \phi(I, p_i), ..., \phi(I, p_n), \\
& \phi_d(p_1 - p_1), ..., \phi_d(p_i - p_j), ..., \phi_d(p_n - p_n)). && (5)
\end{aligned}
$$

4 Initialization of Appearance Model for Human Pose Estimation with Occlusion-Free Body Parts Appearance

Here we introduce our approach constructing local part templates with excluding occluded body part. We write O_k to define a set of occluded body parts in image I_k. The appearance score of a hypothesis z for the appearance modeling can be written as

$$Sa(z) = \sum_{i \in V \setminus O_k} w_i \cdot \phi(I_k, p_i). \quad (6)$$

The above appearance score is used to derive total score of model β given a hypothesis z.

5 Occluded Parts Detection Using Pose Annotation

In this section we propose simple but effective approach to detect occluded body parts. We use 0-1 function $f(l)$ to detect if a line segment corresponding to $l \in E$ of body part i and its parent body part j intersects with another line segments or not as follows:

$$f(l) = \begin{cases} 1 & (l \textit{ intersects with } l',\ \forall l' \in E \setminus l) \\ 0 & \textit{otherwise} \end{cases} \quad (7)$$

If $f(l) = 1$, the two body parts consisting the link l is occluded by the other two body parts of the link l', or link l occludes link l'. We regard each link as a line segment on two dimensional plane so that it is easy to detect intersection of the two links. Although only with this approach we cannot judge which of the two links is occluded, removing both links for modeling appearance of body parts are acceptable because our goal is to exclude occluded body part appearance to be modeled.

After we determine intersecting link l and l' in Eq. 7, for farther detection of occlusion when local part template is available, we can judge which of two links l and l' is occluded as

$$e = \underset{e \in \{l, l'\}}{\operatorname{argmin}}\ w_e \cdot \phi(I, e) \quad (8)$$

Note that we assume appearance of l and l' largely share same image region due to occlusion, therefore the corresponding local part filter of the occluded part gives lower filter response because that image region shows occluding body part.

The simplest alternative approach could be comparing each local part filter response of all body parts with filter response of l at the corresponding image region of l and conclude link l is occluded when one of the other links gives lower filter response. But there is still risk that some local part filters gives lower filter response even when l is actually occluded. By reducing candidate occluding parts by Eq. 7 our approach mitigates such risk and determines which part is occluded one.

6 Experimental Results

In this section we report results of experiments on the proposed method using the Image Parse dataset [4]. The Image Parse dataset contains 305 images with pose-annotation in total with 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 [10], we use full-body skeleton model. 26 parts were used in our implementation; 2 for head, 4 for the torso, 10 for the shoulders to the hands and 10 for the hips to the feet. In this approach the model parameter β is learned in coordinate descent manner alternating between selection of local parts locations z which maximizes $\beta \cdot \psi(I, z)$ and optimizing β given z over an objective function.

We identify occluded body parts in each image in the dataset and skip the occluded body parts to be modeled in initial appearance modeling which is used to choose local parts locations z in the first iteration. In this paper, we select occluded body parts both manually and automatically. Manual selection gives theoretical upper-bound on our performance of the proposed approach in terms of reliance of detecting occlusion. In automatic detection proposed in Eq. 8, local part template is first modeled with appearance data including occluded body parts.

In the proposed method, occlusion detection is required to detect occluded parts as many as possible for occlusion-free appearance modeling. To evaluate whether or not this requirement is achieved, the recall rate of occlusion detection is an important criterion. In the dataset mentioned above, the recall rates of our method with Eq. 7 and Eq. 8 were 42.4 and 23.3, respectively. While these results are not high enough yet, almost all fully-occluded parts, which give a huge negative impact on learning part appearance, could be detected by finding the intersecting line segments of parts in our method. In other words, partially-occluded parts that were not detected by our method are expected to have a less harmful impact on appearance model learning.

We test our approach on four levels of occlusion: (a) body parts with foreshortening are skipped to be modeled in addition to the ones with normal self-occlusion where foreshortening and self-occlusion are manually detected, (b) body parts only with normal self-occlusion are skipped where self-occlusion is manually detected, (c) body parts only with normal self-occlusion are skipped where self-occlusion is automatically detected as in Eq 7, (c) body parts only with normal self-occlusion are skipped where self-occlusion is automatically detected as in Eq. 8. We tested our approach in standard criteria, Average Precision of Keypoints (APK) and Probability of Correction Keypoints (PCK). The result is shown in Table. 1 and Table. 2 with respective criteria. In APK our approach gives same or better performance to base method and in PCK, our method excluding both foreshortening and self-occlusion gives better results to base approach. The results with automatic occlusion detection in the both criteria gives almost same score as results with manual occlusion detection. This means our approach for automatic occlusion detection performs well for appearance modeling as manual detection does.

Our better performance to base approach is likely because our approach successfully learns local part templates which are less distracted by other occluding body parts appearance, and this leads to more accurate inference of body parts configuration in better

Table 1. Comparison of APK. (a) our model (occlusion and foreshortening with manual occlusion detection), (b) our model (only occlusion with manual occlusion detection), (c) our model (only occlusion with automatic occlusion detection as in Eq. 7), (c) our model (only occlusion with automatic occlusion detection as in Eq. 8), (e) base approach

	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
(a) Ours (occlusion and foreshortening)	88.4	80.9	54.5	29.4	72.8	63.8	53.2	63.3
(b) Ours (only occlusion)	86.3	80.9	59.0	30.7	71.5	64.9	54.8	64.0
(c) Ours (only occlusion detected in Eq. 7)	88.9	83.7	58.3	27.4	74.8	65.9	53.1	64.6
(d) Ours (only occlusion detected in Eq. 8)	88.9	83.6	57.9	30.8	74.8	63.3	54.4	64.8
(e) Base approach [3]	87.4	81.4	58.8	29.1	72.0	62.3	52.3	63.3

Table 2. Comparison of PCK. (a) our model (occlusion and foreshortening with manual occlusion detection), (b) our model (only occlusion with manual occlusion detection), (c) our model (only occlusion with automatic occlusion detection as in Eq. 7), (c) our model (only occlusion with automatic occlusion detection as in Eq. 8), (e) base approach

	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
(a) Ours (occlusion and foreshortening)	89.3	84.9	67.3	49.5	81.5	75.4	67.3	73.6
(b) Ours (only occlusion)	89.5	84.1	64.9	45.9	78.5	74.4	66.8	72.0
(c) Ours (only occlusion detected in Eq. 7)	90.2	86.3	68.3	45.9	80.7	76.1	66.8	73.5
(d) Ours (only occlusion detected in Eq. 8)	90.2	85.4	68.5	48.0	79.5	74.1	67.6	73.3
(e) Base approach [3]	90.2	85.1	68.5	46.6	78.3	73.4	65.9	72.6

local optimum through iteration.

7 Conclusion

We have described a model that improves appearance modeling for PSM. We show that removing occluded body parts to be modeled for local part template in model initialization can provide better local optimum through iteration. Our approach is more efficient compared to the previous work discarding whole image including occluded body parts for appearance model. In future work we will investigate effect of removing occluded body parts to be modeled not only in initial appearance modeling but also in later iteration.

References

[1] P. F. Felzenszwalb, Pedro and D. P. Huttenlocher. "Pictorial structures for object recognition." *IJCV*, 61.1 (2005): 55-79.

[2] M. A. Fischler and R. A. Elschlager. "The representation and matching of pictorial structures." *IEEE Transactions on Computers* 22.1 (1973): 67-92.

[3] Y. Yang and D. Ramanan. "Articulated pose estimation with flexible mixtures-of-parts." *CVPR*, 2011.

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

[5] V. Ramakrishna, et al. "Pose Machines: Articulated Pose Estimation via Inference Machines." *ECCV*, 2014.

[6] L. Bourdev, et al. "Detecting people using mutually consistent poselet activations." *ECCV*, 2010.

[7] J. Puwein, et al. "Foreground Consistent Human Pose Estimation using Branch and Bound." *ECCV*, 2014.

[8] P. F. Felzenszwalb, and D. P. Huttenlocher. "Efficient matching of pictorial structures." *CVPR*, 2000.

[9] P. F. Felzenszwalb and D. P. Huttenlocher. "Distance transforms of sampled functions." *TR of Cornell University*, 2004.

[10] P. F. Felzenszwalb, et al. "Object detection with discriminatively trained part-based models." *PAMI*, 32.9 (2010): 1627-1645.

[11] S. Andrews, I. Tsochantaridis, and Thomas Hofmann. "Support vector machines for multiple-instance learning." *NIPS*, 2002.

[12] I. Radwan, et al. "Regression based pose estimation with automatic occlusion detection and rectification." ICME, 2012.

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

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

[15] S. Johnson and M. Everingham. "Learning effective human pose estimation from inaccurate annotation." *CVPR*, 2011.

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

[17] Bosch, Anna, Andrew Zisserman, and Xavier Munoz. "Representing shape with a spatial pyramid kernel." Proceedings of the 6th ACM international conference on Image and video retrieval. ACM, 2007.

[18] Andriluka, Mykhaylo, Stefan Roth, and Bernt Schiele. "Pictorial structures revisited: People detection and articulated pose estimation." Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009.

[19] Ferrari, Vittorio, Manuel Marin-Jimenez, and Andrew Zisserman. "Pose search: retrieving people using their pose." Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009.

[20] Radwan, Ibrahim, et al. "Regression based pose estimation with automatic occlusion detection and rectification." Multimedia and Expo (ICME), 2012 IEEE International Conference on. IEEE, 2012.