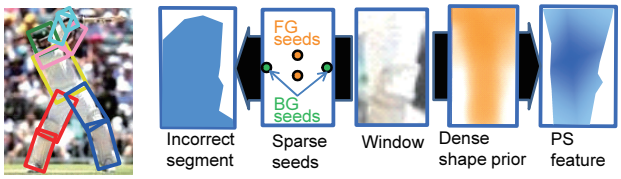
Part-segment Features for Articulated Pose Estimation

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

Abstract

We propose part-segment (PS) features for estimating an articulated pose in still images. The proposed PS feature evaluates the image likelihood of each body part (e.g. head, torso, and arms) robustly to background clutter and nuisance textures on the body and clothing. In contrast to similar segmentation features, part segmentation is improved by part-specific shape priors that are optimized by training images with fully-automatically obtained seeds. The extracted PS feature is fused complementarily with gradient features using discriminative training and adaptive weighting for robust and accurate evaluation of part similarity.



(a) Pose (b) Segments obtained by (left) seeds and (right) shape prior

Figure 1. (a) Parts, depicted by rectangle windows, are configured in their proper locations. (b) While the proposed PS feature is extracted by a dense shape prior that is optimized to each part, correct segmentation is difficult with a weak shape prior (i.e. "Sparse seeds").

1 Introduction

Image features are crucial for articulated pose estimation (Fig. 1 (a)). While most of articulated pose estimation methods employ gradient features such as HOG [4], they include nuisance responses caused by background clutter and textures on a target body.

This work focuses on how to extract only useful responses that represent the boundary of each part based on a shape prior optimized to that part. The boundary is represented by a *part-segment* (PS) feature, as shown in Fig. 1 (b).

However, part segmentation with no shape prior is difficult. The proposed method achieves initial part segmentation with seeds given to each window. Their distribution is automatically determined by training samples. While the distribution of the seeds gives a weak shape prior, a more reliable dense shape prior is acquired from training images with the seeds.

Our contribution is threefold: 1) initial *shape priors* are extracted by segmentation using *part-specific* foreground (FG) and background (BG) seeds trained *automatically* (Sec. 4.1), 2) the shape priors are *refined* and *clustered* for correctly and efficiently computing PS features (Sec. 4.2), and 3) *adaptive weighting* of the PS features with domain adaptation improves their discriminativity (Sec. 4.3).

2 Related Work

Image segments can be obtained by image segmentation such as superpixelization [1]. However, it is not easy to extract each part as one segment because the part might be over-segmented due to textures and shades on a human body. Such over-segmentation can be suppressed by a prior on the configuration/shape of each part in an image. Table 1 summarizes several properties of methods for/using segmentation.

As the prior, the configuration of roughly detected parts is useful (e.g. upper-body [7] detection). In ObjCut [10] and [13, 14, 7], one or more parts are detected initially using features with no segmentation. Depending on the configuration of the detected parts, seeds for segmentation are distributed. It is, however, difficult to distribute the seeds to all the parts by using only the limited detected parts. In addition, error in distributing the seeds propagates to part segmentation.

While several methods [12, 9] achieve part segmentation and detection independently ("DI" in Table 1), each of these methods has functional defects. Segmentation cues [9] require manually-predefined sparse seeds. CHOG [12] also needs a set of manually-given seeds. In addition, they [9, 12] provide only sparse seeds (i.e. a pair of FG and BG pixels). The proposed PS feature is extracted by part-specific dense shape priors optimized by automatically-given seeds and training images; these advantages are shown in "SP", "PS-SP" and "AS" in Table 1.

3 Pose Estimation

An articulated model is represented, in general, by a tree model defined by a set of nodes, $\boldsymbol{V}$, and a set of links each of which connects two nodes, $\boldsymbol{E}$. Each node and link respectively corresponds to a part and a physical connection between parts. The pose parameters of the node are optimized for pose estimation by maximizing the score function below:

$$T(\boldsymbol{P}) = \sum_{i \in \boldsymbol{V}} S_i(\boldsymbol{p}_i) + \sum_{i,j \in \boldsymbol{E}} P_{i,j}(\boldsymbol{p}_i, \boldsymbol{p}_j), \quad (1)$$

where $\boldsymbol{p}_i$ and $\boldsymbol{P}$ denote the pose parameters of i-th part and its set of all parts ($\boldsymbol{P} = \{\boldsymbol{p}_i | \forall i \in \boldsymbol{V}\}$).

A unary term $S_i(\boldsymbol{p}_i)$ is a similarity score of i-th part at $\boldsymbol{p}_i$. In the proposed model, $S_i(\boldsymbol{p}_i)$ is a sum of filter responses using HOG [4] and the PS feature.

$$S_i(\boldsymbol{p}_i) = \left[F_i^T, G_i^T\right] \left[\phi(I, \boldsymbol{p}_i), \varphi(I, \boldsymbol{p}_i, i)\right]^T \quad (2)$$

where F_i and $\phi(I, \boldsymbol{p}_i)$ denote the filter of i-th part and the HOG extracted at $\boldsymbol{p}_i$ in image I, respectively, and G_i and $\varphi(I, \boldsymbol{p}_i, i)$ denote those of the PS feature.

Table 1. Comparison of body/part segmentation methods. Each column shows whether the methods exhibit each property. No init: no pose initialization is required. DT: feature is discriminatively trained. Weight: each pixel/cell in a segment has its weight (i.e. probability to be FG). DI: detection and segmentation are achieved independently. SP: segmentation is achieved using a shape prior. PS-SP: part-specific shape prior is given. AS: seeds are given automatically. NR: negative effects due to noise in segmentation are suppressed.

	No init	DT	Weight	DI	SP	PS-SP	AS	NR
(a) ObjCut [10]					Y			
(b) Parse [13], Better appearance [5]			Y					
(c) Segmentation [14]		Y			Y			
(d) PoseCut [3]	Y				Y			
(e) CHOG [12]	Y	Y	Y	Y				Y
(f) Segmentation cues [9]	Y	Y		Y				
(g) PS feature (Proposed)	Y	Y	Y	Y	Y	Y	Y	Y

A pairwise term $P_{i,j}(\boldsymbol{p}_i, \boldsymbol{p}_j)$ is a spring-based score between i-th and j-th parts, which has a greater value if the configuration of $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$ is highly probable:

$$P_{i,j}(\boldsymbol{p}_i, \boldsymbol{p}_j) \quad = \quad \boldsymbol{w}_{i,j}^T \cdot [d_{i,j}^x, d_{i,j}^{x^2}, d_{i,j}^y, d_{i,j}^{y^2}]^T \qquad (3)$$

$\boldsymbol{w}_{i,j}$ is a weight parameter. $d_{i,j}^x$ and $d_{i,j}^y$ denote $(x_i - x_j)$ and $(y_i - y_j)$, respectively, where $(x_i, y_i) \in \boldsymbol{p}_i$ and $(x_j, y_j) \in \boldsymbol{p}_j$ are the locations of i-th and j-th parts.

In what follows, how to learn G_i and extract $\varphi(I, \boldsymbol{p}_i, i)$ is described.

4 Training of Part-segment Features

4.1 Initial Shape Prior from FG and BG seeds

The shape prior of each part is obtained from its segments in all training images. For extracting the segments, each image is segmented by SLIC superpixelization [1] (Figs. 2 (a), (b), and (c)). Since the region of a part might be over-segmented, segments in each window must be clustered into those of the part of interest and others. This clustering is achieved initially with seeds automatically given by using training data.

4.1.1 Fully-automatic Configuration of Seeds

Training data consists of images and pose annotations. The pose of each part is given as a window and a pair of end-points of a part line. In each part's window, the initial sample colors of FG are collected from segments that cross the part line. The mean color of each segment is denoted by $\boldsymbol{c}_s$ in s-th segment.

Then the distance between each segment's color and its nearest neighbor color in the collected FG samples is computed. The color distances are binarized [11] for dividing the segment colors into FG (i.e. colors with a smaller distance) and BG. If $\boldsymbol{c}_s$ is in FG colors, s-th segment is temporally clustered into FG. This clustering is executed in all parts' windows in all training images. After the window sizes are normalized, the rate of FG in all training images is computed in each pixel of each part window. Pixels with the top/bottom γ % FG rate are extracted as FG/BG seeds.

4.1.2 Segment Clustering with Seeds

The seeds provide a weak shape prior. Segment clustering in each window is re-executed with the seeds:

1. Segments each of which has *only* FG/BG seeds are clustered into FG/BG segments in a window. If either of FG or BG segment is not found, this window is removed in learning an initial shape prior.
2. Remaining segments are clustered into FG or BG based on their nearest neighbor colors of FG and BG segments for part segmentation.
3. The part-segmented window is regarded as an initial binary PS feature (denoted by $\boldsymbol{\varphi}_i$ for i-th part) where FG/BG pixels have 1/0 as pixel values.

Binary PS features of all training images, except those removed in the above step 1, are averaged in each part. The mean is regarded as the initial shape prior (denoted by $\bar{\boldsymbol{\varphi}}_i$ for i-th part): $\bar{\boldsymbol{\varphi}}_i = \left(\sum_i^{N_p} \boldsymbol{\varphi}_i\right) / N_p$, where N_p denotes the number of training images.

Figure 2 (e) shows obtained binary PS features. By comparing them with their respective images (Fig. 2 (f)), it seems segmentation is reasonable. However, a PS feature might be sometimes extracted unsuccessfully, as shown in Fig. 2 (h-upper); its successful example is (h-lower).

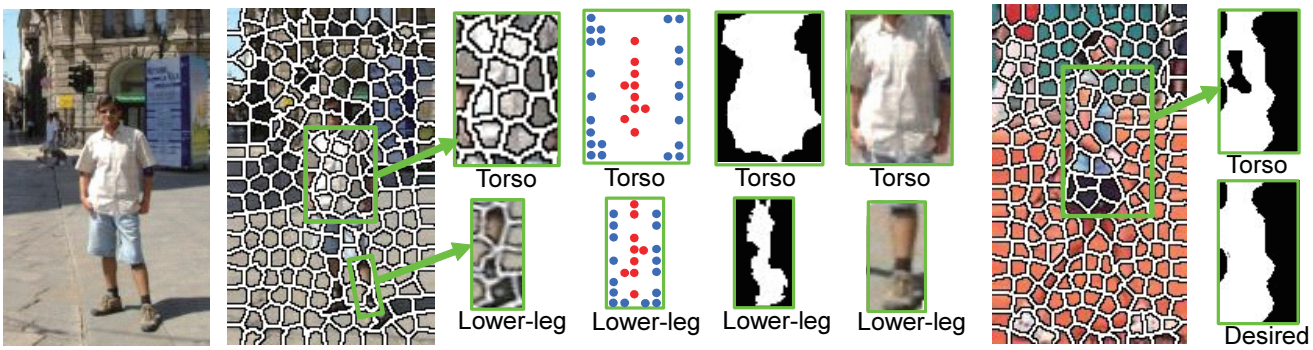
4.2 Segmentation with Updating Shape Prior

For refining the shape prior, a PS feature of i-th part in each training image is updated with $\bar{\boldsymbol{\varphi}}_i$:

1. The mean color of a segment having FG/BG seeds in each window is stored as the sample color of FG/BG, after the size of the shape prior is changed to that of the window.
2. By comparing $\bar{\boldsymbol{\varphi}}_i$ and a segmented window, the mean of pixel values of $\bar{\boldsymbol{\varphi}}_i$ in s-th segment is regarded as the probability that the segment is FG. This probability is denoted by $P_f(s)$.
3. The nearest neighbor of the mean color in the s-th segment is found from the sample colors of FG. The color distance from the nearest neighbor is denoted by $l_f(s)$. $l_b(s)$ for BG is also computed.
4. By deeming $\exp(-l_f(s))$ and $\exp(-l_b(s))$ to be image likelihoods, the Bayes' theorem gives the following probabilities:

$$\begin{aligned} P(f|s) \quad &\propto \quad \exp(-\lambda l_f(s)) P_f(s) \\ P(b|s) \quad &\propto \quad \exp(-\lambda l_b(s))(1 - P_f(s)) \end{aligned}$$

Pixels in s-th segment have the pixel value below: $P(f|s) / \left(P(f|s) + P(b|s)\right)$



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 2. Initial part segmentation in training. (a) Input image. (b) Correct windows of two parts (i.e. torso and left lower-leg) superimposed on a segmented image. (c) Cropped windows of the torso and the left lower-leg. (d) FG and BG seeds, indicated by red and blue circles, respectively. (e) Binary segmentation using the seeds. (f) Parts' windows cropped from (a). (g) Another input image. (h-upper) Torso segment extracted from (g). (h-lower) Desired torso segment in (g).

4.3 Discriminative Training of Adaptively-weighted Gradient and Segment Features

Discriminative training [6] optimizes the model parameters in score (1), namely F_i and G_i in (2) and $\boldsymbol{w}_{i,j}$ in (3).

To improve effects of PS features, an additional weight is given to a PS feature depending on its confidence. The confidence value $C(\varphi(I,\boldsymbol{p}_i,i),i)$ of PS feature $\varphi(I,\boldsymbol{p}_i,i)$ is determined by the subtraction between $\varphi(I,\boldsymbol{p}_i,i)$ and the shape prior of i-th part, $\bar{\varphi}_i$:

$$C(\varphi(I,\boldsymbol{p}_i,i),i) = \exp(-||\varphi(I,\boldsymbol{p}_i,i) - \bar{\varphi}_i||)$$

For efficient training while using $C(\varphi(I,\boldsymbol{p}_i,i),i)$, we train their appearance filters with domain adaptation by redundantly-concatenated features [8]. The feature vector, $[\phi(I,\boldsymbol{p}_i),\varphi(I,\boldsymbol{p}_i,i)]^T$, is changed to either of the followings depending on the confidence value of the PS feature:

$$[\phi(I,\boldsymbol{p}_i),\varphi_1,\varphi_2,\varphi_3]^T = [\phi(I,\boldsymbol{p}_i),\varphi(I,\boldsymbol{p}_i,i),\varphi(I,\boldsymbol{p}_i,i),\mathbf{0}]^T, \quad \text{if } C(\varphi(I,\boldsymbol{p}_i,i),i) < C' \tag{4}$$

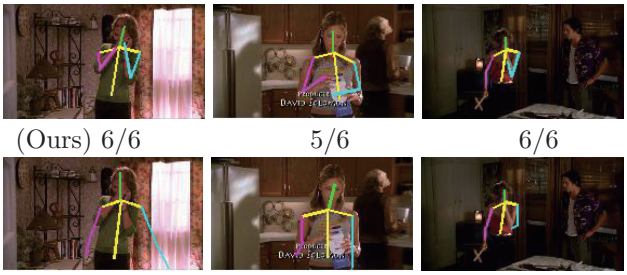
$$[\phi(I,\boldsymbol{p}_i),\varphi_1,\varphi_2,\varphi_3]^T = [\phi(I,\boldsymbol{p}_i),\varphi(I,\boldsymbol{p}_i,i),\mathbf{0},\varphi(I,\boldsymbol{p}_i,i)]^T, \quad \text{if } C' \le C(\varphi(I,\boldsymbol{p}_i,i),i) \tag{5}$$

$C(\varphi(I,\boldsymbol{p}_i,i),i)$ is clustered into two classes by threshold C'. Given the number of the classes, C' was determined by K-means clustering of $C(\varphi(I,\boldsymbol{p}_i,i),i)$ of PS features obtained from all training images; C' coincides with the middle point between the means of two neighboring clusters. With feature vectors (4) and (5), appearance score (2) is rewritten:

$$S_i(\boldsymbol{p}_i) = \left[F_i^T, G_{i,1}^T, G_{i,2}^T, G_{i,3}^T\right][\phi(I,\boldsymbol{p}_i),\varphi_1,\varphi_2,\varphi_3]^T \tag{6}$$

5 Inference with Part-segment Features

In pose inference, PS features are extracted from all possible windows in a test image for optimizing score (1). The PS features are extracted by steps 1–4 described in Sec. 4.2. With the extracted PS features, the appearance score (6) with concatenated features (4) and (5) are used for computing the score (1).



(Ours) 6/6 5/6 6/6

(Base) 4/6 4/6 5/6

Figure 3. Results of our method and the base method[15] in BUFFY. The number of correctly localized parts is shown under each result.

6 Experiments

We tested the proposed PS features with the Image Parse (IP) [13] and the BUFFY stickmen [7] datasets. Negative samples for discriminative training were given from 1218 background images in the INRIA Person database [4].

A human body is modeled with a mixture of non-oriented parts proposed by Yang and Ramanan [15]. This base model[15] divides physically-rigid parts (e.g. limbs) into smaller 26 parts for robustness to in-plane rotation and foreshortening of body parts.

The seeds in each part were given in a 11×11 pixels window, which is scaled with respect to the size of a window.

The results of pose estimation were evaluated quantitatively by the percentage of correctly localized parts (PCP). PCP was implemented by the code in the BUFFY dataset [7] with the strictest interpretation described in [15]. Tables 2 3 and show the results. For comparison, the results obtained by the base model [15] is shown. The effects of the proposed schemes were evaluated with (b) initial binary PS features obtained only by seeds (i.e. shape priors were not used for part segmentation), (c) PS features without domain adaptation (i.e. concatenated features (4) and (5) were not used), and (d) the full PS features obtained by using all schemes proposed in this paper.

Table 2. BUFFY stickmen dataset: Comparative results of PCP. (a) base model [15] (b) our initial binary feature extracted only by seeds, (c) our feature without domain adaptation, and (d) our proposed PS feature.

Model	Head	Torso	Upper-arms	Lower-arms	Total
(a) Mixture of parts [15]	99.2	98.8	97.8	68.6	88.5
(b) Ours by binary feature (by seeds)	99.3	98.9	97.4	68.8	88.4
(c) Ours without adaptation (by shape prior)	99.3	99.3	98.2	70.3	89.3
(d) Ours (full model)	99.3	99.3	98.2	70.3	89.3

Table 3. IP dataset: Comparative results of PCP.

Model	Head	Torso	U-legs	L-legs	U-arms	L-arms	Total
(a) Mixture of parts [15]	99.0	96.1	85.9	79.0	79.0	53.4	79.0
(b) Ours by binary feature (by seeds)	99.0	96.1	87.3	78.5	79.5	52.7	79.1
(c) Ours without adaptation (with shape prior)	99.0	97.6	88.8	79.5	85.9	56.1	82.0
(d) Ours (full model)	99.0	97.6	89.8	80.5	87.3	56.1	82.4

(Ours) 9/10 9/10 10/10 10/10

(Base) 7/10 7/10 8/10 8/10

Figure 4. Results of our method and the base method[15] in Image Parse.

In the comparative experiments, our method had less impacts in BUFFY compared. This might be because 1) many images in BUFFY have low contrast that makes segmentation difficult, and 2) people in BUFFY, who were pictured larger than those in IP, were too over-segmented by SLIC [1]. While more deliberate segmentation methods (e.g. globalPb [2]) might alleviate those problems, they need much computational cost. For example, globalPb took 30 sec or more, while SLIC [1] took around 1 sec for segmentation of each image in IP.

Figures 3 and 4 show examples of results improved by the proposed method. For visualization, 6 and 10 parts, whose joints are a subset of those of full-body 26 parts, are displayed. The rightmost example in Fig. 4 shows a typical case where the PS features could localize a limb (i.e. lower-arms) without being disturbed by a noisy background.

7 Concluding Remarks

This paper proposed the part-segment features for evaluating the shape of each part. In training, the PS features are extracted with automatically trained initial seeds and then refined for improving a shape prior on each part. The extracted features are discriminatively trained, and their adaptive weights with respect to gradient features are also learnt.

Future work includes more efficient extraction and discriminative representation of the PS feature. Reducing its parameters is also important (e.g. automatic selection of the number of superpixels).

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(11):2274–2282, 2012. 1, 2, 4

[2] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(5):898–916, 2011. 4

[3] M. Bray, P. Kohli, and P. H. S. Torr. Posecut: Simultaneous segmentation and 3d pose estimation of humans using dynamic graph-cuts. In *ECCV (2)*, 2006. 2

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 1, 3

[5] M. Eichner and V. Ferrari. Better appearance models for pictorial structures. In *BMVC*, 2009. 2

[6] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1627–1645, 2010. 3

[7] V. Ferrari, M. J. Marín-Jiménez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR*, 2008. 1, 3

[8] H. D. III. Frustratingly easy domain adaptation. In *ACL*, 2007. 3

[9] S. Johnson and M. Everingham. Combining discriminative appearance and segmentation cues for articulated human pose estimation. In *MLVMA*, 2009. 1, 2

[10] M. P. Kumar, P. H. S. Torr, and A. Zisserman. Objcut: Efficient segmentation using top-down and bottom-up cues. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(3):530–545, 2010. 1, 2

[11] N. Otsu. A threshold selection method from gray-level histograms. *IEEE Trans. Sys., Mam., Cyber.*, 9(1):62–66, 1979. 2

[12] P. Ott and M. Everingham. Implicit color segmentation features for pedestrian and object detection. In *ICCV*, 2009. 1, 2

[13] D. Ramanan. Learning to parse images of articulated bodies. In *NIPS*, 2006. 1, 2, 3

[14] D. Ramanan. Using segmentation to verify object hypotheses. In *CVPR*, 2007. 1, 2

[15] Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures of parts. In *IEEE Trans. Pattern Anal. Mach. Intell.*, volume 35, pages 2878–2890, 2013. 3, 4