# ADAPTIVE APPEARANCE LEARNING FOR HUMAN POSE ESTIMATION

*Lei Wang, Xu Zhao*, Yuncai Liu*

Key Laboratory of System Control and Information Processing
Department of Automation, Shanghai Jiao Tong University
800 Dongchuan RD, Shanghai, China

## ABSTRACT

We address the problem of pose estimation in videos. The part detectors play important roles, but traditional template-based detectors (e.g. Histogram of Gradient, HoG) fail at pose estimation due to the high variability in appearance. We present an adaptive representation of appearance and shape for articulated human body. The full representation of human body is based on the flexible mixture-of-parts model. We train a Naive Bayes classifier to obtain a confidence score of estimated pose by the basic mixture model, and based on the confidence we learn an instance-specific appearance model. For between-frame consistency, we design a time-efficient energy function for motion cues instead of complex motion models. We incorporate these models into a framework that allows for efficient inference. Quantitative evaluation of pose estimation conducted on two video datasets demonstrates the effectiveness of the proposed method.

## 1. INTRODUCTION

Human pose estimation is an important task in computer vision area, and the principle purpose of this task is to locate each body part and obtain the body configuration. Pose estimation holds potential to impact many applications that widely range from image understanding to action recognition and human computer interaction. However, this problem is challenging, because of the variations of body shape, pose and appearance. Also, human body is articulated, and there are many degrees of freedom to be estimated.

A classic approach for pose estimation is the pictorial structure model [1, 2], as shown in Fig. 1(a). The whole body is decomposed into local body parts, and each pair of parts is connected by a "spring" as the geometric constrain. The inference of the method is complicated and time-consuming, since a large number of rotated and foreshortened part templates need to be searched to obtain the best configuration. Yang *et al.* [3] proposed a flexible mixture-of-parts model with non-oriented parts, as shown in Fig. 1(b). The model can capture local appearance and body geometry, and can handle rotation and foreshortening implicitly.

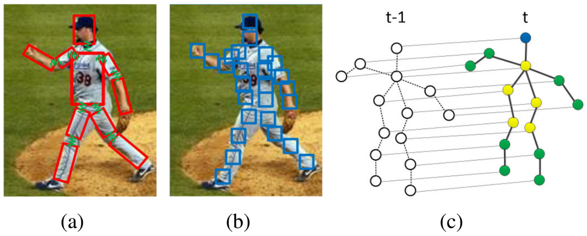

**Fig. 1**. Models for human pose estimation. (a) is the classic pictorial structure model, (b) is the flexible mixture-of-parts model [3], and (c) is our model for video sequences.

In order to deal with the changes in body pose and appearance, a large number of training samples are required for better performance. Pishchulin *et al.* [4] explicitly control pose and shape variations, and learn a general model. Yang *et al.* [3] employ multiple components for each parts. The effectiveness of the general model is validated. However, instance-specific appearance (e.g. color of clothing) can be beneficial to pose estimation, and this specific features are discarded by the general models. The specific features are learned and used for pose estimation in this paper.

Temporal coupling of limb positions is used to reduce search space of body configuration by Ferrari *et al.* [5]. Sapp *et al.* [6] proposed a model with edges within and between frames, involving inference over an ensembles of six models. We adopt a simple function of part displacement as the temporal model, and the inference is efficient.

Our model is shown in Fig. 1(c). All the edges of solid lines are modeled as pairwise relationships, and the dotted lines are only for displaying purpose. The representation is based on the flexible mixture-of-parts model in Fig. 1(b). The full model includes the general model, the instance-specific model, and the simplified temporal model.

This work contributes in two aspects. First, we propose an instance-specific model and a temporal model, and integrate them with the general model into a unified framework that allows for efficient inference and learning. Second, instead of modeling the motion cues as a generative motion model, we

This research has been partially supported by China 973 Program (2011CB302203) and NSFC grants (61273285, 61375019). * indicates corresponding author.

employ a discriminative function to restrict the displacement of body parts, and the effectiveness is validated.

**Related work.** Various features are employed in previous work, such as foreground and background color model [5, 7]. Another kind of color model is color histogram. Contours are used in [8], and descriptor of gradient is applied in [9]. HoG is used for pose estimation in [3].

Temporal limb coupling is used for search space reduction in [5], in which they proposed an integrated spatio-temporal model covering multiple frames. Sapp *et al.* [10] proposed adaptive pose prior for efficient inference. The stretchable model [6] builds pairwise relationships between frames, and then is decomposed into an ensemble of six models which enable tractable inference and learning.

## 2. MODEL

The model is shown in Fig. 1(c), and it is a tree structure. The child parts are independently placed in a coordinate system defined by their parent. The solid lines represent pairwise relationships within and between frames. The dotted lines show the edges that are not modeled to avoid intractable inference.

Assume there is a $K$-part model and a sequence of $T$-frame images from a video. We use superscripts to denote frames $t \in \{1, ..., T\}$, and subscripts to denote body parts $i \in \{1, ..., K\}$. Given an image $I^t$ of the sequence, we write $l_i^t = (x, y)$ for the pixel location of part $i$. As proposed by [3], we employ the same approach and build $m_i \in \{1, ..., M\}$ components for part $i$, where $m_i$ is an indication variable that determines the type of part $i$. To encode the spatial structure of the model, let $G = (V, E)$ be a tree-structured relational graph, where nodes $V$ represent all $K$ parts, and edges $E$ specify pairwise relationships of the nodes.

We would like to score each possible part configuration at the location $l_i^t$. By maximizing the scores, the model can obtain the best configuration. The score we want to maximize for the full model is written as:

$$S(I^t, l, m, t) = \sum_{i \in V} \phi_i + \Omega(m) + \sum_{ij \in E} \psi_{ij} + \sum_{i \in V} \varphi_i^{t-1,t} \ , \quad (1)$$

where $m$ indicate the types of parts. $\Omega(m)$ is score of the part co-occurrence model, and $\sum_{ij \in E} \psi_{ij}$ is score of the deformation model [3]. $\sum_{i \in V} \phi_i$ is score of our appearance model, and $\sum_{i \in V} \varphi_i^{t-1,t}$ is score of the temporal model. These models are detailed as follows.

The global consistency is broken down into pairwise consistency. The co-occurrence model in (1) is composed of local and pairwise score [3]:

$$\Omega(m) = \sum_{i \in V} b_i^{m_i} + \sum_{ij \in E} b_{ij}^{m_i, m_j} \ , \quad (2)$$

where $m_i \in \{1, ..., M\}$ is the type indicator for part $i$, the parameter $b_i^{m_i}$ represents the score of a particular component assignment for part $i$, and $b_{ij}^{m_i, m_j}$ denotes the score of co-occurrence of part $i$ and part $j$.

The score of the deformation model is controlled by the spring models and the relative displacements of parts. The score of deformation model [3, 11] is formulated as follows:

$$\sum_{ij \in E} \psi_{ij} = \sum_{ij \in E} \lambda_{ij}^{m_i, m_j} \cdot \xi(l_i^t - l_j^t) \ , \quad (3)$$

where $\lambda_{ij}^{m_i, m_j}$ is the parameter vector of a spring model governed by component $m_i$ of part $i$ and $m_j$ of part $j$. The displacement vector is written as $\xi(l_i^t - l_j^t) = [dx \ dx^2 \ dy \ dy^2]^T$, where $dx = x_i^t - x_j^t$ and $dy = y_i^t - y_j^t$ are the relative locations of part $i$ and $j$ in frame $t$.

Both general features and instance-specific features contribute to the final performance of pose estimation. One of the most effective general descriptors is HoG [12]. However, some useful information of part appearance has been ignored in the general features. For example, if one wants to obtain the part configurations of a person wearing a red suit, the red color is beneficial to the pose estimation. This color feature is instance-specific, and red color model can not be used to estimate body pose for a person in yellow.

Based on the general and instance-specific features, the score of appearance model $\sum_{i \in V} \phi_i$ in (1) can be expressed as:

$$\begin{aligned}\sum_{i \in V} \phi_i = \sum_{i \in V} [(1 - \theta) \cdot \omega_i^{m_i} \cdot \phi_i^g(I^t, l_i^t) + \\ \theta \cdot \mu_i^{m_i} \cdot \phi_i^s(I^t, l_i^t)] \ ,\end{aligned} \quad (4)$$

where $\phi_i^g(I^t, l_i^t)$ is the general feature vector extracted at location $l_i^t$ of the $t$-th frame image, $\phi_i^s(I^t, l_i^t)$ is the instance-specific feature vector, $\omega_i^{m_i}$ and $\mu_i^{m_i}$ are parameters of the model template, and $\theta$ is the weight which controls the contributions to the overall appearance model.

We apply HoG to form the general features, and use the color models for the instance-specific features. The pixel-level color model is constructed by logistic regression based on samples of body part pixels and background pixels.

For a human body model, there are many parts and each part can vary in appearance or shape, resulting in many degrees of freedom to be estimated. In the experiments, even the similar images can produce totally different body configurations. With inspiration of tracking in videos, the estimated positions of parts can be employed as a prior to parse human body for the next frame. The score of temporal model in (1) is defined as:

$$\sum_{i \in V} \varphi_i^{t-1,t} = \sum_{i \in V} \alpha_i \cdot \left(\frac{1}{1 + \exp^{-\beta_i \cdot d_i^{t-1,t}}} - 0.5\right) \ , \quad (5)$$

where $d_i^{t-1,t}$ is the distance between the relative position of part $i$ with respect to body center in frame $t-1$ and that in frame $t$. The Euclidean distance is adopted for our application. $\beta_i$ is the weight parameter for part $i$, and $\alpha_i$ determines the contribution of part $i$ to the whole model.

Each term of the full model in (1) is either sum over $V$ or sum over $E$. We can rearrange the full model as:

$$S(I^t, l, m, t) = \sum_{i \in V} (\phi_i + \varphi_i^{t-1,t} + b_i^{m_i}) + \sum_{ij \in E} (\psi_{ij} + b_{ij}^{m_i, m_j}) \ , \tag{6}$$

where $\phi_i$ is shown in (4), $\varphi_i^{t-1,t}$ is formulated in (5), $\psi_{ij}$ is in (3), and $b_i^{m_i}$ and $b_{ij}^{m_i,m_j}$ are from (2).

The model involves a large number of parameters. We will illustrate the details of inference and learning in the following Section 3.

## 3. INFERENCE AND LEARNING

In order to estimate body pose of different scales, we compute the scores of the model all over the image pyramid. The maximal score of $S(I^t, l, m, t)$ over all scales and locations of image pyramid is considered as a candidate, and then we trace back the maximizing process and get the body configurations.

### 3.1. Inference

Dynamic programming [13] is an efficient and effective approach. To be specific, we compute the message from the leaves and pass the maximal messages to their parents, and finally to the root part. The message passed to part $j$ from its children is defined as:

$$\mathrm{MS}_j = \sum_{k \in \mathrm{kids}(j)} \max_{l_k, m_k} (\mathrm{MS}_k + \phi_k + \psi_{jk} + \varphi_k^{t-1,t} + b_k^{m_k} + b_{jk}^{m_j, m_k}) \ , \tag{7}$$

where kids$(j)$ denote all the child nodes of part $j$, and $\mathrm{MS}_k$ is the message passed to part $k$ from its child nodes. For a leaf part $z$, the kids$(z)$ is an empty set.

Once the messages are passed to the root part ($j = 1$), the score of the best configuration for root part at location $l$ of image $I^t$ is:

$$\mathrm{score}_r(I^t, l) = \mathrm{MS}_1 + \phi_1 + b_1^{m_1} + \varphi_1^{t-1,t} \ , \tag{8}$$

where $\mathrm{MS}_1$ is the message passed to the root part. The best detection of image $I^t$ can be acquired by maximize the root scores over all locations, while multiple detections (for multiple persons) can be obtained by thresholding the root scores and applying non-maximum suppression [11].

To infer the body configurations from the root scores, we can backtrack the procedure of dynamic programming to determine the location and component of each part. In the experiments, there is only one score for each location $l$ anchored with root part. If multiple detections at the same location are required, one can use the approach of Park *et al.* [13].

### 3.2. Learning

As presented in Section 2, the model involves general features which mainly describe body shapes and poses, and instance-specific features which depict part appearance. The training is divided into two stages.

We intend to estimate the parameters related to general features first, so $\theta$ in (4) and $\alpha$ in (5) are set to 0 in the first stage. Then, the score function of (6) can be written as:

$$S'(I, l, m) = \sum_{ij \in E} (\lambda_{ij}^{m_i, m_j} \cdot \xi(l_i^t - l_j^t) + b_{ij}^{m_i, m_j}) + \sum_{i \in V} (\omega_i^{m_i} \cdot \phi_i^g(I, l_i) + b_i^{m_i}) \ . \tag{9}$$

Note that it is the same score as that of flexible mixtures of parts proposed in [3]. The score $S'(I, l, m)$ is a linear function of parameters $\gamma = (\omega, \lambda, b)$ , and all the general features can be written as a corresponding vector $\Phi(I, l, m)$, with class label $y \in \{1, -1\}$. Then the score function is:

$$S'(I, l, m) = \gamma^T \Phi(I, l, m) \ . \tag{10}$$

Linear Support Vector Machines is used to estimate $\gamma$ by solving the following optimization problem:

$$\min_{\gamma} \ \frac{1}{2}\gamma^T \gamma + C \sum_{i=1}^{n} \max(0, 1 - y_i \cdot \gamma^T \Phi_i) \ . \tag{11}$$

After the parameter $\gamma$ is learned, we want to determine the parameters $\alpha$ and $\beta$ in (5). We employ a simple grid search method to find the values for $\alpha$ and $\beta$.

We adopt a semi-supervised method to train the color model. Each video clip is divided into two parts. The first small part is used for training, and the remaining for testing. The instance-specific model is learned for each video respectively. Based on the learned general pose model without instance-specific features, poses with high confidence scores are detected and estimated for training. The false positives are manually removed from the training data. The instance-specific features are modeled by a pixel-level color model. The score of color model is written as:

$$\mathrm{score}^a = \mu_i \cdot \phi_i^s(I^t, l_i^t) \ , \tag{12}$$

where the $\mu_i$ is trained by logistic regression, with pixels of body part $i$ as positive samples and pixels of its background as negative samples.

## 4. EXPERIMENTS

We report results on VideoPose2.0 dataset [6] and UCF Sports Action dataset [14]. We also utilize Image Parse dataset

**Table 1**. Mean APK and mean PCK of keypoints of shoulder, elbow and wrist on the VideoPose2.0 dataset.

| Method | Mean APK | Mean PCK | Time cost (s) |
|---|---|---|---|
| Yang *et al.* [3] | 69.1 | 78.8 | 1.17 |
| Us HoG+T. | 69.6 | 79.9 | 1.22 |
| Us HoG+C. | 70.7 | 80.0 | 1.31 |
| Us HoG+C.+T. | 69.7 | 81.0 | 1.60 |
| Sapp *et al.* [6] | 82.4 | 87.7 | 340.76 |

[7] and the Buffy Stickmen dataset [5, 15] for training. VideoPose2.0 dataset consists of 44 clips, with 1286 frames. UCF Sports Action dataset is a challenging dataset for action recognition. We choose a few video clips as the testing set and annotate them to evaluate full body pose estimation.

### 4.1. Evaluation

Percentage of correctly estimated keypoints (PCK) and Average Precision of Keypoints (APK) [3] are used as the evaluation measure. A candidate keypoint is considered to be correctly estimated if it falls within $\alpha \cdot \max(h, w)$ pixels of ground-truth keypoint, where $h$ and $w$ are the height and width of the bounding box. APK establishes average precision for each keypoint separately. $\alpha$ is set to $0.2$.

### 4.2. VideoPose2.0 database

VideoPose2.0 dataset is used to estimate upper body pose. The general body model learned from HoG features is applied to estimate primary pose with confidence scores. Then an instance-specific pixel-level color model is learned for each keypoint by samples with high confidence for each video sequence. The models are combined in the proposed framework to estimate the final pose. In order to evaluate the instance-specific model and the temporal model separately, we report results of three approaches, "HoG+C.", "HoG+T." and "HoG+C.+T.", where "HoG" is the general body model, "C." is the pixel-level color model, and "T." is the temporal model. $\theta$ in (4) is set to 0.3 by training, and $\alpha_i$ and $\beta_i$ in (5) are set to 1 and $0.01$, respectively.

The head can be localized with high accuracy, so we focus on parts of shoulder, elbow and wrist. The mean APK and mean PCK are shown in Table 1. We compared our method with those of Yang *et al.* [3] and Sapp *et al.* [6]. The comparison was under the same condition. The image resolution is $370 \times 330$. Our methods achieve better results on both APK and PCK than [3], and consume less time for each frame than [6], as shown in Table 1.

### 4.3. UCF Sports Action dataset

The UCF Sports Action dataset is used to estimate full body pose. The settings are the same as those of VideoPose2.0 dataset, except $\beta_i$ in (5) being 0.1. The mean APK of all keypoints are shown in Table 2. "HoG+C." reports the mean APK of $60.8\%$, in contrast to $52.8\%$ by the method [3]. By "HoG+T.", mean APK $54.3\%$ is reported. Both the color model and the temporal model contribute to the performance of pose estimation. The full model "HoG+C.+T." achieves the best results of mean APK $65.2\%$.

**Table 2**. Mean APK of all the keypoints on the UCF Sports Action dataset.

| Method | Mean APK |
|---|---|
| Yang *et al.* [3] | 52.8 |
| Us HoG+T. | 54.3 |
| Us HoG+C. | 60.8 |
| Us HoG+C.+T. | 65.2 |

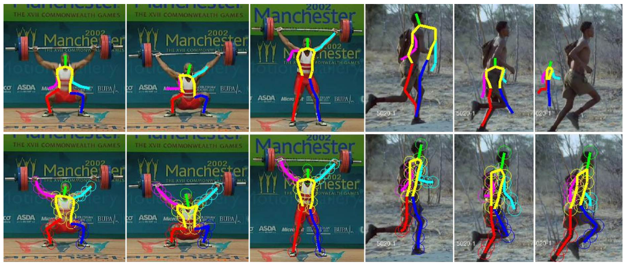

**Fig. 2**. Pose estimation on UCF Sports Action dataset. The upper row shows the results by method [3], and the lower row shows the results that obtained by our full model.

The example images are shown in Fig. 2. Due to the instance-specific model, there are less false positives (keypoints of arm in the three left columns) by our method. As well, the temporal model reduces the confusion of between-frame estimation, as shown in the three right columns.

### 4.4. Analysis

The instance-specific appearance model helps to improve the overall performance, as shown in Table 1 and Table 2. However, if the images fail to provide effective color information, adding color model to the framework will weaken the whole system. The strategy is to manipulate $\theta$ in (4). Another problem lies in the confidence score. In most of the cases, higher confidence score means better pose estimation, but in the experiments some false positives have high confidence.

## 5. CONCLUSION

We have presented a body model that incorporates appearance model, deformation model, co-occurrence model and temporal model for pose estimation in videos. We analyze the detailed effects of color model and temporal model, and demonstrate that both models are effective to estimate body pose. Nevertheless, there is still much space for improvement. More effective instance-specific models are needed for a robust system. Pose estimation for human body with self-occlusion is still a challenging problem.

## 6. REFERENCES

[1] M. A. Fischler and R. A. Elschlager, "The representation and matching of pictorial structures," *IEEE Transactions on Computers*, vol. 22, no. 1, pp. 67–92, Jan. 1973.

[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, Jan. 2005.

[3] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2878–2890, 2013.

[4] L. Pishchulin, A. Jain, M. Andriluka, T. Thormahlen, and B. Schiele, "Articulated people detection and pose estimation: Reshaping the future," in *CVPR*. IEEE, 2012, pp. 3178–3185.

[5] V. Ferrari, M. Marin-Jimenez, and A. Zisserman, "Progressive search space reduction for human pose estimation," in *CVPR*. IEEE, 2008, pp. 1–8.

[6] B. Sapp, D. Weiss, and B. Taskar, "Parsing human motion with stretchable models," in *CVPR*, June 2011, pp. 1281–1288.

[7] D. Ramanan, "Learning to parse images of articulated bodies," in *NIPS*, 2006, vol. 1, p. 7.

[8] P. Srinivasan and J. Shi, "Bottom-up recognition and parsing of the human body," in *CVPR*, June 2007, pp. 1–8.

[9] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," in *CVPR*. IEEE, 2009, pp. 1014–1021.

[10] B. Sapp, C. Jordan, and B. Taskar, "Adaptive pose priors for pictorial structures," in *CVPR*. IEEE, 2010, pp. 422–429.

[11] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[12] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*. IEEE, 2005, vol. 1, pp. 886–893.

[13] D. Park and D. Ramanan, "N-best maximal decoders for part models," in *ICCV*, Nov 2011, pp. 2627–2634.

[14] M.D. Rodriguez, J. Ahmed, and M. Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition," in *CVPR*, June 2008, pp. 1–8.

[15] M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari, "2d articulated human pose estimation and retrieval in (almost) unconstrained still images," *International Journal of Computer Vision*, vol. 99, no. 2, pp. 190–214, 2012.