Torso Orientation: A New Clue for Occlusion-Aware Human Pose Estimation

Yang Yu, Baoyao Yang and Pong C Yuen*
Department of Computer Science
Hong Kong Baptist University
{yangyu, byyang, pcyuen}@comp.hkbu.edu.hk

***Abstract*— Self-occlusion is a challenging problem existing in human pose estimation. In this paper we exploit a new cue to solve this problem: the torso orientation. We describe a technique to automatically detect self-occlusion in training set without visibility label. Given this prior information, we are able to jointly learn an occlusion-aware model to capture the pattern of self-occluded body parts. We evaluate our model on two major datasets, which are both publicly available. The experiment result shows that our model is quite competitive in both of the datasets with the state-of-the-arts. By this way, we illustrate our model's robustness to the self-occlusion problem in human pose estimation.**

Keywords—computer vision; pose estimation; articulated model; self-occlusion

I. INTRODUCTION

The target of 2D human pose estimation is to generate a configuration of human body parts from a still image. It plays a vital role to many other high-level computer vision topics, e.g. 3D human pose reconstruction, activity recognition and human-computer interaction. Most of the state-of-the-art methods adopt the pictorial structure (PS) [1] to describe human body, which is made of two components: the local appearance templates of each part and the geometric constraints between pairs of parts. Although a great amount of pose estimation algorithms has been proposed in recent years, this problem still remains challenging due to the occlusion, appearance diversity, complex background, camera viewpoint and the large variation of body part configuration.

In this paper, an occlusion-aware model is developed to deal with the problem of self-occlusion, using the torso orientation as prior information. In the existing methods of human pose estimation, detection failures often occur while a body part is occluded by the others. To address this problem, we exploit the observation that the self-occlusion pattern is highly related to where the person is facing with respect to the camera viewpoint. With the information of torso orientation, we can infer the position of self-occluded part, since they tend to share similar appearance and deformation pattern under the same condition. For example, if a still image contains a person with his torso heading the right and the two wrists are determined to be overlapped, then it is highly possible that the right wrists is occluding the left one. We develop our model based on the famous flexible mixtures-of-parts (MoP) model [17], which models each body part with a collection of 'types'. We use an extra body part to represent the torso and one more type is assigned to each part to model self-occlusion relationship. Notably, we do not require a visibility label in the training data, but learn it automatically. By experiment we show that our methods not only out-performed the original but also better than other recent methods.

This paper is organized as follows: In Sec.II we introduce the related works to deal with the occlusion problem in pose estimation. In Sec.III we firstly review the structure of the Mixture of Parts model. Then we introduce our idea, namely a self-occlusion handling process using the torso direction as a cue. At last we will illustrate the inference and learning procedure in detail. In Sec.IV we show and analyze the experiment result to support our idea. The key contributions of this paper are:

- A new kind of prior information is introduced to address the self-occlusion problem in human pose estimation.
- A method to generate the occlusion relationship without requiring the visibility ground truth annotation.

II. RELATED WORK

The most popular model in human pose estimation in recent year is the part-based model, which considers both of the local appearance of each part and the pairwise spatial constraints on adjacent. However, the input information is somewhat insufficient to determine the global configuration of human body, since the local appearance is often confused by the background or similar parts e.g. the limbs look similar to their symmetrical correspondents. Moreover, the large degree of freedom of the human pose often leads to ambiguity of the spatial constraints. Yang and Ramanan proposed the MoP model [17] that considers a collection of the appearance under different conditions, which is one of the most successful models in performance. Observing that the MoP model usually fails when there is occlusion in test images due to the misdetection of appearance of certain parts, we proposed a method to improve it by introducing an occlusion prior. Unlike some other methods, we don't require visibility ground truth labels at training stage.

The existing methods aiming to handle the occlusion problem in pose estimation can be categorized into two groups: (1) miscellaneous spatial models and (2) occlusion pattern

*Corresponding author

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

modeling. For the first group, Sigal and Black [3] extended the tree model with additional occlusion constraints to encode the occlusion relationship between parts. Moreover, Tran and Forsyth [4] connected all parts from upper body together to generate the "full relational model" of human body. Since the extra loops are introduced to the original tree structure, it will not be able to implement dynamic programming in inference stage, which makes it a NP-hard problem. Apart from the loopy models, Wang and Mori [6] combined multiple tree models together to reason the occlusion in pose estimation. Based on this idea, Johnson et al. [7] partitioned the space of human pose by clustering the relative positions between each joint and neck joint. In a recent paper, Chen [8] exploited the idea that the visible nodes form a subtree under occlusion, which generates a more reasonable solution space. Also, a very recent work uses an "unrolling" technique to accelerate the inference on non-tree structure [5].

Although a lot of work has been done on the spatial relationship, there is an inherent dilemma in this family of method: The traditional tree structure is simple and straightforward, but usually fails to capture the high level interaction between body parts; the graphical models with loops allow for complicated part relationship, but they are not eligible for efficient and exact inference process [2]. Meanwhile some works aim to utilize the occlusion patterns. There are proposed approaches aiming to model occlusion by segmenting the feature map [9, 10]. While the "poselet" was introduced to directly capture the pattern of body parts interaction [11], Desai and Ramanan extend it to model the human body and occluding objects altogether [12]. Also, the grammar-based models [13] and the strongly supervised deformable part models [14] contain explicit occlusion part templates. Ghiasi et al improved this idea by using a mixture of templates to model the occlusion pattern through a non-parametric way [15]. Additionally, Radwan et al. [16] adopted Twin-GP regression as a post process to rectify the 2D pose estimation.

III. THE MODEL

A. Mixture of Parts(MoP) Model

Given an image I, we can define a graph **G = (V, E)** as a PS model. The nodes in **V** represents the parts, while the edges in **E** stand for the pairwise spatial constraints between the neighboring parts. Each part *i* is presented by two parameters combined: (i) the pixel location ***p*** and (ii) the mixture type ***t***. Therefore, the whole collection of $p_i = (x_i, y_i)$ and t_i can sufficiently describe a global human pose configuration, which is denoted as **(p, t)**, where **p** $=[p_1 ... p_n]^T$ and **t** $=[t_1 ... t_n]^T$, **n = |V|**. The types can be defined by the relative position or semantic classes of parts, e.g. the vertical versus horizontal arms and the open versus close hands. Now the score of a configuration can be defined as following equations:

$$S(I,p,t) = S(t) + \\ \sum_{i\in V} \omega_i^{t_i} \bullet \qquad\qquad {}_{i,j\in E} \qquad \bullet \qquad \textbf{(1)}$$

Where the former unary term $\phi(I, p_i)$ is the feature vector and the later binary term $\psi(p_i - p_j) = [dx, dx^2, dy, dy^2]^T$, measures the deformation, and $dx = x_i - x_j, dy = y_i - y_j$. The parameter $\omega_i^{t_i}$ is the appearance template to be learnt. And the parameter $\omega_{ij}^{t_i,t_j}$ is the deformation model, which can be interpreted as a "spring" that connects a particular pair of parts *i, j* with respect to their types (t_i, t_j) .

The first term $S(t)$ is the compatibility function for part types, which can be divided into two terms as well:

$$S(t) = \sum_{i\in V} b_i^{t_i} + \sum_{i,j} b_{ij}^{t_i,t_j} \qquad \textbf{(2)}$$

The parameter $b_i^{t_i}$ prefers particular type assignment of part *i*, and the parameter $b_{ij}^{t_i,t_j}$ is biased on certain co-occurrences of connected parts' types. An example is that if two parts are form the same rigid limb, then $b_{ij}^{t_i,t_j}$ would be bias to that they have the same orientation.

Notably, the original model uses the tree structure to represent human body parts. Alternatively, we adopted a star-structure with an extra parts correspondent to the torso type to better capture the kinematic constraints, as shown in Figure.1.

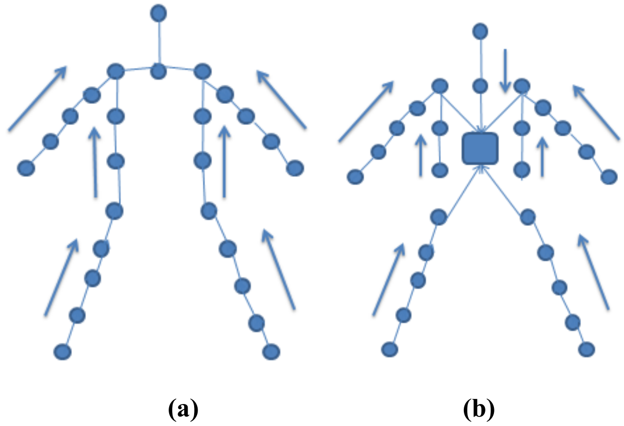

(a) **(b)**

Figure 1. This figure shows the difference of graphical model between (a) the original base method and (b) our modified model. Note that the original model has 26 parts and or model has an extra torso part.

B. Our Model

It is intuitive that the pattern of self-occlusion is highly related to the torso orientation: Due to some kinematic constraints of human body, the diversity of limbs is limited given a specific direction of torso. Therefore, the status of torso can imply the limbs' location to some extend. In order to model the body part under the presence of self-occlusion, we assigned an extra type to each part, the "occluded" type. We expect that this extra type can be used to pick out the occluded part. By this mean, we can model particular patterns of them respectively.

To acquire the occlusion relationship, we use the information of torso orientation instead of the ground truth visibility labels. Our basic idea is that the torso, as the most inflexible part of human body, can imply the layer of other human parts regarding to the camera viewpoint. Thus, it will be reasonable to speculate the occlusion between two parts that have close ground truth locations.

Inspired by this idea, we add an extra torso part to the original model, denoted by (p_0, t_0). The position p_0 can be defined by the barycenter of its children or it can be defined by a weighted average of certain combination of parts, while the parameter t_0 belongs to a set T_0 of torso direction. Although we uses four directions in our experiment, namely facing left, right, to the camera and away from camera, the set of T_0 can be amplified by finer classification of torso direction, e.g. facing up versus facing down.

In practice, we classify the torso types by the physiological constraints of human body, including the ratio of torso length and shoulder width, the direction of knees and elbows and the relative position of head and neck, etc.

With the new clue give above, we follow the inference and learning paradigm used in the base model [17].

C. Model Inference

The inference process corresponds to the maximization of the model score defined by equation (1) in terms of **p** and **t**. Given that our graphical model is tree-structured, we can implement the inference algorithm efficiently by using the dynamical programming. The score of leaves can be computed in a sliding-window fashion exhaustively, and then be passed to their parents. Thus the score of a configuration can be obtained accordingly. The message passing process can be described by following:

$$score_i(t_i, p_i) = b_i^{t_i} + \omega_{t_i}^{i} \bullet \qquad \sum_{k \in kids(i)} m_k(t_i, p_i) \tag{3}$$

$$\begin{aligned} m_i(t_j, p_j) = {} & \max_{t_i} b_{i,j}^{t_i,t_j} + \\ & \max_{p_i} score(t_i, p_i) + \omega_{i,j}^{t_i,t_j} \bullet \qquad {}_{i}) \end{aligned} \tag{4}$$

Where the kids(i) means the children set of part i, and the part j is the parent node of part i. While the equation (6) computes the local score for all pixel locations of p_i and for all possible mixture types t_i, the equation (7) derives the location with the highest score over all location p_j and all possible type t_j. Thus the score of the leaf nodes can be passed along the tree and eventually reached the root part. Instead of simply picking up the best root score, non-maximum suppression is adopted, that is, for each spatial neighborhood, a local maximum is picked. By this way, multiple detections can be spread out through the image to increase the chance of true detection. Given the previous outcome of $\arg\max$, it will be efficient to obtain the location and type of each part in each maximal configuration by back tracking.

In terms of computation, for each part one has to loop over $L \times T$ possible parent locations and types, and compute the maximum over $L \times T$, meaning the computation $O(L^2T^2)$ for each part, where the L denotes all possible locations and the T denotes all mixture types. In this case, $\psi(p_i, p_j)$ is a quadratic function, thus the inner maximization (4) can be efficiently computed for each type combination in $O(L)$ with a max-convolution or distance transform. Since T^2 distance transforms must be done, the message passing cost is $O(LT^2)$ for each part.

D. Model Learning

The model learning process follows the supervised learning paradigm. The input data contains the positive examples $\{I_n, p_n, t_n\}$ and negative examples $\{I_n\}$. To define a structured prediction objective function like in [26], we denote $z_n = (p_n, t_n)$. Since the parameters $\beta = (\omega, b)$ in (1) are linear, the equation (1) can be rewritten as $S(I, z) = \beta \bullet$. The learning form is as following:

$$\begin{aligned} & \arg \min_{\omega, \varepsilon_i \geq 0} \frac{1}{2} \beta \bullet \qquad {}_{n} \quad \varepsilon_n \\ & s.t. \forall n \in pos, \beta \bullet \qquad \geq 1 - \varepsilon_n \\ & \forall n \in neg, \forall z, \beta \bullet \qquad \lesssim -1 + \varepsilon_n \end{aligned} \tag{5}$$

As illustrated in above, the positive examples should score better than1, while the negative ones ought to score be less than -1 for all configurations of parts and types. The slack variables ε_n are designed to punish the violations of these constraints in the objective function.

While implementing the learning process in practice, the ground truth label of visibility is unnecessary. To acquire the type of torso, we use some inherent physiological constraints in human body to obtain the torso types. Given the type of torso, we can speculate the layer of human body with respect to the camera viewpoint. Afterwards, we use the ground truth box of each part to determine if it's overlapped. If so, the parts on the top layer of human body are very likely to occlude the parts behind it.

For those which are not occluded, the K-means technique is used to generate part type from relative position. In our experiment, the relative position of each part and its parent is clustered, and types for all part *I* can be defined as demonstrated above.

IV. EXPERIMENT

In this section, we evaluate the proposed torso-direction model on two popular pose estimation benchmarks. The results on the some benchmarks of other state-of–the-art approaches are compared to show difference. In terms of performance measure, we use the popular PCP metric [24], which stands for "Percentage of Correct Pose". It measures the accuracy by matching the connection between joints with the ground truth.

Datasets: To evaluate our model, we used two publicly available pose estimation dataset containing miscellaneous human body configuration and camera viewpoint: (1) The "Leeds Sports Poses" (LSP) dataset [7], containing 2000 images show people involved in various sports, where we use former1000 images for training and later 1000 images for testing. (2) The "Image Parsing" (IP) dataset [25] containing 305 images of fully visible people performing a diversity of activities, where we use first 100 images for training and the rest 205 for testing.

Implement details: We extended the base model with our self-occlusion handling strategy. Firstly, as shown in Figure 1(b), 27 body parts are derived from the annotation. For each body part, except for the torso part, we define 7 mixture types i.e. one more extra type than the original model. For the torso part, we define 4 mixture types, corresponding to the body facing left/right and to/away the camera. The non-person images from INRIA dataset are used as the negative examples. The part filter of our model is based on the rigid HOG templates [23].

A. Results on LSP

Table I shows the results of our model and other state-of-the-art models in term of PCP metric on LSP dataset. To illustrate the ability to deal with the highly flexible parts, we pick out the average PCP score of limbs of each model. As we can see, our method outperforms all other methods under a standard PCP metric using PC annotation, especially in the detection of limbs. Please note that we do not use the Observer-Centric (OC) annotation, i.e. left/right body parts are marked according to the camera viewpoint. The OC annotation is not able to tell if the person is facing the camera or turn his back around. Although the OC annotation reduces complexity in training stage, it does not provide enough information to deal with the self-occlusion. Therefore, in our torso type definition these two situations are separated, as we use the Person-Centric annotation to training a discriminative model to address the self-occlusion problem.

TABLE I. PCP[a] VALUES ON LSP DATASET

Method	Upper Leg	Lower leg	Upper Arm	Lower Arm	Avg Limbs	Avg All
Our Approach	**78.1**	66.1	61.2	**42.7**	**62.0**	**69.1**
Base-Model [17]	74.9	63.3	61.3	41.6	60.3	67.8
Fu et al.[5]	74.2	66.8	**62.5**	41.3	61.2	66.9
Wang et al.[6]	74.0	**69.8**	48.9	32.2	56.2	62.8
Johnson et al.[20]	74.5	66.5	53.7	37.5	58.0	62.7
Tian et al.[18]	69.9	60.0	51.9	32.9	53.7	61.3
Dantone et al.[19]	66.5	61.0	45.1	24.7	49.3	55.5

[a] The threshold of PCP metric is 0.5

B. Results on Image Parse

Table II shows the PCP score of our model and other four techniques, including our base model. As we can see, given the threshold of 0.5, the base model performs better than ours. However, if we restrict the PCP threshold to a more strict level, we can show that our model outperforms the original model in the average PCP score. This means that our model has a better performance when we have a higher requirement of the detection accuracy. The comparison between these two methods under different PCP threshold is illustrated in Figure.2. If we consider another frequently used criterion, Percentage of Correct Keypoints (PCK) [27], our model is also better than the base model on the whole dataset. Moreover, our model excels more significantly while we only consider the self-occluded images, as shown in table III.

Additionally, in term of qualitative results, Figure.3 shows the qualitative comparison between our model and the base model. In the first row, because the baseball player is side viewed, while the base model failed to find his left arm, it will take a speculation according to the global configuration. But in our model, the pattern of torso direction can be detected and used to infer the occlusion between his left arm and his body. Similarly, in the second row, the left arm is occluded by the person's left shoulder. In our model, we can reason its location out of this occlusion pattern.

TABLE II. PCP VALUES ON IP DATASET

Method	Upper Leg	Lower Leg	Upper Arm	Lower Arm	Avg Limbs	Avg All
Our Approach	**89.9**	**82.8**	74.4	50.9	74.5	79.3
Base-Model[17]	88.4	80.8	**81.3**	**53.4**	**76.0**	**80.6**
Tian et al.3 Layers[21]	85.1	76.1	71.0	45.1	69.3	74.4
Tian et al. 4 Layers	81.2	71.0	69.5	39.0	65.1	71.0
Johnson et al.[20]	73.4	65.4	64.7	46.9	62.6	66.2
Andriluka et al.[22]	63.2	55.1	47.6	31.7	49.4	55.2

TABLE III. PCK[a] VALUES ON IP DATASET

PCK	Avg All Limbs	Avg All	Avg SO Limbs	Avg SO
OUR	**65.5**	**73.3**	**54.8**	**62.9**
BASE	64.6	73.1	52.2	60.7

[a] The threshold of PCK is 0.1

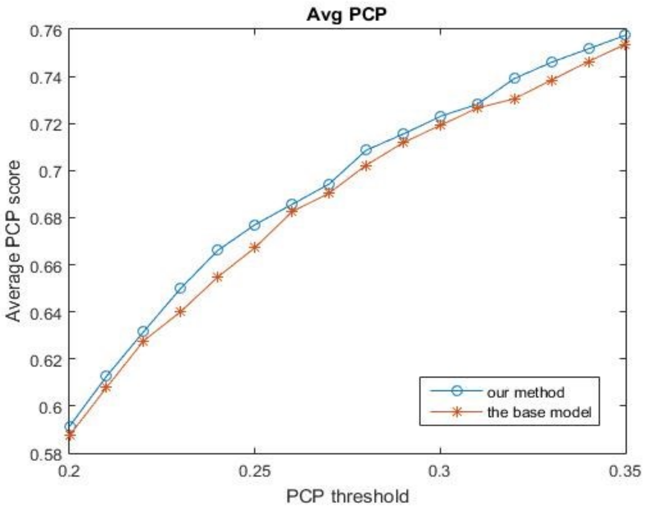


Figure 2. The average PCP score of our model and the original model under the threshold of 0.35.

V. CONCLUSION AND FUTURE WORK

In this paper we introduce a new prior into human pose estimation, which is the torso orientation. Given the torso orientation, we have a new cue to learn an occlusion-aware model. By this way, we address the self-occlusion problem at the testing stage. We demonstrate that our extended model is competitive compared to other cutting edge techniques by experiment. In the later future, we will try to exploit a hierarchy of overlapping body parts. And use the layer information to eliminate the ambiguity caused by self-occlusion.

VI. ACKNOWLEDGEMENT

This project is partially supported by Hong Kong RGC General Research Fund HKBU 212313 and HKBU 11202514.

REFERENCES

[1] Felzenszwalb, Pedro F., and Daniel P. Huttenlocher. "Pictorial structures for object recognition." International Journal of Computer Vision 61.1 (2005): 55-79.

[2] Koller, Daphne, and Nir Friedman. Probabilistic graphical models: principles and techniques. MIT press, 2009.

[3] Sigal, Leonid, and Michael J. Black. "Measure locally, reason globally: Occlusion-sensitive articulated pose estimation." Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. Vol. 2. IEEE, 2006.

[4] Tran, Duan, and David Forsyth. "Improved human parsing with a full relational model." Computer Vision–ECCV 2010. Springer Berlin Heidelberg, 2010. 227-240.

[5] Fu, Lianrui, Junge Zhang, and Kaiqi Huang. "Beyond Tree Structure Models: A New Occlusion Aware Graphical Model for Human Pose Estimation." Proceedings of the IEEE International Conference on Computer Vision. 2015.

[6] Wang, Yang, and Greg Mori. "Multiple tree models for occlusion and spatial constraints in human pose estimation." Computer Vision–ECCV 2008. Springer Berlin Heidelberg, 2008. 710-724.

[7] Johnson, Sam, and Mark Everingham. "Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation." BMVC. Vol. 2. No. 4. 2010.

[8] Chen, Xianjie, and Alan L. Yuille. "Parsing occluded people by flexible compositions." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

[9] Wang, Xiaoyu, Tony X. Han, and Shuicheng Yan. "An HOG-LBP human detector with partial occlusion handling." Computer Vision, 2009 IEEE 12th International Conference on. IEEE, 2009.

[10] Gao, Tianshi, Benjamin Packer, and Daphne Koller. "A segmentation-aware object detection model with occlusion handling." Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011.

[11] Bourdev, Lubomir, Subhransu Maji, and Jitendra Malik. "Describing people: A poselet-based approach to attribute classification." Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE, 2011.

[12] Desai, Chaitanya, and Deva Ramanan. "Detecting actions, poses, and objects with relational phraselets." Computer Vision–ECCV 2012. Springer Berlin Heidelberg, 2012. 158-172.

[13] Girshick, Ross B., Pedro F. Felzenszwalb, and David A. Mcallester. "Object detection with grammar models." Advances in Neural Information Processing Systems. 2011.

[14] Azizpour H, Laptev I. Object detection using strongly-supervised deformable part models[M]//Computer Vision–ECCV 2012. Springer Berlin Heidelberg, 2012: 836-849.

[15] Ghiasi, Golnaz, et al. "Parsing occluded people." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014.

[16] Radwan, Ibrahim, Abhinav Dhall, and Roland Goecke. "Monocular image 3d human pose estimation under self-occlusion." Proceedings of the IEEE International Conference on Computer Vision. 2013.

[17] Yang, Yi, and Deva Ramanan. "Articulated pose estimation with flexible mixtures-of-parts." Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011.

[18] Tian, Yuandong, C. Lawrence Zitnick, and Srinivasa G. Narasimhan. "Exploring the spatial hierarchy of mixture models for human pose estimation." Computer Vision–ECCV 2012. Springer Berlin Heidelberg, 2012. 256-269.

[19] Dantone, Matthias, et al. "Human pose estimation using body parts dependent joint regressors." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013.

[20] Johnson, Sam, and Mark Everingham. "Learning effective human pose estimation from inaccurate annotation." Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011.

[21] Tian, Yuandong, C. Lawrence Zitnick, and Srinivasa G. Narasimhan. "Exploring the spatial hierarchy of mixture models for human pose estimation." Computer Vision–ECCV 2012. Springer Berlin Heidelberg, 2012. 256-269.

[22] 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.

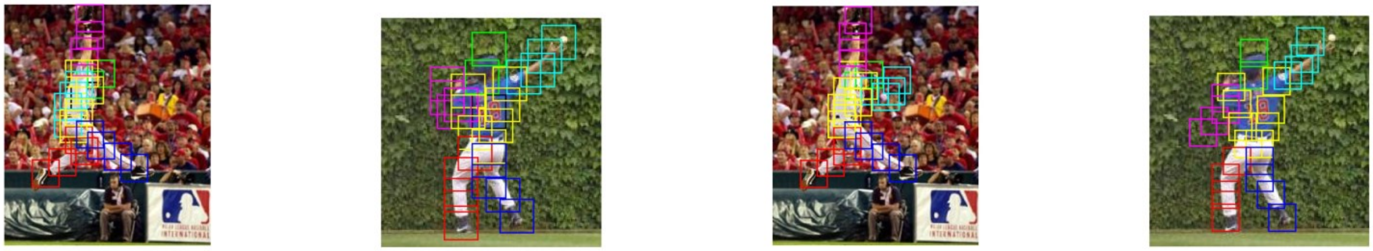
[23] Dalal, Navneet, and Bill Triggs. "Histograms of oriented gradients for human detection." Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. Vol. 1. IEEE, 2005.

[24] Ferrari, Vittorio, Manuel Marin-Jimenez, and Andrew Zisserman. "Progressive search space reduction for human pose estimation." Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, 2008.

[25] Ramanan, Deva. "Learning to parse images of articulated bodies." Advances in neural information processing systems. 2006.

[26] Felzenszwalb, Pedro F., et al. "Object detection with discriminatively trained part-based models." Pattern Analysis and Machine Intelligence, IEEE Transactions on 32.9 (2010): 1627-1645.

[27] Yang, Yi, and Deva Ramanan. "Articulated human detection with flexible mixtures of parts." Pattern Analysis and Machine Intelligence, IEEE Transactions on 35.12 (2013): 2878-2890.



Our model

Base model

Figure 3. The qualitative results of our model and the based model.