CONTEXT AWARE MODEL FOR ARTICULATED HUMAN POSE ESTIMATION

Lianrui Fu, Junge Zhang, Kaiqi Huang

National Lab of Pattern Recognition (NLPR)
Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing 100190, China

ABSTRACT

Simple tree model prevails for 2D pose estimation for its simplicity and efficiency. However, the limited kinetic constraints often lead to double-counting and damage the accuracy of leaf parts, and this is largely ignored in previous work. In this paper, we propose a novel enhanced tree model which incorporates both local kinetic constraints and global contextual constraints among non-adjacent parts. By introducing virtual parts, we are able to model richer constraints within a tree structure and dynamic programming can be utilized for efficient inference. Experiments on public benchmarks show that our method is more effective in tackling double counting problem and can improve the localization accuracy, especially for the challenging lower limbs.

***Index Terms*—** Articulation, Pose Estimation, Part Based Model, Mixture, Degree of Freedom

1. INTRODUCTION

The task of 2D human pose estimation is to detect the presence of human and localize their body parts. It is important for action recognition, human computer interaction (HCI), and video analysis etc. The task is challenging due to cluttered background, articulation and occlusion.

The most successful pictorial structured model(PSM) [1] is first introduced to pose estimation by Felzenszwalb and Huttenlocher [2]. The human body configuration is represented as a collection of independent parts with pairwise connections. The pairwise part relationships are embodied in tree models [3, 4, 5, 6, 7, 8], multi-tree model [9] or loopy graphic models [10, 11, 12, 13, 14].

Tree models prevail for its simplicity and exact inference. For instance, Yang and Ramanan [4] proposed the tree-structured flexible mixture parts(FMP) model which can capture pairwise spatial relations between locally connected parts, and it was followed by the hierarchical tree model [6] and latent tree model [7]. However, existing tree structured models are insufficient in capturing the relationships of non-connected body parts, such as symmetric limbs. This often leads to confuse between left and right limbs and cause the so called double counting problem. Fig. 1 reflects that the simple tree structured FMP is prune to double counting with large deformation and partial occlusion.

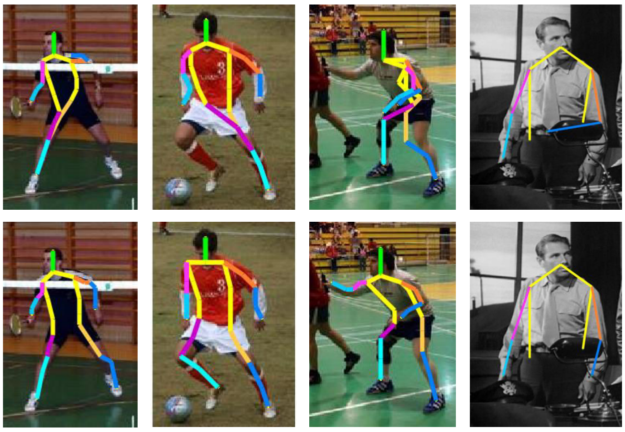

Fig. 1. Human pose estimation results from Yang et.al [4] (top row) and ours (bottom row).

To overcome this issue, Xiao et.al [15] proposed to use multiple model and recombine the detection results. Wang et.al [9] utilized multi-tree model to alleviate the limitations of a single tree-structured model. Some researchers adopted loopy-graph models [10, 12, 13, 14] and even fully connected graph [11]. Though loopy models allow more complex relationships among parts, they can only get approximate solutions with high computational cost.

Our context aware model improves the flexible mixture-of-parts model [4] in three aspects. First, we model not only local kinetic constraints but also global contextual constraints among non-adjacent symmetric parts. This is helpful when there is weak image evidence for one side of the body parts, such as occlusion. Then the context aware model maintains in tree structure by introducing virtual parts, so that dynamic programming can be utilized for efficient inference. Further, we propose to use phraselet clustering to learn the local mixtures for each part to encode global context. This differs from that of clustering parts according to the relative position from the parent node [4] or by the part appearance [7]. Compared with those loopy graphs of [10] and [14] which also model constraints among non-adjacent parts, our model is tree structured and can be inferred efficiently with dynamic programming. Contrast to previous tree structure, our estimation is more effective in tackling double counting problem and can improve the localization accuracy of challenging lower limbs significantly.

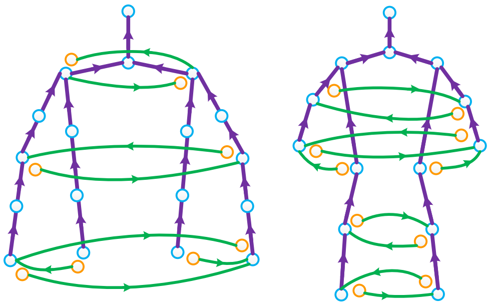

Fig. 2. Structures of the proposed models for upper-body pose and full-body pose estimation. The nodes colored in blue and orange denote real parts and virtual parts respectively. The purple edges represent the kinetic constraints between physically connected parts, and the green edges are the enhanced edges used for modeling spacial context among non-adjacent parts. The arrows show the direction of message passing.

2. OUR APPROACH

In this section we will first briefly overview the Flexible Mixtures-of-Parts model [4], and then introduce the representation of our context aware model as well as the phraselet clustering, finally we will describe the inference and learning procedure.

2.1. Pictorial Structured Mixtures of Parts

Given an image I, let $G = (V, E)$ be the tree structured FMP model, where V is the set of parts and E is the set of pairwise constraints between connected parts. Each part i is parameterized by (p_i, t_i), where $p_i = (x_i, y_i)$ is the part location and t_i is the mixture type of templets for part i. Let $(\mathbf{p}, \mathbf{t})$ represent the pose configuration, where $\mathbf{p} = [p_1, \cdots, p_{|v|}]^T$ and $\mathbf{t} = [t_1, \cdots, t_{|v|}]^T$. The goal is to maximize the pose configuration score $S(I, \mathbf{p}, \mathbf{t})$, which is composed of part appearance score $S^A(I, \mathbf{p}, \mathbf{t})$ and deformation score $S^D(I, \mathbf{p}, \mathbf{t})$ as follows:

$$S(I, \mathbf{p}, \mathbf{t}) = S^A(I, \mathbf{p}, \mathbf{t}) + S^D(I, \mathbf{p}, \mathbf{t}) \tag{1}$$

Part Appearance Score. The appearance score is the summation of part filter response and compatibility biases.

$$S_A(I, \mathbf{p}, \mathbf{t}) = \sum_{i \in V} S_i^A(I, p_i, t_i) = \sum_{i \in V} \left[\alpha_i^{t_i} \cdot \phi(I, p_i) + \beta_i^{t_i}\right] \tag{2}$$

where $\alpha_i^{t_i}$ is the part filter parameters, $\beta_i^{t_i}$ is the bias term for each mixture type and occlusion state and $\phi(I, p_i)$ is the part appearance, such as Histogram of Gradients(HOG) [16] feature in this paper.

Deformation Score. The deformation score is as follows:

$$\begin{aligned} S_D(I, \mathbf{p}, \mathbf{t}) &= \sum_{(i,j) \in E} S_{ij}^D(I, p_i, p_j, t_i) \\ &= \sum_{(i,j) \in E} \left[\gamma_{ij}^{t_i} \cdot \psi(p_i - p_j) + \delta_{ij}^{t_i t_j}\right] \end{aligned} \tag{3}$$

where $\gamma_{ij}^{t_i}$ is the deformation parameters for each pair of connected parts. The deformation $\psi(p_i - p_j) = \left[dx \; dx^2 \; dy \; dy^2\right]^T$, where $dx = x_i - x_j$ and $dy = y_i - y_j$ are the relative location of child part i with respect to its parent part j. $\delta_{ij}^{t_i t_j}$ is the deformation bias.

2.2. Context Aware Model

Fig. 2 shows the structures of the proposed context aware models for upper-body and full-body pose estimation respectively. Compared with the FMP model, the proposed model is characterized by the following three aspects.

Enhanced Edges. One shortcoming of the FMP model is the lack of constraints among non-adjacent parts. In the FMP model [4], as the pairwise geometric constraints merely exist within physically adjacent local parts, the part appearance only embodies local geometric constraint between child node and its parent node. Thus the non-connected symmetric parts turn to explain the same region and the model prunes to double counting when there is large deformation or occlusion (as shown in the top row of Fig. 1). Our context aware model incorporates both local constraints and long range contextual information. The deformation score can be formulated as

$$S_D(I, \mathbf{p}, \mathbf{t}) = \sum_{(i,j) \in E} S_{ij}^D(I, p_i, p_j, t_i) + \sum_{(k,l) \in E'} S_{kl}^D(I, p_k, p_l, t_k) \tag{4}$$

where E' denote the enhanced edges (colored in green) depicted in Fig. 2. The long range constraints make our model more robust to large deformation and occlusion (as shown in the bottom row of Fig. 1).

Virtual Parts. In the FMP model, the relative position of leaf parts to the parent are much more diverse than those of the higher-level parts in the kinetic tree. This will decrease the localization accuracy of leaf nodes dramatically. To overcome this issue, we emphasize the loss of localization error of lower-level parts with virtual parts V' (colored in orange in Fig. 2) in our model and the part appearance score becomes as follows

$$S_A(I, \mathbf{p}, \mathbf{t}) = \sum_{i \in V} S_i^A(I, p_i, t_i) + \sum_{i \in V'} S_k^A(I, p_k, t_k) \tag{5}$$

To achieve efficient inference, we make the assumption that each virtual part and its corresponding real part are spacially independent. Thus the score of virtual parts can pass along the enhanced edges and the entire structure of our model is still in the form of tree.

Phraselet Clustering As shown in Fig. 3, the FMP clusters local parts only according to the relative position from their parents, the right elbows with different global upper pose configurations(left: frontal view with separated arms v.s. right: side view with overlapping arms). This will confuse the two different kinds of pose configurations and encourage double counting. To alleviate this challenge, we propose to cluster local parts by modeling both local and long range interactions between parts. Specifically, the local parts are clustered according to the relative position from all the parents and children (including virtual parts) of the part. The proposed method is similar to the relational phraselets by Desai and Ramanan [17], and we call our method as phraselet clustering. However, the difference is that they model the interaction between people and objects while we use visual phraselets to model the interactions between different body parts.

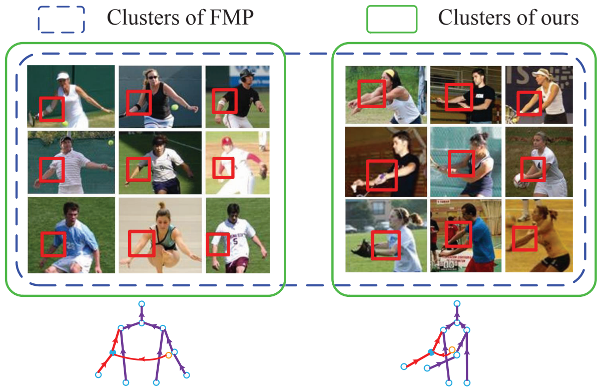

Fig. 3. Phraselet clustering for the right elbow. Right elbows with different pose configuration will be clustered together by the FMP [4].

2.3. Inference and Learning

Our goal is to maximize the score of the enhanced tree model as follows:

$$(\mathbf{p}^*, \mathbf{t}^*) = \arg\max \Bigg[\sum_{i \in V} S_i^A(I, p_i, t_i) + \sum_{(i,j) \in E} S_{ij}^D(I, p_i, p_j, t_i) + \sum_{k \in V'} S_k^A(I, p_k, t_k) + \sum_{(k,l) \in E'} S_{kl}^D(I, p_k, p_l, t_k) \Bigg] \tag{6}$$

As mentioned above, our model is still tree structured with virtual parts and enhanced edges. We can take the advantage of dynamic programming by passing messages from leaf nodes to the root nodes for efficient inference.

Given training data with labeled positive examples, i.e. images containing people with annotated part locations and learned mixture types to be $\{(I_n, \mathbf{p}_n, \mathbf{t}_n) \mid n \in pos\}$. Denote the model parameter as $\mathbf{w} = [\alpha_i^{t_i}, \ldots, \beta_i^{t_i} \ldots, \gamma_{ij}^{t_i} \ldots, \delta_{ij}^{t_i t_j}]^T$, which is a concatenation of HOG filters $\alpha_i^{t_i}$, part appearance bias $\beta_i^{t_i}$, deformation parameters $\gamma_{ij}^{t_i}$ and deformation bias $\delta_{ij}^{t_i t_j}$. Let $\Phi(I_n, \mathbf{p}_n, \mathbf{t}_n)$ be the concatenation of all the features with the same order. The parameter $\mathbf{w}$ can be learned with structural SVM [18] as follows:

$$\begin{aligned} &\underset{\mathbf{w},\, \xi_n \geq 0}{\text{argmin}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_n \xi_n \\ &s.t. \quad \mathbf{w} \cdot \Phi(I_n, \mathbf{p}_n, \mathbf{t}_n) \geq 1 - \xi_n \quad \forall n \in pos, \\ &\qquad\;\; \mathbf{w} \cdot \Phi(I_n, \mathbf{p}, \mathbf{t}) \leq -1 + \xi_n \quad \forall n \in neg, \forall(\mathbf{p}, \mathbf{t}). \end{aligned}$$

We optimize this objective function using dual coordinate descent [19]. The formulation above forces the exponential number of negative examples to be scored lower than -1. We follow the advice of [4] to search for hard negative examples from images without person.

3. EXPERIMENTS AND COMPARASION

The proposed approach is evaluated on three public datasets: Buffy [20], LSP [21] and FLIC [22], which are popular benchmark datasets. The training procedure and setting are the same as [4] which is chosen as our baseline method, e.g., the non-person images of INRIA Person dataset [16] is used for hard negative example mining.

3.1. Evaluation Criteria

The most popular criterion for human pose estimation is the Percentage of Correct Parts (PCP) measure, where estimated part end points must be within half of the part length from the ground truth part end points [20]. We adopt the more strict criterion "PCP-strict": single output and "both end points to be correct" as described in [23] for the LSP dataset. As most of the previous results on the Buffy dataset are evaluated with the original PCP-average criterion [20], we also adopt the same criterion.

Though PCP was the most widely used metric for evaluation, it has the drawback of penalizing short limbs, such as lower arms, which are usually more difficult to detect. An alternative is the Percentage of Corrected Keypoints (PCK) measure [23], which is adopted by most of the literatures. We use the PCK criterion to precisely evaluate the localization accuracy of body joints on the FLIC dataset.

3.2. Evaluation Results

LSP dataset The LSP dataset contains sport images with various pose and we use the observer-centric annotations as suggested in [24]. Table 1 compares our approach with some state-of-the-art methods. Some of the average PCP outperform us because they utilize deep structures [25, 26], and some others use stronger feature and prior [27]. However, our method is much better at localizing lower limbs such as forearms. Fig. 4 shows some detection results compared with that of [4].

Table 1. Test results on LSP dataset.

Method	Head	Torso	U.Leg	L.Leg	U.Arm	L.Arm	Avg
Ours	78.8	86.2	74.5	71.1	62.1	**46.9**	67.4
Ramakrishna et al. [26]	84.3	88.1	**79.0**	**73.6**	62.8	39.5	67.8
Pishchulin et al. [27]	**85.6**	**88.7**	78.8	73.4	61.5	44.9	**69.2**
Ouyang et al. [25]	83.1	85.8	76.5	72.2	**63.3**	46.6	68.6
Eichner et al. [24]	80.1	86.2	74.3	69.3	56.5	37.4	64.3
Pishchulin et al. [8]	78.1	87.5	75.7	68.0	54.2	33.9	62.9
Yang & Ramanan [4]	77.1	84.1	69.5	65.6	52.5	35.9	60.8
Yang & Ramanan [23]	79.3	82.9	70.3	67.0	56.0	39.8	62.8
Andriluka et al. [3]	74.9	80.9	67.1	60.7	46.5	26.4	55.7

Buffy dataset Table 2 shows quantitative results of PCP-average. Our model without phraselet clustering is on par with that of Yang & Ramanan [4]. And the model with phraselet clustering is 2.5% better in overall PCP-average.

Table 2. PCP-average on Buffy dataset.

Method	Head	Torso	U.Arm	L.Arm	Avg
Ours with PC[a]	**100**	**100**	**97.2**	**73.0**	**90.1**
Ours without PC	100	99.3	96.7	66.5	87.4
Yang & Ramanan [4]	99.6	98.9	95.1	68.5	87.6
Sapp et al. [28]	81.9	85.1	77.6	53.6	72.8
Eichner et al. [29]	83.4	84.0	70.5	50.9	68.2
Andriluka et al. [3]	81.3	77.2	67.5	35.1	62.6

[a] PC = Phraselet Clustering

From Fig. 5 we observed that the position of virtual parts is closer to its corresponding real part when the model is trained

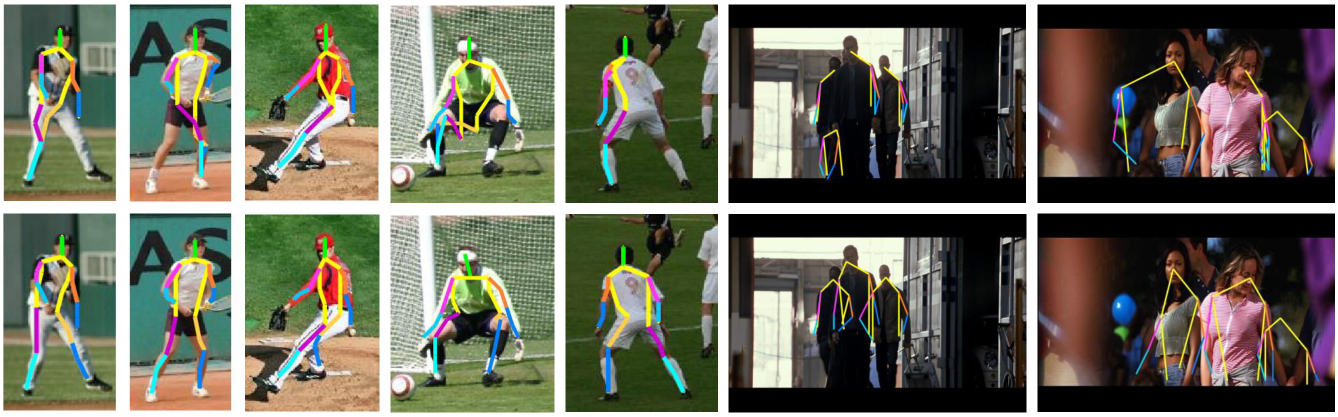

Fig. 4. Comparison of detection results. The first row for FMP [4] and the second row for ours. The full-body images are from the LSP dataset and the upper-body images are from the FLIC dataset.

with phraselet clustering. It reflects that phraselet clustering is complementary to the spacially independent assumption betwee

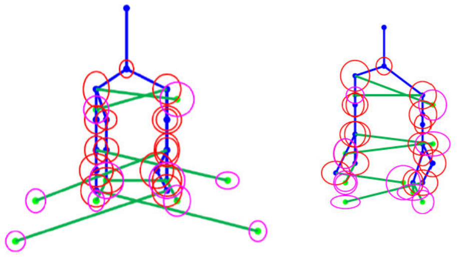

Fig. 5. The visualization of skeletons of trained models on Buffy dataset without (left) and with (right) Phraselet Clustering. The blue edges and green edges represent kinetic constraints and the enhanced edges respectively. The red and magenta ellipses show the variance of each child part relative to its parent. We only show one pose configuration for clearance by placing parts at their best-scoring location relative to their parent.

FLIC dataset The FLIC dataset contains images of real life scenes and is challenging in the localization of elbows and wrists. We compare with several state-of-the-art models whose codes are available. The result of MODEC [22] is derived from the model trained by the authors. The model of FMP [4] is retrained on FLIC training set. Since the training code of Eichner et al. [29] is not available, we use the provided model for test.

As shown in Fig 6, our method outperforms MODEC [22] by 9.8% and 9.0% in AUC[1] respectively on elbows and wrists. It reflects that the modeling of long range interactions between physically unconnected parts(e.g., left and right wrists) is beneficial for the localization of lower arms.

4. CONCLUSION

In this paper, we propose a novel context aware model which incorporates both local kinetic constraints and global contextual constraints of non-adjacent parts. The model can maintain in the form of tree structure by introducing virtual parts, thus dynamic programming can be utilized for efficient inference. Experiments on public benchmarks show that effectiveness of our method in tackling double counting problem and improving the localization accuracy of challenging lower limbs.

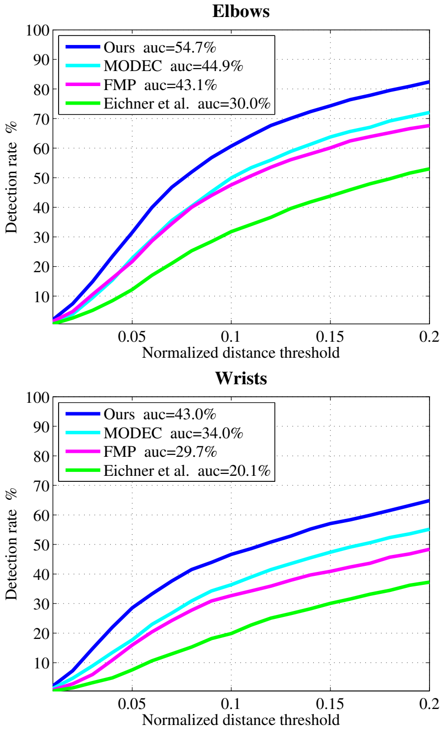

Fig. 6. PCK on FLIC dataset for the most challenging parts: elbows and wrists.

5. ACKNOWLEDGMENT

This work is funded by the National Basic Research Program of China (Grant No. 2012CB316302), National Natural Science Foundation of China (Grant No. 61322209 and Grant No. 61175007).

[1] Here AUC means the average detection rate for normalized distance threshold to be within $0 \sim 0.2$.

6. REFERENCES

[1] Martin A. Fischler and Robert A. Elschlager, "The representation and matching of pictorial structures," *IEEE Transactions on Computers*, vol. 100, no. 1, pp. 67–92, 1973.

[2] Pedro F. Felzenszwalb and Daniel P. Huttenlocher, "Pictorial structures for object recognition," *IJCV*, vol. 61, no. 1, pp. 55–79, 2005.

[3] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," in *CVPR*, 2009, pp. 1014–1021.

[4] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *CVPR*, 2011, pp. 1385–1392.

[5] Min Sun and Silvio Savarese, "Articulated part-based model for joint object detection and pose estimation," in *ICCV*, 2011, pp. 723–730.

[6] Yuandong Tian, C. Lawrence Zitnick, and Srinivasa G. Narasimhan, "Exploring the spatial hierarchy of mixture models for human pose estimation," in *ECCV*, 2012, pp. 256–269.

[7] Fang Wang and Yi Li, "Beyond physical connections: Tree models in human pose estimation," in *CVPR*, 2013, pp. 596–603.

[8] Leonid Pishchulin, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele, "Poselet conditioned pictorial structures," in *CVPR*, 2013, pp. 588–595.

[9] Yang Wang and Greg Mori, "Multiple tree models for occlusion and spatial constraints in human pose estimation," in *ECCV*, 2008, pp. 710–724.

[10] Tai-Peng Tian and Stan Sclaroff, "Fast globally optimal 2d human detection with loopy graph models," in *CVPR*, 2010, pp. 81–88.

[11] Duan Tran and David A. Forsyth, "Improved human parsing with a full relational model," in *ECCV (4)*, 2010, pp. 227–240.

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

[13] Y. Wang, D. Tran, and Z.C. Liao, "Learning hierarchical poselets for human parsing," in *CVPR*, 2011, pp. 1705–1712.

[14] Min Sun, Murali Telaprolu, Honglak Lee, and Silvio Savarese, "An efficient branch-and-bound algorithm for optimal human pose estimation," in *CVPR*, 2012, pp. 1616–1623.

[15] Yi Xiao, Huchuan Lu, and Shifeng Li, "Posterior constraints for double-counting problem in clustered pose estimation," in *ICIP*, 2012, pp. 5–8.

[16] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005, pp. 886–893.

[17] Chaitanya Desai and Deva Ramanan, "Detecting actions, poses, and objects with relational phraselets," in *ECCV*, 2012, pp. 158–172.

[18] Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun, "Large margin methods for structured and interdependent output variables," *JMLR*, vol. 6, pp. 1453–1484, 2005.

[19] Deva Ramanan, "Dual coordinate solvers for large-scale structural svms," *CoRR*, vol. abs/1312.1743, 2013.

[20] Vittorio Ferrari, Manuel J. Marłn-Jimnez, and Andrew Zisserman, "Progressive search space reduction for human pose estimation," in *CVPR*, 2008.

[21] Sam Johnson and Mark Everingham, "Clustered pose and nonlinear appearance models for human pose estimation," in *BMVC*, 2010, pp. 12.1–12.11.

[22] Ben Sapp and Ben Taskar, "Modec: Multimodal decomposable models for human pose estimation," in *CVPR*, 2013, pp. 3674–3681.

[23] Yi Yang and Deva Ramanan, "Articulated human detection with flexible mixtures of parts," *PAMI*, vol. 35, no. 12, pp. 2878–2890, 2013.

[24] Marcin Eichner and Vittorio Ferrari, "Appearance sharing for collective human pose estimation," in *ACCV*, 2012, pp. 138–151.

[25] Wanli Ouyang, Xiao Chu, and Xiaogang Wang, "Multi-source deep learning for human pose estimation," in *CVPR*, 2014, pp. 2337–2443.

[26] Varun Ramakrishna, Daniel Munoz, Martial Hebert, J.A. Bagnell, and Yaser Sheikh, "Posemachines: Articulated pose estimation via inference machines," in *ECCV*, 2014, pp. 33–47.

[27] Leonid Pishchulin, Mykhaylo Andriluka, Peter Gehler, and Schiele Bernt, "Strong appearance and expressive spatial models for human pose estimation," in *ICCV*, 2013, pp. 3487–3494.

[28] Benjamin Sapp, Alexander Toshev, and Ben Taskar, "Cascaded models for articulated pose estimation," in *ECCV*, 2010, pp. 406–420.

[29] Marcin Eichner and Vittorio Ferrari, "Better appearance models for pictorial structures.," in *BMVC*, 2009.