Beyond Tree Structure Models: A New Occlusion Aware Graphical Model for Human Pose Estimation

Lianrui Fu[1,2,4], Junge Zhang[1,2,4] and Kaiqi Huang[1,2,3,4]

[1]Center for Research on Intelligent Perception and Computing
[2]National Laboratory of Pattern Recognition
[3]CAS Center for Excellence in Brain Science and Intelligence Technology
[4]Institute of Automation, Chinese Academy of Sciences

{lianrui.fu, jgzhang, kaiqi.huang}@nlpr.ia.ac.cn

Abstract

Occlusion is a main challenge for human pose estimation, which is largely ignored in popular tree structure models. The tree structure model is simple and convenient for exact inference, but short in modeling the occlusion coherence especially in the case of self-occlusion. We propose an occlusion aware graphical model which is able to model both self-occlusion and occlusion by the other objects simultaneously. The proposed model structure can encodes the interactions between human body parts and objects, and hence enables it to learn occlusion coherence from data discriminatively.

We evaluate our model on several public benchmarks for human pose estimation including challenging subsets featuring significant occlusion. The experimental results show that our method obtains comparable accuracy with the state-of-the-arts, and is robust to occlusion for 2D human pose estimation.

1. Introduction

Human pose estimation from still image is a challenging problem in computer vision. It is key to many visual tasks, e.g., action recognition, clothes parsing and human-computer interaction. This problem is still challenging due to large deformation, illumination, camera viewpoint, cluttered background and occlusion.

Recent progress on human pose estimation is ascribed to the pictorial structured model especially simple tree structure [53, 40, 44, 48]. Although these methods perform well on images with rare occlusion, they may fail when the body parts are occluded by some other body parts(self-occlusion) or the other objects(other-occlusion). Fig. 1(c) depicts that the famous flexible mixtures-of-parts model(FMP) [53] fails under occlusion. The tree structured model is simple, yet fails to model the interaction between unconnected body parts, and the interactions between human body and objects. However, these interactions are important cues for occlusion reasoning. The question is, how can we model such interactions for occlusion reasoning?

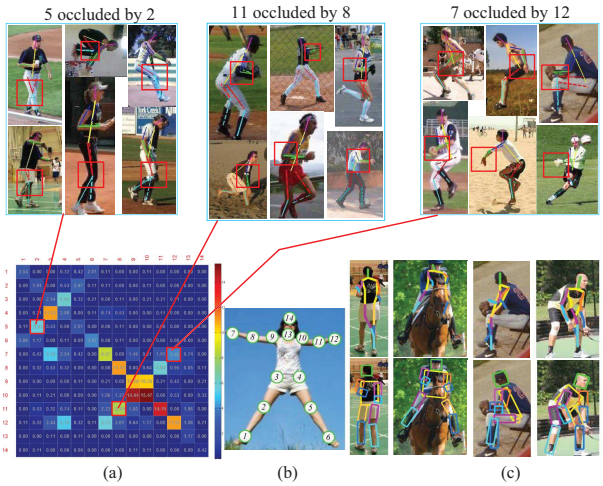

Figure 1. Occlusions in Leeds Sports dataset. (a) Visualization of the occlusion relation matrix. The color of diagonal squares reflects the probability of other-occlusion for each joint. Hotter color means heavier occlusion. The hotness of each nondiagonal square in row i and column j indicates the probability of joint i being occluded by joint j. The image groups on the top are the instances with the same self-occlusion relationship. (b) The sequence number of body joints in Leeds Sports dataset. (c) Human pose estimation results with occlusions from FMP [53](the first row) and ours(second row).

There are mainly two types of occlusion for human pose estimation: other-occlusion (occluded by objects) and self-occlusion (occluded by some other body parts). Other-occlusion appears when some objects block the view and this will damage the local appearance of body parts and cause failure detection. Take the images of the second column of Fig 1(c) for example, the FMP fails because the left knee and hip of the rider are occluded by the horse. However, only a few body parts are frequently occluded such as lower arms and legs while head is mostly visible(see the

diagonal elements in the occlusion relation matrix).

In contrast to other-occlusion, self-occlusion appears when the body parts occlude each other due to viewpoint or pose deformation. In this case, the same region in 2D image has to be explained as different body parts. This is less likely to damage local appearance of the occludee, yet will cause the ambiguity of pose configuration as there is no interaction between the occluder and the occludee. For example, as seen from the third and fourth column of Fig 1(c), FMP fails to capture the correct pose configuration under self-occlusion.

Statistics on the Leeds Sport dataset (LSP) [24] show that 47.2% of the images have one more body joints invisible and 16.7% have more than three body joints occluded. Among all these invisible joints, 67.4% are self-occluded while the rest are other-occluded. Existing works mainly focus on handling other-occlusion, self-occlusion is often ignored or treated in the same manner as other-occlusion. We argue that the occluder in self-occlusion can not be treated as noise as that in other-occlusion. How can we model both kinds of occlusion in an unified framework simultaneously?

Motivated by these above, we propose a novel occlusion aware graphical model which explicitly model both self-occlusion and other-occlusion to improve the robustness to occlusion. We evaluate our model on several public benchmarks for human pose estimation and test on the challenging subsets with significant occlusion. The results verify the proposed method's effectiveness to address occlusion problem and it has obtained comparable accuracy to previous state-of-the-art methods on public datasets. In particular, our method performs much better than previous methods on those datasets with heavy occlusion.

2. Related Work

Recent approaches on human pose estimation mainly focus on richer model structure, stronger feature representation and specific challenges such as occlusion.

The most popular modern approaches for human pose estimation is based on the pictorial structured model(PSM) [16]. In the PSM model, human body configuration is represented as a collection of independent parts with pairwise connections. The pairwise part relationships are embodied in tree models [2, 53, 40, 44, 48, 30], multi-tree model [50] or loopy models [43, 47, 38, 51, 41]. Tree models prevail for their simplicity and exact inference. However, they are insufficient in capturing high-order spacial relationships among body parts and the message passing tends to break down when occlusion occurs. Loopy models allow more complex relationships among parts, but require approximate inference iteratively. Our occlusion aware graphical model is able to model such interactions among parts with efficient approximate inference.

In addition to model structure, some adopt strong feature and middle level representation. For instance, Convolutional Neural Networks (CNNs) [27] are used to extract more powerful features [46, 21] and Poselets [4], Deformable Part based Model (DPM) [13] are adopted to generate richer middle level representations with strong pose priors [30, 29]. Some incorporates CNN part detectors and graphical models with either piecewise training [7] or joint training [45]. In contrast to modeling pairwise constraints, some [9, 32] adopt layered random forest to incorporate rich spatial interactions among multiple parts. However, there is no explicit modeling of occlusion in these approaches.

In terms of handling occlusion of pose estimation, body part visibility is usually modeled as binary variable in either part level or image level. Some previous object detection approaches [49, 17] model occlusion with segmentation of image feature map. Part level occlusion reasoning is frequently used to model more complicated occlusions. For instance, the supervised part models [3] includes visibility variable for each part but imposes no constraints on the visibility of different parts in the model. Similarly, Hejrati et al. [22] extend the flexible mixtures-of-part model [53] with part level occlusion reasoning for 3D car alignment. Desai et al. [10] model the interactions between human and objects which can capture the occlusion relationships. Wang et al. [50] propose to combine multiple tree framework for occlusion reasoning. The And-Or graph model [36] also incorporates visibility into the part node. The grammar-based model [20] in people detection includes explicit occlusion part templates but enforces more structure in the pattern of occlusion. The strongly supervised deformable model [19], by contrast, tries to sidestep the structure learning problem and automatically learn valid occlusion patterns from data in a non-parametric way. The very recent flexible compositions [6] model visible parts with subtrees and learn occlusion cues with CNNs.

Most of the work above mainly focus on other-occlusion while self-occlusion is often ignored or treated the same manner as other-occlusion(as noise). There are only a few works trying to model self-occlusion. Sigal et al. [39] propose to use pixel level hidden binary variables for self-occlusion reasoning. Some others try to model self-occlusion in a holistic manner. Yang et al. [55] model self-occlusion of pedestrian in a joint shape and appearance tracking framework. Radwan et al. [31] treat self-occlusion reasoning as post process with Twin-GP regression for 2D pose rectification. However, our model learns the part-level occlusion relationships from data and infers the occlusion states of parts explicitly. Our model is more flexible and can encode more complex interactions between parts.

3. Occlusion Aware Graphical Model

In this section, we will first introduce the proposed occlusion aware graphical model, and then describe the inference and learning procedure of our model.

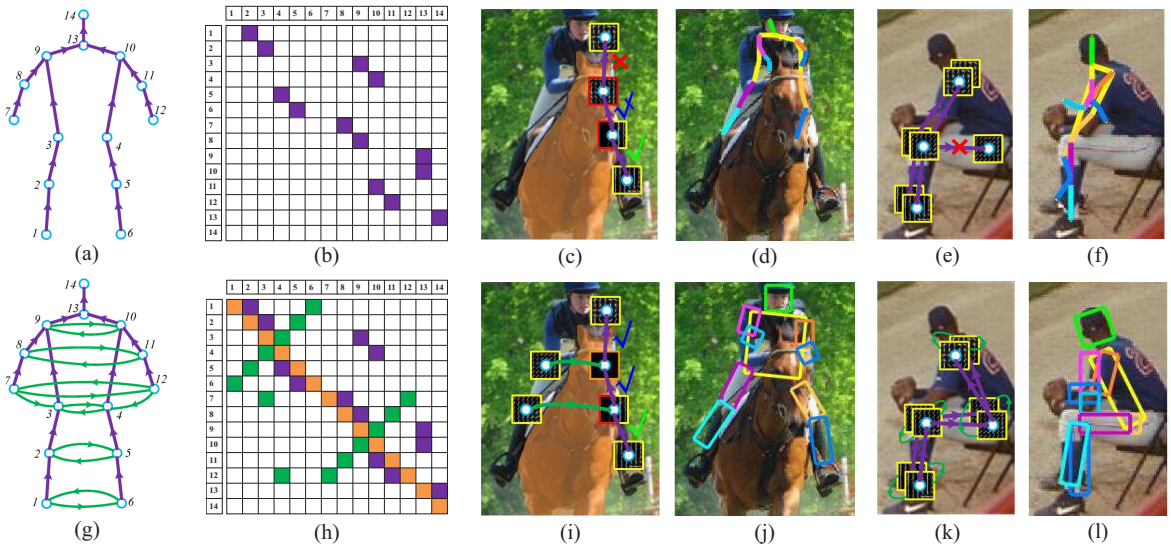

Figure 2. Comparison of the proposed model structures with respect to that of FMP [53]. (a) Model structure of FMP, which is a kinetic tree. (b) The kinetic constrains represented in adjacent matrix of body parts. (c) The message passing of FMP breaks down under other-occlusion. (d) Failure detection of FMP under other-occlusion. (e) The message passing of FMP breaks down under self-occlusion. (f) Failure detection of FMP under self-occlusion. (g) Model structure of the proposed method. (h) The constraints and occlusion representation in adjacent matrix of body parts. (i) The message passing of the proposed method under other-occlusion. (j) Result of the proposed method under other-occlusion. (k) The message passing of the proposed model under self-occlusion. (l) Result of the proposed method under self-occlusion. The skeletons in (a) and (g) both are in a front view. The charts and pictures are best viewed in color.

3.1. Model Structure

Let $p_i = (x_i, y_i)$ be the pixel location for part i, $t_i \in 1, \ldots, T$ be the local mixture component of part i, $o_i \in \{0, 1, 2\}$ be the occlusion state("0" for visible, "1" for self-occlusion and "2" for occlusion by the other objects) and $\mathcal{J} = \{\boldsymbol{j}_i\}$ be the pose configuration where $\boldsymbol{j}_i = (p_i, o_i, t_i)$. Given an input image I, the posterior of a pose configuration of parts is

$$P(\mathcal{J}|I) \propto \exp\left[\sum_{i \in V} s_{ao}(I|\boldsymbol{j}_i) + \sum_{k,l \in E} s_{do}(\boldsymbol{j}_k, \boldsymbol{j}_l)\right] \quad (1)$$

The unary term $s_{ao}(I|\boldsymbol{j}_i)$ models the appearance of each part i, V is the set of part nodes and E is the set of edges in the model. The appearance varies with view point change, articulation as well as occlusion. To model these variations, $\boldsymbol{j}_i = (p_i, o_i, t_i)$ specifies the part appearance with respect to part localization p_i, part occlusion state o_i, and part mixture type t_i which encodes the rotation and size of part i.

The pairwise term $s_{do}(\boldsymbol{j}_k, \boldsymbol{j}_l)$ models the geometric deformation constraints as well as occlusion relations between body parts k and l on an occlusion-aware graph G, e.g., the left knee is probably occluded by the right knee while the left and right arms are less likely to occlude each other in Fig. 1(a). However, it is hard to model such subtle relations in a tree structured model such as in [53, 40, 48].

Fig 2 compares the structure of our occlusion aware graphical model ((g) and (h)) and that of FMP [53] ((a) and (b)). The proposed model differs from FMP in two aspects: first, each part of the model contains occlusion states which indicate whether the part is visible, other-occluded (squares in the diagonal elements in the adjacent matrix of body parts, colored in orange) or self-occluded; second, in contrast to merely considering the kinetic constrains (purple edges/squares in (a), (b), (g) and (h)) between nearby parts, our model encodes richer interactions between parts (green edges/squares in (g) and (h)) that are closely related to self-occlusion. We call these green edges enhanced edges with respect to the purple edges representing kinetic constraints.

The goal of our occlusion aware graphical model is to maximize the posterior as follows:

$$P(\mathcal{J}|I) \propto \exp\left(S(I, p, o, t)\right) \quad (2)$$

This is equivalent to maximize the score of pose configuration score $S(I, p, o, t)$, which is composed of part appearance score and deformation score.

$$S(I, p, o, t) = S_{ao}(I, p, o, t) + S_{do}(I, p, o, t) \quad (3)$$

Part appearance score: The part appearance score is a summation of part filter response and compatibility biases.

$$S_{ao}(I, p, o, t) = \sum_{i \in V} \left[\alpha_i^{t_i} \cdot \phi(I, p_i, o_i) + \beta_i^{t_i}(o_i)\right] \quad (4)$$

where $\alpha_i^{t_i}$ is the part filter parameters and $\beta_i^{t_i}(o_i)$ is the bias term for each mixture type and occlusion state. The part appearance $\phi(I, p_i, o_i)$ is defined as

$$\phi(I, p_i, o_i) = \begin{cases} \phi(I, p_i), & if\ o_i = 0, 1 \\ 0, & if\ o_i = 2 \end{cases}$$

This indicates that we set the part score to be zero only when it is occluded by some other objects. This differs from those approaches that treat both self-occlusion and other occlusion as noise and prune the local part score. In our method, the pattern of self-occlusion can be captured for further inference even when the body part is invisible (occluded by some other body part).

Deformation score: The deformation score is as follows:

$$S_{do}(I,p,o,t) = \sum_{(i,j)\in E} \left[\gamma_{ij}^{t_i t_j} \cdot \psi(p_i - p_j) + \delta_{ij}^{t_i t_j}(o_i, o_j)\right] \quad (5)$$

where $\gamma_{ij}^{t_i t_j}$ is the deformation parameters for each pair of connected parts. The part deformation $\psi(p_i - p_j) = \begin{bmatrix} dx & dx^2 & dy & dy^2 \end{bmatrix}^T$, where $dx = x_i - x_j$ and $dy = y_i - y_j$, the relative location of part i with respect to j. $\delta_{ij}^{t_i t_j}$ encodes the occlusion coherence between body parts.

Note that the edges in our model not only contain kinetic constraints between nearby parts but also incorporate interactions between parts which can help reasoning occlusion relationships. As shown in Fig. 2(i), when the left hip of the rider is occluded by the head of the horse, the score of visible parts(on the right leg of the rider) can be passed through green edges (see Fig. 2(g) and (j)). Similarly, when the body parts are occluded by the other parts of the person in Fig. 2(k), the occluder and the occludee can pass the occlusion relationship to each other, so the occluder-occludee part pair can explain the same region without mutual exclusion. However, the FMP model can not handle these issues and often fails under other-occlusion and self-occlusion (see Fig. 2(d) and (f)).

In the following subsection, we will introduce how the subtle information is passed to the corresponding parts and benefits the inference of occluded parts.

3.2. Model Inference

As described above, the structure of our model is a graph which contains loops. Inference on general loopy graphs is a NP-hard problem. Many approximate methods, such as Loopy Belief Prorogation [52], Branch and Bound [43] and Dual Decomposition [26], need to iteratively infer on tractable structures many times until converge. However, our model contains large number of parameters and needs to mine huge amount of negative examples. Alternatively Ramanan [34] propose to use tree-model for generating candidate pose configurations and scoring the configurations using more complex non-tree constraints. Inspired by this, we first unroll the graphical model into a tree model to generate candidate pose hypothesis, and then rescore the candidate pose configurations with graphical model.

Model unrolling For any part $\boldsymbol{j}_i$ with out-degree (number of connections pointing to the other parts) $\nu_i > 1$, we generate $\nu_i - 1$ virtual parts and unroll the enhanced edges to form a computation tree similar to [42](See Fig 3(b)). As the parent of each virtual part is real part in our model, unrolling for our model is equivalent to the effect of single iteration of loopy belief propagation at the root node. The unrolled tree model is then used for generating and selecting candidate pose hypotheses.

Pose selection The goal of our graphical model is to

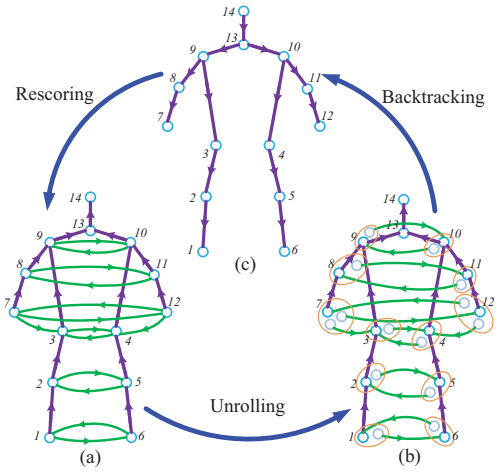

Figure 3. Inference of the graphical model. (a)The graphical model, (b)The unrolled computation tree for approximate message passing, (c)The nodes backtracked via the tree structure

maximize the posterior $P(\mathcal{J}|I)$ in Eq. (2), i.e.,

$$\mathcal{J}_m = \underset{\mathcal{J}}{\operatorname{argmax}} \sum_{i\in V} s_{ao}(I|\boldsymbol{j}_i) + \sum_{k,l\in E} s_{do}(\boldsymbol{j}_k, \boldsymbol{j}_l) \quad (6)$$

Instead of passing message on a loopy graph, we pass message on the unrolled tree structure to generate root hypothesis. This allows us to employ dynamic programming to pass message from leaf nodes to the root node efficiently. The optimization over the unrolled model can be formulated as:

$$\mathcal{J}'_m = \underset{\mathcal{J}}{\operatorname{argmax}} \sum_{i\in V} s_{ao}(I|\boldsymbol{j}_i) + \sum_{i\in V'} s_{ao}(I|\boldsymbol{j}_i) + \sum_{k,l\in E} s_{do}(\boldsymbol{j}_k, \boldsymbol{j}_l) \quad (7)$$

This equals to adding part appearance weights to the nodes with more connections.

Suppose the number of possible root hypotheses to be L in the test image. We sort them by the score and choose top L_σ hypotheses with the highest score. $\sigma = L_\sigma / L$ is the ratio of hypotheses selection. We assume that the optimal hypothesis is included in the selected top-σ hypotheses of the unrolled configurations.

Backtracking and rescoring As soon as the top-σ hypotheses of root node are determined, the optimal configuration can be obtained by backtracking directly from the root node to the leaf nodes. We only backtrack the child node from actual parent node (e.g., node 1 is backtracked from node 2 rather than node 6 in Fig 3) as the parent near the root node is more reliable. We will recompute the score of the pose configurations with graphical model and rerank the hypothesis.

Experimental results in the later sections will show that the performance almost does not change when the ratio $\sigma > 0.01$. We set $\sigma = 0.01$ for all the evaluations.

Computation Let L be the number of possible part locations, T be the number of mixture types, K be the number of real parts and K_v be the number of virtual parts. The complexity of message passing is $O((K-1)LT^2)$ with dynamic programming and distance transform [14] for the tree structured model of FMP [53]. For our model, the complexity becomes $O((K + K_v - 1)L(3T)^2) =$

$O(9(K + K_v - 1)LT^2)$, which is slower than the FMP. However, the backtracking and rescoring procedure can be very fast as we only process the selected L_σ hypotheses. The average detection speed of our method is 4.5 seconds per image for the LSP dataset on single 3.4GHz CPU compared with 1.2 seconds per image of the FMP.

3.3. Model Learning

Learning local mixtures

In the learning of mixture types of local parts, there are several approaches. In the latent tree model [48], local part mixtures are learned by clustering part appearance yet without considering the structure of human body. In contrast, the FMP model [53] learns part mixtures by clustering the relative position from the parent for each part in the kinetic tree. This is because each part is only constraint to its parent in the articulated tree structure. It is reasonable for higher level parts such as head and torso. However, this makes it hard to capture the varying occlusion relationships between non-adjacent parts (especially for lower limbs). In our model, many parts have multiple interactions with some other parts. We learn the part mixture types from the relative position from all the parents and children in the graphical model. In this way, we can not only capture the local part deformation but also encode the global pose deformation. This will benefit the localization of occluded parts and the lower level parts which contain much uncertainty of freedom in a tree structured model. We use the simple k-means clustering with multiple runs and choose the one with minimum objective function.

Learning occlusion coherence To model spatial coherence among part occlusions, we utilize two sources of occlusion samples. One is from the label of part occlusion states and the other from synthetic occlusion patterns. For the labeled invisible part, we distinguish it as self-occluded if there is some other visible body part with more than 50% overlapped with it. The self-occlusion relationships will be captured by the enhanced edges in our model. We find that more than half the invisible body parts in Leeds Sports Dataset are occluded by the other body parts due to articulation and viewpoint. And the number of instances with other-occlusion is relatively small. To balance the two different types in the training sample, we synthesize samples with occlusion by the other objects. We utilize occlusion masks to generate synthetic samples similarly as [18]. And we only use samples with seldom occlusions for synthesization. During training, the occlusion relationship between parts as well as the occlusion pattern are learned and encoded in the model.

Learning parameters Given the pose configuration $\mathcal{J} = \{\boldsymbol{j}_i\}$ and the image I, the configuration score can be computed using Eq.(3), Eq.(4) and Eq.(5). For the linear property, the total score of configuration $\mathcal{J}$ in image I can be simplified as:

$$S(I, \mathcal{J}) = w \cdot \Phi(I, \mathcal{J}) \tag{8}$$

$w = [\alpha_i^{t_i}, \cdots, \beta_i^{t_i}(o_i) \cdots, \gamma_{ij}^{t_i t_j} \cdots, \delta_{ij}^{t_i t_j}(o_i, o_j)]$.

where w is the concatenation of all the parameters including $\alpha_i^{t_i}$, $\beta_i^{t_i}(o_i)$, $\gamma_{ij}^{t_i t_j}$ and $\delta_{ij}^{t_i t_j}(o_i, o_j)$. $\Phi(I, \mathcal{J})$ is the concatenation of all the features with the same order. For the bias terms $\beta_i^{t_i}(o_i)$ and $\delta_{ij}^{t_i t_j}(o_i, o_j)$, the corresponding dimensions of $\Phi(I, \mathcal{J})$ are set to be 1. For mixture types and occlusion states which are not activated, the corresponding dimensions in $\Phi(I, \mathcal{J})$ are filled with 0.

In this way, the proposed occlusion aware graphical model can be linearly parameterized, allowing efficient training using a large margin objective. The optimization function can be written as:

$$\underset{w}{\operatorname{argmin}} \frac{1}{2} w^T w + C \sum_n \max(0, 1 - y_n \langle w, \Phi(I, \mathcal{J}) \rangle) \tag{9}$$

where $y_n \in \{1, -1\}$, $y_n = 1$ if $n \in pos$, and $y_n = -1$ if $n \in neg$. This is a standard structural SVM learning problem, which can be solved by the cutting pane solver like $\text{SVM}^{\text{struct}}$ [23] or the stochastic gradient descent(SGD) solver. In this paper, we turn to use dual coordinate descent QP solver of [35] as we should meet the requirement of parameters constraints, e.g., the coefficents of part deformation in $\gamma_{ij}^{t_i t_j}$ should be negative for generic distance distance transform [14]. The body part position, visibility and local spacial configurations are completely specified during training.

4. Experimental Evaluation

This section describes our experimental setup, presents a comparative performance evaluation on human pose estimation benchmarks and analyze the influence of parameter settings.

Datasets For comprehensive evaluation on public benchmarks, we firstly evaluate the proposed approach on the popular LSP [24] dataset, and then we test it on the PARSE [33] dataset with the model trained on LSP dataset for generalization ability, finally we evaluate our method on the FLIC dataset [37] with 11 points Upper body annotations from popular Hollywood movies. As this paper intends to address the problem of human pose estimation with occlusion, we specifically design an experiment on occluded images for better explaining our approach. We choose subset images with occlusions from LSP and the new challenging MPII [1] for detailed analysis of the robustness to occlusion. Tab. 1 lists the dataset used for evaluation in our work.

Dataset	#train	#test	#points	POJ[1]	scene	Pose variation
LSP [24]	1000	1000	14	16.7%	sports	large
PARSE [33]	100	205	14	–	diverse	most upright
FLIC [37]	3987	1016	11	–	feature film	frontal
LSP [24]-sub	1000	468	14	20.7%	sports	large
MPII [1]-sub	1500	698	16	44.1%	diverse	large

[1] POJ = Percentage of Occluded Joints.

Table 1. Datasets used in our experiments.

	Method	Head	Torso	Leg		Arm		Avg Limbs	Avg All
				Upper	Lower	Upper	Lower		
Person-Centric	Our approach	**87.5**	91.5	74.2	66.8	**62.5**	**41.3**	**61.2**	**66.9**
	Toshev et al. [46]	–	–	**77**	**71**	56	38	61	–
	Wang et al. [48]	86.0	91.9	74.0	69.8	48.9	32.2	56.2	62.8
	Johnson et al. [25]	74.6	88.1	74.5	66.5	53.7	37.5	58.0	62.7
	Tian et al. [44]	87.8	**95.8**	69.9	60.0	51.9	32.9	53.7	61.3
	Yang et al. [53]	87.4	92.6	66.4	57.7	50.0	30.4	51.1	58.9
	Dantone et al. [9]	79.2	81.6	66.5	61.0	45.1	24.7	49.3	55.5
	Johnson et al. [24]	62.9	78.1	65.8	58.8	47.4	32.9	51.2	55.1
Observer-Centric	Our approach	77.7	85.4	75.0	71.9	62.1	**48.8**	64.2	67.7
	Ramakrishna et al. [32]	84.3	88.1	**79.0**	**73.6**	62.8	39.5	63.7	67.8
	Pishchulin et al. [29]	**85.6**	**88.7**	78.8	73.4	61.5	44.9	**64.6**	**69.2**
	Ouyang et al. [28]	83.1	85.8	76.5	72.2	**63.3**	46.6	**64.6**	68.6
	Eichner et al. [12]	80.1	86.2	74.3	69.3	56.5	37.4	59.4	64.3
	Pishchulin et al. [30]	78.1	87.5	75.7	68.0	54.2	33.9	57.9	62.9
	Yang et al. [53]	77.1	84.1	69.5	65.6	52.5	35.9	55.9	60.8
	Yang et al. [54]	79.3	82.9	70.3	67.0	56.0	39.8	58.3	62.8
	Andriluka et al. [2]	74.9	80.9	67.1	60.7	46.5	26.4	50.2	55.7

Table 2. Percentage of Correct Parts (PCP) at 0.5 on LSP for our method as well as state-of-the-art approaches. All the results are from the authors's papers respectively except that the Person-Centric(PC) results and Observer-Centric(OC) results of [53] are from [48] and [12] respectively. All the PC results are evaluated with the "PCP-average" measure while all the OC results are evaluated with the "PCP-strict" measure as in most of the literature. The detailed description of "PCP-average" and "PCP-strict" measure can be found in [54].

Criteria The most widely used criterion for human pose estimation is the Percentage of Correct Parts (PCP) measure, which evaluates the localization accuracy of body parts(sticks of skeleton). Another frequently used criterion is the Percentage of Correct Keypoints(PCK) measure, which evaluates the localization accuracy of each body joint. It is recommended to refer to [15] and [54] for more details.

4.1. Implementation detail

In the experiments, we take the FMP [53] as baseline. To enable a fair comparison of our models, our implementation uses the same settings of [53]: we use the same number of parts and identical amount of mixtures for each part. The non-person images from INRIA person dataset [8] are used as negative samples. For FLIC dataset [37], there are only annotations of upper body joints yet without occlusion state. We create 2 other-occluded samples synthetically as in [18] for each image. The joints and edges in the legs are pruned and the occlusions states are limited to model other-occlusion only.

4.2. Comparison with the Other Methods

The Leeds Sports dataset Tab. 2 shows the results of our model with the state-of-the-art approaches on the LSP dataset with Person-Centric and Observer-Centric annotations respectively. Please note that Toshev et al. [46] use additional 10000 images from LSP extend dataset [25] for training. This is due to the huge number of parameters to be learned in the CNN model. In the experiments, Andriluka's approach [2], Yang and Ramanan's approach [53] and our method are trained on the 1000 training images of the LSP dataset [24]. As shown in Tab. 2, our method performs comparable to the state-of-the-art method. Especially, our approach is better in detecting legs and arms which are prone to be occluded.

In terms of Observer-Centric annotation, the approach of Pishchulin et al. [29] performs better in localizing torso and head, this is mainly because they used strong poselet detectors as prior. The method of Ouyang et al. [28] uses deep model and takes the result of [54] as input. The performance of our method are lower but close to the state-of-the-art approach of Pishchulin et al. [29] and Ouyang et al. [28]. However, there is an ambiguity between frontal person and back person for the Observer-Centric annotation. This may confuse self-occlusion relationships for our model and hence may hurt the performance of our model.

Fig. 4 shows the detection results of the proposed method compared with the baseline method of Yang and Ramanan [53] as well as the DeepPose of Toshev et al. [46]. The detection results reflects that the DeepPose model is good at capturing global configurations of human body, yet sometimes locate the body parts inaccurately. There are two possible reasons for this: one is the normalization of image size to fit into ConvNet [27] and the other is the smoothing effect of the convolution.Our method can locate the body part more accurate in fine scales and is robust to occlusion.

Cross test on Image Parse dataset In order to measure the generalization ability of the proposed model, we test our method on the PARSE dataset as shown in Tab 3. Pishchulin's approach [30] used the LSP+PARSE training set when evaluated on the PARSE dataset. Both Johnson's approach [25] and Toshev's DeepPose [46] included 10,000

Figure 4. Comparison of detection results in LSP dataset.

extra training samples when evaluated on the PARSE dataset. In the experiment, Yang et al.'s approach [53], Ouyang et al.'s approach [28] and our method are trained on the 1000 training images of the LSP dataset [24]. Compared with the approach [53], our approach improves the accuracy by 10.9% over Yang et al.'s method in PCP on average on the PARSE dataset for LPS-PARSE cross test. The result shows good generalization ability of our method.

Method	Head	Torso	U.Leg	L.Leg	U.Arm	L.Arm	Avg
Our approach	88.3	**90.7**	75.4	66.8	**71.9**	**51.2**	70.9
Toshev et al. [46]	–	–	88	**75**	71	50	–
Ouyang et al. [28]	**89.3**	78.0	**89.3**	72.0	67.8	47.8	**71.0**
Johnson et al. [24]	76.1	78.1	73.4	65.4	64.7	46.9	66.2
Wang et al. [48]	78.7	88.3	75.2	71.8	60.0	35.9	65.3
Yang et al. [53]	70.0	78.8	66.0	61.1	61.0	37.4	60.0

Table 3. Cross test results on PARSE dataset with models trained on LSP dataset.

The FLIC dataset Compared with LSP and PARSE datasets, the FLIC dataset features real life scenes and is challenging in the localization of elbows and wrists. We also test our method for upper body pose estimation on the large FLIC dataset [37]. We compare with several state-of-the-art models whose codes are available. The result of MODEC [37] is derived from the model trained by the authors. We retrain the FMP model of Yang et al. [53] on the FLIC training set and obtained comparable results as in [37]. The training code of Eichner et al. [11] is not available, thus we use the provided model for test.

As most of the people are not centred in the image in the FLIC dataset, Eichner et al. [11] propose to use OpenCV face detector and DPM [13] upper body detector for rough detection first. The method of MODEC [37] utilized the poselet [5] torso detector for initial detection. However, the approach of ours and Yang [53] do not use the other detectors for initial detection. We follow the evaluation measure of [37] which is similar to the Percentage of Corrected Keypoints(PCK) [54] criterion except that the height of torso is chosen for normalization. As shown in Fig 5, our method outperforms MODEC [37] by 6.3% and 3.2% in AUC[1] respectively on elbows and wrists. The result shows that the modeling of interactions between physically unconnected parts(e.g., left and right wrists) will benefit the localization of lower arms.

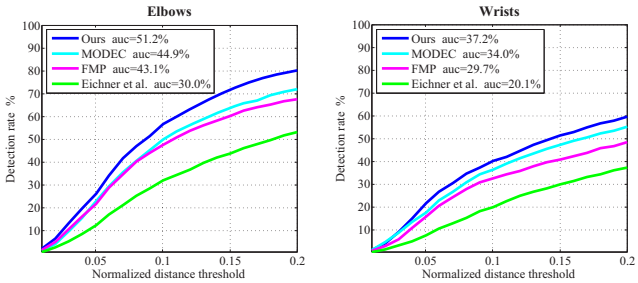

Figure 5. Test results on FLIC dataset. We compared results for the most challenging parts: elbows and wrists. Best viewed in color.

4.3. Experiments on Occlusion

As our model focuses on the problem of occlusion handling in human pose estimation, we specifically design experiments to test the robustness on occlusion. We select images with occlusion from LSP [24] dataset and the MPII [1] dataset for detailed analysis.

Occluded Leeds Sports We evaluate our method on a subset of the LSP [24] test set consisting of 468 images with one more joints occluded. Tab. 4 shows the performance of our method as well as the baseline under different levels of occlusions. It reflects that the performance of FMP [53] drops quickly with more occluded joints. However, the performance of our method only drops slightly when there are less than 4 joints occluded.

# occluded joints	1	2	3	4	5
Ours	67.5	62.9	61.4	53.3	43.6
FMP [53]	59.6	52.7	50.3	47.5	39.1
# test images	174	133	105	37	19

Table 4. Analysis of performance on the LPS occluded subset. Both models are evaluated on the Observer-Centric view.

Occluded MPII We evaluated on a subset of the newly published challenging MPII [1] pose dataset. The selected subset consists of 2198 images with severe occlusion(44.1% of the joints) and is suitable for the evaluation of robustness to occlusion. Though PCP was the most frequently used metric for evaluation, it has the drawback of penalizing shorter limbs. For better evaluation of per joint detection, we adopt the Percentage of Correct Keypoints(PCK) for analysis. Fig. 6 illustrates the performance of our method v.s. the baseline on the Occluded MPII dataset. The chart shows that our method performs better than the FMP approach when there is heavy occlusion.

4.4. Analysis of our model

We design two experiments to better understand the influence of parameter settings on the performance of our model. We evaluate the parameters on the LSP dataset and take the FMP [53] as baseline.

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

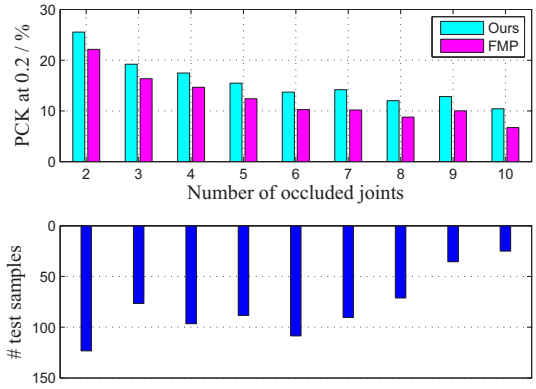

Figure 6. Analysis of occlusion robustness on the MPII subset for the proposed method and the baseline method of FMP.

The learning method of mixture types Tab. 5 shows the performance gain when the local mixtures are learned with our method instead of that of the FMP [53]. It reflects our approach benefits more for lower level parts (limbs) which contain much uncertainty of freedom in a tree structured model.

Parts	Head	Torso	U.Leg	L.Leg	U.Arm	L.Arm	Limbs	Avg
%	0.6	0.4	1.6	1.7	5.8	4.2	3.3	2.8

Table 5. The PCP gain of our approach of learning mixture types w.r.t that of FMP's on the LSP dataset.

The effect of occlusion modeling In terms of occlusion modeling, we considered both self-occlusion and other-occlusion in the proposed model. It is worth analyzing how each feature of the model contribute to the boost of performance.

Model	Head	Torso	U.Leg	L.Leg	U.Arm	L.Arm	Limbs	Avg
FMP	87.4	**92.6**	66.4	47.7	50.0	30.4	51.1	58.9
FMP+O	87.5	92.3	69.2	55.2	52.8	34.7	53.0	60.4
G+S	87.3	91.6	70.1	59.3	59.7	38.2	56.8	63.4
G+S+O	**87.5**	91.5	**74.2**	**66.8**	**62.5**	**41.3**	**61.2**	**66.9**

Table 6. The comparison of PCP(%) with different model structures on LSP dataset.

Tab. 6 shows the result of different model structures: (1)FMP [53], the tree structured model. (2)FMP+O, the tree structured model with other-occlusion reasoning only. (3)G+S, our graphical model with self-occlusion handling only. (4)G+S+O, our graphical model with both self-occlusion and other-occlusion reasoning. We noticed that the localization accuracy of torso and head does not improve since they are rarely occluded. It is observed that the introduce of occlusion states is helpful for improving the accuracy of limbs(especially lower limbs) which are frequently occluded by objects. For instance, there is 3.4% improvement on average PCP of limbs for the FMP+O model v.s. the FMP model, and 2.6% for the G+S+O model v.s. the G+S model. On the other hand, the edges between non-connected body parts can significantly improve the overall PCP(e.g., 5.9% for G+S compared with FMP and 5.4% for G+S+O compared with FMP+O). This is mainly because the constraints among non-connected parts can eliminate double-counting and improve the PCP of limbs.

The influence of parameter σ In section 3.2, we assume that the optimal hypothesis is included in the selected top-σ hypotheses of the unrolled configurations. We analyse how different ratio of σ affects the performance of our method. Tab. 7 reflects the effect of such setting. It shows that the performance almost does not change when the ratio $\sigma > 0.01$.

σ	0.001	0.002	0.005	0.01	0.02	0.05	0.1
PCP	65.8	66.2	67.4	67.7	67.7	67.8	67.9

Table 7. The influence of σ on the performance on LSP dataset for the proposed method.

Fig. 7 qualitatively analyzes the oracle accuracy and the actual accuracy of our method with different number of hypotheses per image. The oracle accuracy reflects the upper bound of our method with the given number of hypotheses selected.

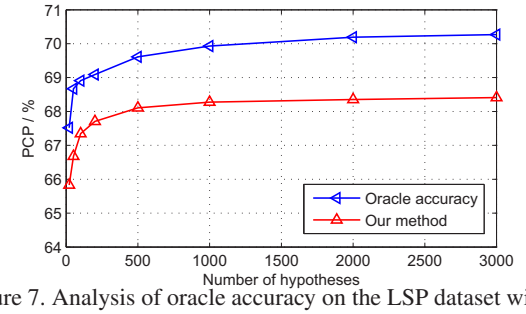

Figure 7. Analysis of oracle accuracy on the LSP dataset with different number of hypotheses selected per image.

5. Conclusion and Future Work

In this paper, we have proposed an occlusion aware graphical model to model both self-occlusion and other-occlusion in human pose estimation. Beyond tree structure model, we explicitly capture the high-order interactions among parts, enabling occlusion handling, especially self-occlusion. We demonstrate that part level occlusion reasoning is important for human pose estimation as occlusion coherence and stronger structural constraints can be embedded in such model. The experimental results show comparable performance of our method compared with the state-of-the-arts. Our method especially obtains promising performance in human pose estimation with occlusion. In the later future, we will try to combine stronger feature representation such as CNN feature to boost the performance of our model further.

Acknowledgment

This work is funded by the National Basic Research Program of China (Grant No. 2012CB316302), National Natural Science Foundation of China (Grant No. 61403387, Grant No. 61322209 and Grant No. 61175007), the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant XDA06040102).

References

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, pages 3686–3693, 2014. 5, 7

[2] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR*, pages 1014–1021, 2009. 2, 6

[3] H. Azizpour and I. Laptev. Object detection using strongly-supervised deformable part models. In *ECCV*, pages 836–849, 2012. 2

[4] L. D. Bourdev, S. Maji, and J. Malik. Describing people: A poselet-based approach to attribute classification. In *ICCV*, pages 1543–1550, 2011. 2

[5] L. D. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, pages 1365–1372, 2009. 7

[6] X. Chen and A. Yuille. Parsing occluded people by flexible compositions. In *CVPR*, 2015. 2

[7] X. Chen and A. L. Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *NIPS*, pages 1736–1744, 2014. 2

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005. 6

[9] M. Dantone, J. Gall, C. Leistner, and L. V. Gool. Human pose estimation using body parts dependent joint regressors. In *CVPR*, pages 3041–3048, 2013. 2, 6

[10] C. Desai and D. Ramanan. Detecting actions, poses, and objects with relational phraselets. In *ECCV*, pages 158–172, 2012. 2

[11] M. Eichner and V. Ferrari. Better appearance models for pictorial structures. In *BMVC*, 2009. 7

[12] M. Eichner and V. Ferrari. Appearance sharing for collective human pose estimation. In *ACCV*, pages 138–151, 2012. 6

[13] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 32(9):1627–1645, 2010. 2, 7

[14] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1):55–79, 2005. 4, 5

[15] V. Ferrari, M. J. Marłn-Jimnez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR*, 2008. 6

[16] M. A. Fischler and R. A. Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on Computers*, 100(1):67–92, 1973. 2

[17] T. Gao, B. Packer, and D. Koller. A segmentation-aware object detection model with occlusion handling. In *CVPR*, pages 1361–1368, 2011. 2

[18] G. Ghiasi and C. Fowlkes. Occlusion coherence: Localizing occluded faces with a hierarchcial deformable part model. In *CVPR*, pages 1899–1906, 2014. 5, 6

[19] G. Ghiasi, Y. Yang, D. Ramanan, and C. Fowlkes. Parsing occluded people. In *CVPR*, pages 2401–2408, 2014. 2

[20] R. B. Girshick, P. F. Felzenszwalb, and D. A. McAllester. Object detection with grammar models. In *NIPS*, pages 442–450, 2011. 2

[21] G. Gkioxari, B. Hariharan, R. Girshick, and J. Malik. Using k-poselets for detecting people and localizing their keypoints. In *CVPR*, pages 3582–3589, 2014. 2

[22] M. Hejrati and D. Ramanan. Analyzing 3d objects in cluttered images. In *NIPS*, pages 602–610, 2012. 2

[23] T. Joachims, T. Finley, and C.-N. J. Yu. Cutting-plane training of structural svms. *Machine Learning*, pages 27–59, 2009. 5

[24] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, pages 12.1–12.11, 2010. 2, 5, 6, 7

[25] S. Johnson and M. Everingham. Learning effective human pose estimation from inaccurate annotation. In *CVPR*, pages 1465–1472, 2011. 6

[26] N. Komodakis, N. Paragios, and G. Tziritas. MRF energy minimization and beyond via dual decomposition. *PAMI*, 33(3):531–552, 2011. 4

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1106–1114, 2012. 2, 6

[28] W. Ouyang, X. Chu, and X. Wang. Multi-source deep learning for human pose estimation. In *CVPR*, pages 2337–2443, 2014. 6, 7

[29] L. Pishchulin, M. Andriluka, P. Gehler, and S. Bernt. Strong appearance and expressive spatial models for human pose estimation. In *ICCV*, pages 3487–3494, 2013. 2, 6

[30] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Poselet conditioned pictorial structures. In *CVPR*, pages 588–595, 2013. 2, 6

[31] I. Radwan, A. Dhall, and R. Goecke. Monocular image 3d human pose estimation under self-occlusion. In *ICCV*, pages 1888–1895, 2013. 2

[32] V. Ramakrishna, D. Munoz, M. Hebert, J. Bagnell, and Y. Sheikh. Posemachines: Articulated pose estimation via inference machines. In *ECCV*, pages 33–47, 2014. 2, 6

[33] D. Ramanan. Learning to parse images of articulated bodies. In *NIPS*, pages 1129–1136, 2006. 5

[34] D. Ramanan. Part-based models for finding people and estimating their pose. In *Visual Analysis of Humans - Looking at People.*, pages 199–223. Springer, 2011. 4

[35] D. Ramanan. Dual coordinate solvers for large-scale structural svms. *CoRR*, abs/1312.1743, 2013. 5

[36] B. Rothrock, S. Park, and S.-C. Zhu. Integrating grammar and segmentation for human pose estimation. In *CVPR*, pages 3214–3221, 2013. 2

[37] B. Sapp and B. Taskar. Modec: Multimodal decomposable models for human pose estimation. In *CVPR*, pages 3674–3681, 2013. 5, 6, 7

[38] B. Sapp, D. Weiss, and B. Taskar. Parsing human motion with stretchable models. In *CVPR*, pages 1281–1288, 2011. 2

[39] L. Sigal and M. J. Black. Measure locally, reason globally: Occlusion-sensitive articulated pose estimation. In *CVPR*, pages 2041–2048, 2006. 2

[40] M. Sun and S. Savarese. Articulated part-based model for joint object detection and pose estimation. In *ICCV*, pages 723–730, 2011. 1, 2, 3

[41] M. Sun, M. Telaprolu, H. Lee, and S. Savarese. An efficient branch-and-bound algorithm for optimal human pose estimation. In *CVPR*, pages 1616–1623, 2012. 2

[42] S. C. Tatikonda and M. I. Jordan. Loopy belief propagation and gibbs measures. In *UAI*, pages 493–500, 2002. 4

[43] T.-P. Tian and S. Sclaroff. Fast globally optimal 2d human detection with loopy graph models. In *CVPR*, pages 81–88, 2010. 2, 4

[44] Y. Tian, C. L. Zitnick, and S. G. Narasimhan. Exploring the spatial hierarchy of mixture models for human pose estimation. In *ECCV*, pages 256–269, 2012. 1, 2, 6

[45] J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *NIPS*, pages 1799–1807, 2014. 2

[46] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, pages 1653–1660, 2014. 2, 6, 7

[47] D. Tran and D. A. Forsyth. Improved human parsing with a full relational model. In *ECCV (4)*, pages 227–240, 2010. 2

[48] F. Wang and Y. Li. Beyond physical connections: Tree models in human pose estimation. In *CVPR*, pages 596–603, 2013. 1, 2, 3, 5, 6, 7

[49] X. Wang, T. X. Han, and S. Yan. An HOG-LBP human detector with partial occlusion handling. In *ICCV*, pages 32–39, 2009. 2

[50] Y. Wang and G. Mori. Multiple tree models for occlusion and spatial constraints in human pose estimation. In *ECCV*, pages 710–724, 2008. 2

[51] Y. Wang, D. Tran, and Z. Liao. Learning hierarchical poselets for human parsing. In *CVPR*, pages 1705–1712, 2011. 2

[52] Y. Weiss. Correctness of local probability propagation in graphical models with loops. *Neural Computation*, 12(1):1–41, 2000. 4

[53] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, pages 1385–1392, 2011. 1, 2, 3, 4, 5, 6, 7, 8

[54] Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures of parts. *PAMI*, 35(12):2878–2890, 2013. 6, 7

[55] Y. Yang and G. Sundaramoorthi. Modeling self-occlusions in dynamic shape and appearance tracking. In *ICCV*, pages 201–208, 2013. 2