Parsing Occluded People by Flexible Compositions

Xianjie Chen
University of California, Los Angeles
Los Angeles, CA 90095
cxj@ucla.edu

Alan Yuille
University of California, Los Angeles
Los Angeles, CA 90095
yuille@stat.ucla.edu

Abstract

This paper presents an approach to parsing humans when there is significant occlusion. We model humans using a graphical model which has a tree structure building on recent work [?, ?] and exploit the prior that, even in presence of occlusion, the visible nodes form a connected subtree of the graphical model. We call each connected subtree a flexible composition of object parts. This involves a novel method for learning occlusion cues. During inference we need to search over a mixture of different flexible models. By exploiting part sharing, we show that this inference can be done extremely efficiently requiring only twice as many computations as searching for the entire object (i.e., not modeling occlusion). We evaluate our model on the standard benchmarked "We Are Family" Stickmen dataset and obtain significant performance improvements over the best alternative algorithms.

1. Introduction

Parsing humans into parts is an important visual task with many applications such as activity recognition [?, ?]. A common approach is to formulate this task in terms of graphical models where the graph nodes and edges represent human parts and their spatial relationships respectively. This approach is becoming successful on benchmarked datasets [?, ?]. But in many real world situations many human parts are occluded. Standard methods are partially robust to occlusion by, for example, using a latent variable to indicate whether a part is present and paying a penalty if the part is not detected, but are not designed to deal with significant occlusion. One of these models [?] will be used in this paper as a *base model*, and we will compare to it.

In this paper, we observe that part occlusions often occur in regular patterns. The visible parts of a human tend to consist of a subset of connected parts even when there is significant occlusion (see Figures 1 and 2(a)). In the terminology of graphical models, the visible (non-occluded) nodes
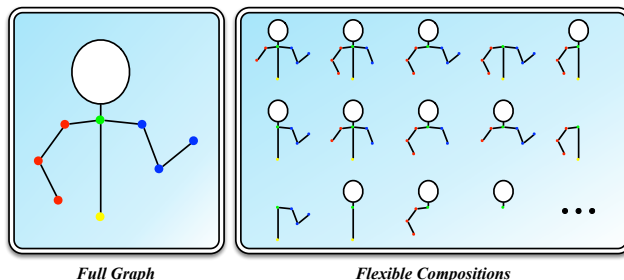
Figure 1: An illustration of the *flexible compositions*. Each connected subtree of the *full graph* (include the full graph itself) is a flexible composition. The flexible compositions that do not have certain parts is suitable for the people with those parts occluded.

form a connected subtree of the full graphical model (following current models, for simplicity, we assume that the graphical model is treelike). This prior is not always valid (*i.e.*, the visible parts of humans may form two or more connected subsets), but our analysis (see Section 6.4) suggests it's often true. In any case, we will restrict ourselves to it in this paper, since verifying that some isolated pieces of body parts belong to the same person is still very difficult for vision methods, especially in challenging scenes where multiple people occlude one another (see Figure 2).

To formulate our approach we build on the base model [?], which is the state of the art on several benchmarked datasets [?, ?, ?], but is not designed for dealing with significant occlusion. We explicitly model occlusions using the prior above. This means that we have a mixture of models where the number of components equals the number of connected subtrees of the graph, which we call *flexible compositions*, see Figure 1. The number of flexible compositions can be large (for a simple chain like model consisting of N parts, there are $N(N + 1)/2$ possible compositions). Our approach exploits the fact there is often local evidence for the presence of occlusions, see Figure 2(b). We propose a novel approach which learns occlusion cues, which can break the links/edges, between adjacent parts in the graphical model. It is well known, of course, that there are lo-

1

cal cues such as T-junctions which can indicate local occlusions. But although these occlusion cues have been used by some models (*e.g.*, [**?**, **?**]), they are not standard in graphical models of objects.

We show that efficient inference can be done for our model by exploiting the sharing of computation between different flexible models. Indeed, the complexity is only doubled compared to recent models where occlusion is not explicitly modeled. This rapid inference also enables us to train the model efficiently from labeled data.

We illustrate our algorithm on the standard benchmarked "We Are Family" Stickmen (WAF) dataset [**?**] for parsing humans when significant occlusion is present. We show strong performance with significant improvement over the best existing method [**?**] and also outperform our base model [**?**]. We perform diagnostic experiments to verify our prior that the visible parts of a human tend to consist of a subset of connected parts even when there is significant occlusion, and quantify the effect of different aspects of our model.

2. Related work

Graphical models of objects have a long history [**?**, **?**]. Our work is most closely related to the recent work of Yang and Ramanan [**?**], Chen and Yuille [**?**], which we use as our base model and will compare to. Other relevant work includes [**?**, **?**, **?**, **?**].

Occlusion modeling also has a long history. Psychophysical studies (*e.g.*, Kanizsa [**?**]) show that T-junctions are a useful cue for occlusion. But with a few exceptions, *e.g.*, [**?**], there has been little attempt to model the spatial patterns of occlusions. Instead it is more common to design models so that they are robust in the presence of occlusion, so that the model is not penalized very much if an object part is missing. Girshick et. al. [**?**] and Supervised-DPM [**?**] model the occluded part (background) using extra templates. And they rely on a root part (*i.e.*, the holistic object) that never takes the status of "occluded". When there is significant occlusion, modeling the root part itself is difficult. Ghiasi et. al. [**?**] advocates modeling the occlusion area (background) using more templates (mixture of templates), and localizes every body parts. It may be plausible to "guess" the occluded keypoints of face (*e.g.*, [**?**, **?**]), but seems impossible for body parts of people, due to highly flexible human poses. Eichner and Ferrari [**?**] handles occlusion by modeling interactions between people, which assumes the occlusion is due to other people.

Our approach models object occlusion effectively uses a mixture of models to deal with different occlusion patterns. There is considerable work which models objects using mixtures to deal with different configurations, see Poselets [**?**] which uses many mixtures to deal with different object configurations, and deformable part models (DPMs) [**?**]

where mixtures are used to deal with different viewpoints.

To ensure efficient inference, we exploit the fact that parts are shared between different flexible compositions. This sharing of parts has been used in other work, e.g., [**?**]. Other work that exploits part sharing includes compositional models [**?**] and AND-OR graphs [**?**, **?**].

3. The Model

We represent human pose by a graphical model $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where the nodes $\mathcal{V}$ corresponds to the parts (or joints) and the edges $\mathcal{E}$ indicate which parts are directly related. For simplicity, we impose that the graph structure forms a $K-$node tree, where $K = |\mathcal{V}|$. The pixel location of part part i is denoted by $\mathbf{l}_i = (x, y)$, for $i \in \{1, \ldots, K\}$.

To model the spatial relationship between neighboring parts $(i, j) \in \mathcal{E}$, we follow the base model [**?**] to discretize the pairwise spatial relationships into a set indexed by t_{ij}, which corresponds to a mixture of different spatial relationships.

In order to handle people with different degrees of occlusion, we specify a binary occlusion decoupling variable $\gamma_{ij} \in \{0, 1\}$ for each edge $(i, j) \in \mathcal{E}$, which enables the subtree $\mathcal{T}_j = (\mathcal{V}(\mathcal{T}_j), \mathcal{E}(\mathcal{T}_j))$ rooted at part j to be decoupled from the graph at part i (the subtree does not contain part i, *i.e.*, $i \notin \mathcal{V}(\mathcal{T}_j)$). This results in a set of flexible compositions of the graph, indexed by set $\mathcal{C}_{\mathcal{G}}$. These compositions share the nodes and edges with the full graph $\mathcal{G}$ and each of themselves forms tree graph (see Figure 1). The compositions that do not have certain parts is suitable for the people with those parts occluded.

In this paper, we exploit the prior that body parts tend to be connected even in the presence of occlusion, and do not consider the cases when people are separated into isolated pieces, which is very difficult. Handling these cases typically requires non-tree models, *e.g.*,[**?**], and thus does not have exact and efficient inference algorithms. Moreover, verifying whether some isolated pieces of people belong to the same person is still very difficult for vision methods, especially in challenging scenes where multiple people usually occlude one another (see Figure 2(a)).

For each flexible composition $\mathcal{G}_c = (\mathcal{V}_c, \mathcal{E}_c), c \in \mathcal{C}_{\mathcal{G}}$, we will define a score function $F(\mathbf{l}, \mathbf{t}, \mathcal{G}_c | \mathbf{I}, \mathcal{G})$ as a sum of appearance terms, pairwise relational terms, occlusion decoupling terms and decoupling bias terms. Here $\mathbf{I}$ denotes the image, $\mathbf{l} = \{\mathbf{l}_i | i \in \mathcal{V}\}$ is the set of locations of the parts, and $\mathbf{t} = \{\mathbf{t}_{ij}, \mathbf{t}_{ji} | (i, j) \in \mathcal{E}\}$ is the set of spatial relationships.

Appearance Terms: The appearance terms make use of the local image measurement within patch $\mathbf{I}(\mathbf{l}_i)$ to provide evidence for part i to lie at location $\mathbf{l}_i$. They are of form:

$$A(\mathbf{l}_i | \mathbf{I}) = w_i \phi(i | \mathbf{I}(\mathbf{l}_i); \boldsymbol{\theta}), \tag{1}$$
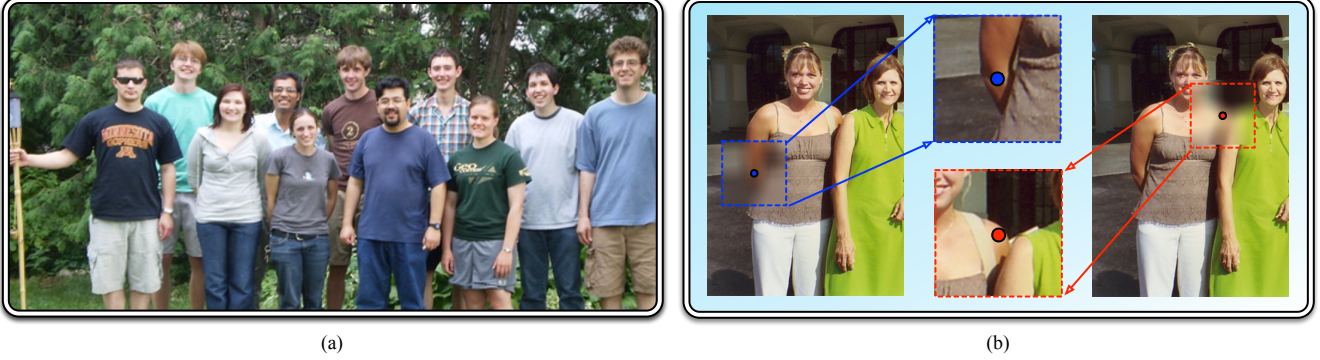
Figure 2: Motivation. (a): In real world scenes, people are usually significantly occluded (or truncated). Requiring the model to localize a fixed set of body parts while ignoring the fact that different people have different degrees of occlusion (or truncation) is problematic. (b): The absence of body parts evidence can help to predict occlusion, *e.g.*, the right wrist of the lady in brown can be inferred as occluded because of the absence of suitable wrist near the elbow. However, absence of evidence is not evidence of absence. It can fail in some challenging scenes, for example, even though the left arm of the lady in brown is completely occluded, there is still strong image evidence of suitable elbow and wrist at the plausible locations due to the confusion caused by nearby people (*e.g.*, the lady in green). In both situations, the local image measurements near the occlusion boundary (*i.e.*, around the right elbow and left shoulder), *e.g.*, in a image patch, can reliably provide evidence of occlusion.

where $\phi(.|.;\boldsymbol{\theta})$ is the (scalar-valued) appearance term with $\boldsymbol{\theta}$ as its parameters (specified in Section 3.1), and w_i is a scalar weight parameter.

Image Dependent Pairwise Relational (IDPR) Terms: We follow the base model [**?**] to use image dependent pairwise relational (IDPR) terms, which gives stronger spatial constraints between neighboring parts $(i,j) \in \mathcal{E}$. Stronger spatial constraints reduce the confusion from the nearby people and clustered background, which helps to better infer occlusion.

More formally, the relative positions between parts i and j are discretized into several types $t_{ij} \in \{1, \ldots, T_{ij}\}$ (*i.e.*, a mixture of different relationships) with corresponding mean relative positions $\mathbf{r}_{ij}^{t_{ij}}$ plus small deformations which are modeled by the standard quadratic deformation term. They are given by:

$$
\begin{aligned}
R(\mathbf{l}_i, \mathbf{l}_j, t_{ij}, t_{ji}|\mathbf{I}) = &\langle \mathbf{w}_{ij}^{t_{ij}}, \boldsymbol{\psi}(\mathbf{l}_j - \mathbf{l}_i - \mathbf{r}_{ij}^{t_{ij}})\rangle \\
&+ w_{ij}\varphi^s(t_{ij}, \gamma_{ij} = 0|\mathbf{I}(\mathbf{l}_i); \boldsymbol{\theta}) \\
&+ \langle \mathbf{w}_{ji}^{t_{ji}}, \boldsymbol{\psi}(\mathbf{l}_i - \mathbf{l}_j - \mathbf{r}_{ji}^{t_{ji}})\rangle \\
&+ w_{ji}\varphi^s(t_{ji}, \gamma_{ji} = 0|\mathbf{I}(\mathbf{l}_j); \boldsymbol{\theta})
\end{aligned}
\quad (2)
$$

where $\boldsymbol{\psi}(\Delta \mathbf{l} = [\Delta x, \Delta y]) = [\Delta x\ \Delta x^2\ \Delta y\ \Delta y^2]^{\mathsf{T}}$ are the standard quadratic deformation features, $\varphi^s(., \gamma_{ij} = 0|.; \boldsymbol{\theta})$ is the Image Dependent Pairwise Relational (IDPR) term with $\boldsymbol{\theta}$ as its parameters (specified in Section 3.1). IDPR terms are only defined when both part i and j are visible (*i.e.*, $\gamma_{ij} = 0$ and $\gamma_{ji} = 0$). Here $\mathbf{w}_{ij}^{t_{ij}}, w_{ij}, \mathbf{w}_{ji}^{t_{ji}}, w_{ji}$ are the weight parameters, and the notation $\langle.,.\rangle$ specifies dot product and boldface indicates a vector.

Image Dependent Occlusion Decoupling (IDOD) Terms: These IDOD terms capture our intuition that the visible part

i near the occlusion boundary (and thus is a leaf node in each flexible composition) can reliably provide occlusion evidence using *only* local image measurement (see Figure 2(b) and Figure 3). More formally, the occlusion decoupling score for decoupling the subtree $\mathcal{T}_j$ from the full graph at part i is given by:

$$
D_{ij}(\gamma_{ij} = 1, \mathbf{l}_i|\mathbf{I}) = w_{ij}\varphi^d(\gamma_{ij} = 1|\mathbf{I}(\mathbf{l}_i); \boldsymbol{\theta}), \quad (3)
$$

where $\varphi^d(\gamma_{ij} = 1|.; \boldsymbol{\theta})$ is the Image Dependent Occlusion Decoupling (IDOD) term with $\boldsymbol{\theta}$ as its parameters (specified in Section 3.1), $\gamma_{ij} = 1$ indicates subtree $\mathcal{T}_j$ is decoupled from the full graph. Here w_{ij} is the scalar weight parameter shared with the IDPR term.

Decoupling Bias Term: The decoupling bias term captures our intuition that the absence of evidence for suitable body part can help to predict occlusion. We specify a scalar bias term b_i for each part i as a learned measure for the absence of good part appearance, and also the absence of suitable spatial coupling with neighboring parts (our spatial constraints are also image dependent).

The decoupling bias term for decoupling the subtree $\mathcal{T}_j = (\mathcal{V}(\mathcal{T}_j), \mathcal{E}(\mathcal{T}_j))$ from the full graph at part i, is defined as the sum of all the bias terms associated with the parts in the subtree, *i.e.*, $k \in \mathcal{V}(\mathcal{T}_j)$. They are of form:

$$
B_{ij} = \sum_{k \in \mathcal{V}(\mathcal{T}_j)} b_k \quad (4)
$$

The Model Score: The model score for a person is the maximum score of all the flexible compositions $c \in \mathcal{C}_{\mathcal{G}}$, therefore the index c of the flexible composition is also a random variable that need to be estimated, which is different from the standard graphical models with single graph structure.
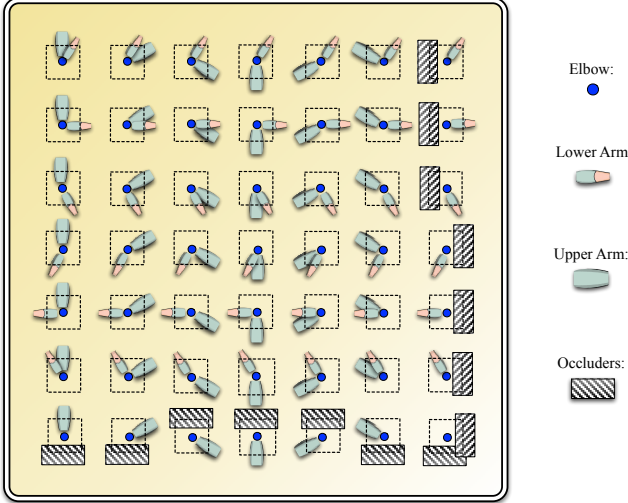
Figure 3: Different occlusion decoupling and spatial relationships between the elbow and its neighbors, *i.e.*, wrist and shoulder. The local image measurement around a part (*e.g.*, the elbow) can reliably predict the relative positions of its neighbors when they are not occluded, which is demonstrated in the base model [?]. In the case when the neighboring parts are occluded, the local image measurement can also reliably provide evidence for the occlusion.

The score $F(\mathbf{l}, \mathbf{t}, \mathcal{G}_c | \mathbf{I}, \mathcal{G})$ for each flexible composition $c \in \mathcal{C}_{\mathcal{G}}$ is a function of the locations $\mathbf{l}$, the pairwise spatial relation types $\mathbf{t}$, the index of the flexible composition c, the structure of the full graph $\mathcal{G}$, and the input image $\mathbf{I}$. It is given by:

$$
\begin{aligned}
F(\mathbf{l}, \mathbf{t}, \mathcal{G}_c | \mathbf{I}, \mathcal{G}) = &\sum_{i \in \mathcal{V}_c} A(\mathbf{l}_i | \mathbf{I}) \\
&+ \sum_{(i,j) \in \mathcal{E}_c} R(\mathbf{l}_i, \mathbf{l}_j, t_{ij}, t_{ji} | \mathbf{I}) \\
&+ \sum_{(i,j) \in \mathcal{E}_c^d} (B_{ij} + D_{ij}(\gamma_{ij} = 1, \mathbf{l}_i | \mathbf{I}))
\end{aligned}
\tag{5}
$$

where $\mathcal{E}_c^d = \{(i,j) \in \mathcal{E} | i \in \mathcal{V}_c, j \notin \mathcal{V}_c\}$ is the edges that is decoupled to generate the composition $\mathcal{G}_c$. See Section 5 for the learning of the model parameters.

3.1. Deep Convolutional Neural Network (DCNN) for Image Dependent Terms

Our model has three kinds of terms that depend on the local image patches: the appearance terms, IDPR terms and IDOD terms. This requires us to have a method that can efficiently extract information from a local image patch $\mathbf{I}(\mathbf{l}_i)$ for the presence of the part i, as well as the occlusion decoupling evidence $\gamma_{ij} = 1$ of its neighbors $j \in \mathcal{N}(i)$, where $j \in \mathcal{N}(i)$ if, and only if, $(i,j) \in \mathcal{E}$. When a neighboring part j is not occluded, *i.e.* $\gamma_{ij} = 0$, we also need to extract

information for the pairwise spatial relationship type t_{ij} between parts i and j.

Extending the base model [?], we learn the distribution for the state variables i, t_{ij}, γ_{ij} conditioned on the image patches $\mathbf{I}(\mathbf{l}_i)$. We'll first define the state space of this distribution.

Let g be the random variable which denotes which part is present $g = i$ for $i \in \{1, ..., K\}$ or $g = 0$ if no part is present (*i.e.*, the background). We define $\mathbf{m}_{g\mathcal{N}(g)} = \{m_{gk} | k \in \mathcal{N}(g)\}$ to be the random variable that determines the pairwise occlusion decoupling and spatial relationships between part g and all its neighbors $\mathcal{N}(g)$, and takes values in $\mathcal{M}_{g\mathcal{N}(g)}$. If $g = i$ has one neighbor j (*e.g.*, the wrist), then $\mathcal{M}_{i\mathcal{N}(i)} = \{0, 1, \ldots, T_{ij}\}$, where value 0 represent part j is occluded, *i.e.*, $\gamma_{ij} = 1$ and values $v \neq 0$ represents part j is not occluded and has corresponding spatial relationship types with part i, *i.e.*, $\gamma_{ij} = 0, t_{ij} = v$. If $g = i$ has two neighbors j and k (*e.g.*, the elbow), then $\mathcal{M}_{i\mathcal{N}(i)} = \{0, 1, \ldots, T_{ij}\} \times \{0, 1, \ldots, T_{ik}\}$ (Figure 3 illustrates the space $\mathcal{M}_{i\mathcal{N}(i)}$ for the elbow when $T_{ik} = T_{ij} = 6$). If $g = 0$, then we define $\mathcal{M}_{0\mathcal{N}(0)} = \{0\}$.

The full space can be written as:

$$
\mathcal{U} = \cup_{g=0}^{K} \{g\} \times \mathcal{M}_{g\mathcal{N}(g)}
\tag{6}
$$

The size of the space is $|\mathcal{U}| = \sum_{g=0}^{K} |\mathcal{M}_{g\mathcal{N}(g)}|$. Each element in this space corresponds to the background or a part with a kind of occlusion decoupling configurations of all its neighbors and the types of its pairwise spatial relationships with its visible neighbors.

With the space of the distribution defined, we use a single Deep Convolutional Neural Network (DCNN) [?], which is efficient and shares features between different parts, to learn the conditional probability distribution $p(g, \mathbf{m}_{g\mathcal{N}(g)} | \mathbf{I}(\mathbf{l}_i); \boldsymbol{\theta})$. See Section 5 for more details.

We specify the appearance terms $\phi(.|.; \boldsymbol{\theta})$, IDPR terms $\varphi^s(., \gamma_{ij} = 0|.; \boldsymbol{\theta})$ and IDOD terms $\varphi^d(\gamma_{ij} = 1|.; \boldsymbol{\theta})$ in terms of $p(g, \mathbf{m}_{g\mathcal{N}(g)} | \mathbf{I}(\mathbf{l}_i); \boldsymbol{\theta})$ by marginalization:

$$
\phi(i|\mathbf{I}(\mathbf{l}_i); \boldsymbol{\theta}) = \log(p(g = i|\mathbf{I}(\mathbf{l}_i); \boldsymbol{\theta}))
\tag{7}
$$
$$
\varphi^s(t_{ij}, \gamma_{ij} = 0|\mathbf{I}(\mathbf{l}_i); \boldsymbol{\theta}) = \log(p(m_{ij} = t_{ij}|g = i, \mathbf{I}(\mathbf{l}_i); \boldsymbol{\theta}))
\tag{8}
$$
$$
\varphi^d(\gamma_{ij} = 1|\mathbf{I}(\mathbf{l}_i); \boldsymbol{\theta}) = \log(p(m_{ij} = 0|g = i, \mathbf{I}(\mathbf{l}_i); \boldsymbol{\theta}))
\tag{9}
$$

4. Inference

To estimate the optimal configuration for each person, we search for the flexible composition $c \in \mathcal{C}_{\mathcal{G}}$ with the configurations of the locations $\mathbf{l}$ and types $\mathbf{t}$ that maximize the model score: $(c^*, \mathbf{l}^*, \mathbf{t}^*) = \arg\max_{c, \mathbf{l}, \mathbf{t}} F(\mathbf{l}, \mathbf{t}, \mathcal{G}_c | \mathbf{I}, \mathcal{G})$.

Let $\mathcal{C}_{\mathcal{G}}^i \subset \mathcal{C}_{\mathcal{G}}$ be the subset of the flexible compositions that have node i present (Obviously, $\cup_{i \in \mathcal{V}} \mathcal{C}_{\mathcal{G}}^i = \mathcal{C}_{\mathcal{G}}$), and we

will consider the compositions that have the part with index 1 present first, *i.e.*, $\mathcal{C}_{\mathcal{G}}^1$.

For all the flexible compositions $c \in \mathcal{C}_{\mathcal{G}}^1$, we set part 1 as root. We will use dynamic programming to compute the best score over all these flexible compositions for each root location $\mathbf{l}_1$.

After setting the root, let $\mathcal{K}(i)$ be the set of children of part i in the full graph ($\mathcal{K}(i) = \emptyset$, if part i is a leaf). We use the following algorithm to compute the maximum score of all the flexible compositions $c \in \mathcal{C}_{\mathcal{G}}^1$:

$$S_i(\mathbf{l}_i|\mathbf{I}) = A(\mathbf{l}_i|\mathbf{I}) + \sum_{k \in \mathcal{K}(i)} m_{ki}(\mathbf{l}_i|\mathbf{I}) \tag{10}$$

$$B_{ij} = b_j + \sum_{k \in \mathcal{K}(j)} B_{jk} \tag{11}$$

$$m_{ki}(\mathbf{l}_i|\mathbf{I}) = \max_{\gamma_{ik}}((1 - \gamma_{ik}) \times m_{ki}^s(\mathbf{l}_i|\mathbf{I})$$
$$+ \gamma_{ik} \times m_{ki}^d(\mathbf{l}_i|\mathbf{I})) \tag{12}$$

$$m_{ki}^s(\mathbf{l}_i|\mathbf{I}) = \max_{\mathbf{l}_k, t_{ik}, t_{ki}} R(\mathbf{l}_i, \mathbf{l}_k, t_{ik}, t_{ki}|\mathbf{I}) + S_k(\mathbf{l}_k|\mathbf{I}) \tag{13}$$

$$m_{ki}^d(\mathbf{l}_i|\mathbf{I}) = D_{ik}(\gamma_{ik} = 1, \mathbf{l}_i|\mathbf{I}) + B_{ik}, \tag{14}$$

where $S_i(\mathbf{l}_i|\mathbf{I})$ is the score of the subtree $\mathcal{T}_i$ with part i each location $\mathbf{l}_i$, and is computed by collecting the messages from all its children $k \in \mathcal{K}(i)$. Each child computes two kinds of messages $m_{ki}^s(\mathbf{l}_i|\mathbf{I})$ and $m_{ki}^d(\mathbf{l}_i|\mathbf{I})$ that convey information to parent for deciding whether to decouple it (and its followed subtree), *i.e.*, Equation 12.

Intuitively, the message computed by Equation 13 measures how well we can find a child part k that not only shows strong evidence of part k (*e.g.*, an elbow) and couples well with the other parts in the subtree $\mathcal{T}_k$ (*i.e.*, $S_k(\mathbf{l}_k|\mathbf{I})$), but also is suitable for the part i at location $\mathbf{l}_i$ based on the local image measurement (encoded in the IDPR terms). The message computed by Equation 14 measures the evidence to decouple $\mathcal{T}_k$ by combining the local image measurements around part i (encoded in IDOD terms) and the learned occlusion decoupling bias.

The following lemma states each $S_i(\mathbf{l}_i|\mathbf{I})$ computes the maximum score for the set of flexible compositions $\mathcal{C}_{\mathcal{T}_i}^i$ that's within the subtree $\mathcal{T}_i$ and have part i at $\mathbf{l}_i$. In other words, we consider an object that's only composed with the parts in the subtree $\mathcal{T}_i$ (*i.e.*, $\mathcal{T}_i$ is the full graph) and $\mathcal{C}_{\mathcal{T}_i}^i$ is the set of flexible compositions of the graph $\mathcal{T}_i$ that have part i present. Since at root part (*i.e.*, i = 1), we have $\mathcal{T}_1 = \mathcal{G}$, once the messages are passed to the root part, $S_1(\mathbf{l}_1|\mathbf{I})$ gives the best score for all the flexible compositions $c \in \mathcal{C}_{\mathcal{G}}^1$ that have part 1 at $\mathbf{l}_1$.

Lemma 1.

$$S_i(\mathbf{l}_i, \mathbf{I}) = \max_{c \in \mathcal{C}_{\mathcal{T}_i}^i} \left\{ \max_{\mathbf{l}_{/i}, \mathbf{t}} F(\mathbf{l}_i, \mathbf{l}_{/i}, \mathbf{t}, \mathcal{G}_c|\mathbf{I}, \mathcal{T}_i) \right\} \tag{15}$$

Proof. We will prove the lemma by induction from leaf to root.

Basis: The proof is trivial when node i is a leaf node.

Inductive step: Assume for each child $k \in \mathcal{K}(i)$ of the node i, the lemma holds. Since we do not consider the case that people are separated into isolated pieces, each flexible composition at node i (*i.e.*, $c \in \mathcal{C}_{\mathcal{T}_i}^i$) is composed of part i and the flexible compositions from the children (*i.e.*, $\mathcal{C}_{\mathcal{T}_k}^k, k \in \mathcal{K}(i)$) that are not decoupled. Since the graph is a tree, the best scores of the flexible compositions from each child can be computed separately, by $S_i(\mathbf{l}_k, \mathbf{I}), k \in \mathcal{K}(i)$ as assumed. These scores are then passed to node i (Equation 13). At node i the algorithm can choose to decouple a child for better score (Equation 12). Therefore, the best score at node i is also computed by the algorithm. By induction, the lemma holds for all the nodes. $\square$

By Lemma 1, we can efficiently compute the best score for all the compositions with part 1 present, *i.e.*, $c \in \mathcal{C}_{\mathcal{G}}^1$, at each locations of part 1 by dynamic programming (DP). These scores can be thresholded to generate multiple estimations with part 1 present in an image. The corresponding configurations of locations and types can be recovered by the standard backward pass of DP until occlusion decoupling, *i.e.* $\gamma_{ik} = 1$ in Equation 12. All the decoupled parts are inferred as occluded and thus do not have location or pairwise type configurations.

Since $\cup_{i \in \mathcal{V}} \mathcal{C}_{\mathcal{G}}^i = \mathcal{C}_{\mathcal{G}}$, we can get the best score for all the flexible compositions of the full graph $\mathcal{G}$ by computing the best score for each subset $\mathcal{C}_{\mathcal{G}}^i, i \in \mathcal{V}$. More formally:

$$\max_{c \in \mathcal{C}_{\mathcal{G}}, \mathbf{l}, \mathbf{t}} F(\mathbf{l}, \mathbf{t}, \mathcal{G}_c|\mathbf{I}, \mathcal{G}) = \max_{i \in \mathcal{V}} (\max_{c \in \mathcal{C}_{\mathcal{G}}^i, \mathbf{l}, \mathbf{t}} F(\mathbf{l}, \mathbf{t}, \mathcal{G}_c|\mathbf{I}, \mathcal{G})) \tag{16}$$

This can be done by repeating the DP procedure K times, letting each part take its turn as the root. However, it turns out the messages on each edge only need to be computed twice, one for each direction. This allows us to implement an efficient message passing algorithm, which is of twice (instead of K times) the complexity of the standard one-pass DP, to get the best score for all the flexible compositions.

Computation: As discussed above, the inference is of twice the complexity of the standard one-pass DP. Moreover, the max operation over the locations $\mathbf{l}_k$ in Equation 13, which is a quadratic function of $\mathbf{l}_k$, can be accelerated by the generalized distance transforms [?]. The resulting approach is very efficient, takes $O(2T^2LK)$ time once the image dependent terms are computed, where T is the number of spatial relation types, L is the total number of locations, and K is the total number of parts in the model. This analysis assumes that all the pairwise spatial relationships have the same number of types, *i.e.*, $T_{ij} = T_{ji} = T, \forall (i, j) \in \mathcal{E}$.

The computation of the image dependent terms is also efficient. They are computed over all the locations by a single DCNN. The DCNN is applied in a sliding window fashion by considering the fully-connected layers as 1×1 convolutions [?], which naturally shares the computations common to overlapping regions.

5. Learning

We learn our model parameters from the images containing occluded people. The visibility of each part (or joint) is labeled, and the locations of the visible parts are annotated. We adopt a supervised approach to learn the model by first deriving the occlusion decoupling labels γ_{ij} and type labels t_{ij} from the annotations.

Our model consists of three sets of parameters: the mean relative positions $\mathbf{r} = \left\{ \mathbf{r}_{ij}^{t_{ij}}, \mathbf{r}_{ji}^{t_{ji}} | (i,j) \in \mathcal{E} \right\}$ of different pairwise spatial relation types; the parameters $\boldsymbol{\theta}$ for the image dependent terms, *i.e.*, the appearance terms, IDPR and IDOD terms; and the weight parameters $\mathbf{w}$(*i.e.*, $w_i, \mathbf{w}_{ij}^{t_{ij}}, w_{ij}, \mathbf{w}_{ji}^{t_{ji}}, w_{ji}$), and bias parameters $\mathbf{b}$ (*i.e.*, b_k). They are learnt separately by the K-means algorithm for $\mathbf{r}$, DCNN for $\boldsymbol{\theta}$, and linear Support Vector Machine (SVM) [?] for $\mathbf{w}$ and $\mathbf{b}$.

Derive Labels and Learn Mean Relative Positions: The ground-truth annotations give the part visibility labels $\mathbf{v}^n$, and locations $\mathbf{l}^n$ for visible parts of each person instance $n \in \{1, \ldots, N\}$. For each neighboring parts $(i,j) \in \mathcal{E}$, we derive $\gamma_{ij}^n = 1$ if and only if part i is visible but part j is not, *i.e.*, $v_i^n = 1$ and $v_j^n = 0$. Let $\mathbf{d}_{ij}$ be the relative position from part i to its neighbor j, if both of them are visible. We cluster the relative positions over the training set $\left\{ \mathbf{d}_{ij}^n | v_i^n = 1, v_j^n = 1 \right\}$ to get T_{ij} clusters (in the experiments $T_{ij} = 8$ for all pairwise relations). Each cluster corresponds to a set of instances of part i that share similar spatial relationship with its visible neighboring part j. Therefore, we define each cluster as a pairwise spatial relation type t_{ij} from part i to j in our model, and the type label t_{ij}^n for each training instance is derived based on its cluster index. The mean relative position $\mathbf{r}_{ij}^{t_{ij}}$ associated with each type is defined as the the center of each cluster. In our experiments, we use K-means by setting $K = T_{ij}$ to do the clustering.

Parameters of Image Dependent Terms: After deriving the occlusion decoupling label and pairwise spatial type labels, each local image patch $\mathbf{I}(\mathbf{l}^n)$ centered at an annotated (visible) part location is labeled with category label $g^n \in \{1, \ldots, K\}$, that indicates which part is present, and also the type labels $\mathbf{m}_{g^n \mathcal{N}(g^n)}^n$ that indicate its pairwise occlusion decoupling and spatial relationships with all its neighbors. In this way, we get a set of labeled patches $\left\{ \mathbf{I}(\mathbf{l}^n), g^n, \mathbf{m}_{g^n \mathcal{N}(g^n)}^n | v_{g^n}^n = 1 \right\}$ from the visible parts of each labeled people, and also a set of background patches

$\{\mathbf{I}(\mathbf{l}^n), 0, 0\}$ sampled from negative images, which do not contain people.

Given the labeled part patches and background patches, we train a $|\mathcal{U}|$-way DCNN classifier by standard stochastic gradient descent using softmax loss. The final $|\mathcal{U}|$-way softmax output is defined as our conditional probability distribution, *i.e.*, $p(g, \mathbf{m}_{g \mathcal{N}(g)} | \mathbf{I}(\mathbf{l}_i); \boldsymbol{\theta})$. See Section 6.2 for the details of our network.

Weight and Bias Parameters: Given the derived occlusion decoupling labels γ_{ij}, we can associate each labeled pose with a flexible composition c^n. For the poses that's separately into several isolated compositions, we use the composition with the most number of parts. The location of each visible part in the associated composition c^n is given by the ground-truth annotation, and the pairwise spatial types of it are derived above. We can then compute the model score of each labeled pose as a linear function of the parameters $\boldsymbol{\beta} = [\mathbf{w}, \mathbf{b}]$, so we use a linear SVM to learn these parameters:

$$
\begin{aligned}
\min_{\boldsymbol{\beta}, \xi} \quad & \frac{1}{2} \langle \boldsymbol{\beta}, \boldsymbol{\beta} \rangle + C \sum_n \xi_n \\
\text{s.t.} \quad & \langle \boldsymbol{\beta}, \boldsymbol{\Phi}(c^n, \mathbf{I}^n, \mathbf{l}^n, \mathbf{t}^n) \rangle + b_0 \geq 1 - \xi_n, \forall n \in \text{pos} \\
& \langle \boldsymbol{\beta}, \boldsymbol{\Phi}(c^n, \mathbf{I}^n, \mathbf{l}^n, \mathbf{t}^n) \rangle + b_0 \leq -1 + \xi_n, \forall n \in \text{neg}
\end{aligned}
$$

where b_0 is the scalar SVM bias, C is the cost parameter, and $\boldsymbol{\Phi}(c^n, \mathbf{I}^n, \mathbf{l}^n, \mathbf{t}^n)$ is a sparse feature vector representing the n-th example and is the concatenation of the image dependent terms (calculated from the learnt DCNN), spatial deformation features, and constants 1s for the bias terms. The above constraints encourage the positive examples (pos) to be scored higher than 1 (the margin) and the negative examples (neg), which we mine from the negative images using the inference method described above, lower than -1. The objective function penalizes violations using slack variables ξ_i.

6. Experiments

This section describes our experimental setup, presents comparison benchmark results, and gives diagnostic experiments.

6.1. Dataset and Evaluation Metrics

We perform experiments on the standard benchmarked dataset: "We Are Family" Stickmen (WAF) [?], which contains challenging group photos, where several people often occlude one another (see Figure 5). The dataset contains 525 images with 6 people each on average, and is officially split into 350 images for training and 175 images for testing. Following [?, ?], we use the negative training images from the INRIAPerson dataset [?] (These images do not contain people).
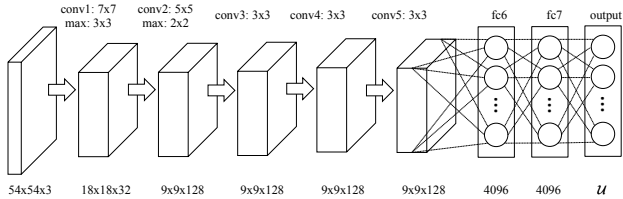
Figure 4: An illustration of the DCNN architecture used in our experiments. It consists of five convolutional layers (conv), 2 max-pooling layers (max) and three fully-connected layers (fc) with a final $|\mathcal{U}|$-way softmax output. We use the rectification (ReLU) non-linearity, and the dropout technique described in [**?**].

Method	AOP	Torso	Head	U.arms	L.arms	mPCP
Ours	**84.9**	**88.5**	**98.5**	**77.2**	**71.3**	**80.7**
Multi-Person [**?**]	80.0	86.1	97.6	68.2	48.1	69.4
Ghiasi et. al. [**?**]	74.0	-	-	-	-	63.6
One-Person [**?**]	73.9	83.2	97.6	56.7	28.6	58.6

Table 1: Comparison of PCP and AOP on the WAF dataset. Our method improves the PCP performance on all parts, and significantly outperform the best previously published result [**?**] by 11.3% on mean PCP, and 4.9% on AOP.

We evaluate our method using the official toolkit of the dataset [**?**] to allow comparison with previous work. The toolkit implements a version of occlusion-aware Percentage of Correct Parts (PCP) metric, where an estimated part is considered correctly localized if the *average* distance between its endpoints (joints) and ground-truth is less than 50% of the length of the ground-truth annotated endpoints, and an occluded body part is considered correct if and only if the part is also annotated as occluded in the ground-truth.

We also evaluate the Accuracy of Occlusion Prediction (AOP) by considering occlusion prediction over all people parts as a binary classification problem. AOP does not care how well a part is localized, but is aimed to show the percentage of parts that have its visibility status correctly estimated.

6.2. Implementation detail

DCNN Architecture: The layer configuration of our network is summarized in Figure 4. In our experiments, the patch size of each part is 54×54. We pre-process each image patch pixel by subtracting the mean pixel value over all the pixels of training patches. We use the Caffe [**?**] implementation of DCNN.

Data Augmentation: We augment the training data by rotating and horizontally flipping the positive training examples to increase the number of training part patches with different spatial configurations with its neighbors. We follow [**?, ?**] to increase the number of parts by adding the midway points between annotated parts, which results in 15 parts on the WAF dataset. Increasing the number of parts produce more training patches for DCNN, which helps to reduce overfitting. Also covering a person with more parts is good for modeling foreshortening [**?**].

Part-based Non-Maximum Suppression: Using the proposed inference algorithm, a single image evidence of a part can be used multiple times in different estimations. This may produce duplicated estimations for the same person. We use a greedy part-based non-maximum suppression [**?**] to prevent this. There is a score associated to each estimation. We sort the estimations by their score and start from

the highest scoring estimation and remove the ones whose parts overlap significantly with the corresponding parts of any previously selected estimations. In the experiments, we require the interaction over union between the corresponding parts in different estimation to be less than 60%.

Other Settings: We use the same number of spatial relationship types for all pairs of neighbors in our experiments. They are set as 8, *i.e.*, $T_{ij} = T_{ji} = 8, \forall (i,j) \in \mathcal{E}$.

6.3. Benchmark results

Table 1 compares the performance of our method with the state of the art methods using the PCP and AOP metrics on the WAF benchmark, which shows our method improves the PCP performance on all parts, and significantly outperform the best previously published result [**?**] by 11.3% on mean PCP, and 4.9% on AOP. Figure 5 shows some estimation results of our model on the WAF dataset.

6.4. Diagnostic Experiments

Prior Verification: We analyze the test set of the WAF dataset using ground truth annotation, and find that 95.1% of the people instances have their visible parts form a connected subtree. This verifies our prior that visible parts of a human tend to form a connected subtree, even in the presence of significant occlusion.

Term Analysis: We design the following experiments to better understand each design component in our model.

Firstly, our model is designed to handle different degrees of occlusion by efficiently searching over large number of flexible compositions. When we do not consider occlusion and use a single composition (*i.e.*, the full graph), our model reduces to the base model [**?**]. So we perform an diagnostic experiment by using the base model [**?**] on the WAF dataset, which will infer every part as visible and localize them.

Secondly, we model occlusion by combining the cue from absence of evidence for body part and local image measurement around the occlusion boundary, which is encoded in the IDOD terms. So we perform the second diagnostic experiment by removing the IDOD terms(*i.e.*, in Equation 3, we have $\varphi^d(\gamma_{ij} = 1|.; \boldsymbol{\theta}) = 0$). In this case, the model handles occlusion only by exploiting the cue from absence of evidence for body part.

Figure 5: Results on the WAF dataset. We show the parts that are inferred as visible, and thus have estimated configurations, by our model.

Method	AOP	Torso	Head	U.arms	L.arms	mPCP
Base Model [?]	73.9	81.4	92.6	63.6	47.6	66.1
FC	82.0	87.0	**98.6**	72.7	67.5	77.7
FC+IDOD	**84.9**	**88.5**	98.5	**77.2**	**71.3**	**80.7**

Table 2: Diagnostic Experiments PCP and AOP results on the WAF dataset. Using flexible compositions (*i.e.*, *FC*) significantly improves our base model [?] by 11.6% on PCP and 8.1% on AOP. Adding *IDOD* terms (*FC+IDODs*, *i.e.*, the full model) further improves our PCP performance to 80.7% and AOP performance to 84.9%, which is significantly higher than the state of the art methods.

We show the PCP and AOP performance of the diagnostic experiments in Table 2. As is shown, flexible compositions (*i.e.*, *FC*) significantly outperform a single composition (*i.e.*, the base model [?]), and adding *IDOD* terms improves the performance significantly (see the caption for details).

7. Conclusion

This paper develops a new graphical model for parsing people. We introduce and experimentally verify on the WAF dataset (see Section 6.4) a novel prior that the visible body parts of human tend to form a connected subtree, which we define as a flexible composition, even with the presence of significant occlusion. This is equivalent to modeling people as a mixture of flexible compositions. We define novel occlusion cues and learn them from data. We show very efficient inference can be done for our model by exploiting part sharing so that computing over all the flexible compositions takes only twice that of the base model [?]. We evaluate on the WAF dataset and show we significantly outperform current state of the art methods [?, ?]. We also show big improvement over our base model, which does not model occlusion explicitly.

8. Acknowledgements

References

[1] H. Azizpour and I. Laptev. Object detection using strongly-supervised deformable part models. In *European Conference on Computer Vision (ECCV)*, 2012.

[2] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *International Conference on Computer Vision (ICCV)*, 2009.

[3] X. P. Burgos-Artizzu, P. Perona, and P. Dollár. Robust face landmark estimation under occlusion. In *International Conference on Computer Vision (ICCV)*, 2013.

[4] X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, and A. Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *Computer Vision and Pattern Recognition (CVPR)*, 2014.

[5] X. Chen and A. Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[6] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 1995.

[7] J. Coughlan, A. Yuille, C. English, and D. Snow. Efficient deformable template detection and localization without user initialization. *Computer Vision and Image Understanding*, 2000.

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition (CVPR)*, 2005.

[9] M. Eichner and V. Ferrari. We are family: Joint pose estimation of multiple persons. In *European Conference on Computer Vision (ECCV)*, 2010.

[10] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2010.

[11] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision (IJCV)*, 2005.

[12] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *Computer Vision and Pattern Recognition (CVPR)*, 2008.

[13] M. A. Fischler and R. A. Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on Computers*, 1973.

[14] G. Ghiasi and C. C. Fowlkes. Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model. In *Computer Vision and Pattern Recognition (CVPR)*, 2014.

[15] G. Ghiasi, Y. Yang, D. Ramanan, and C. C. Fowlkes. Parsing occluded people. In *Computer Vision and Pattern Recognition (CVPR)*, 2014.

[16] R. B. Girshick, P. F. Felzenszwalb, and D. A. Mcallester. Object detection with grammar models. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

[17] Y. Jia. Caffe: An open source convolutional architecture for fast feature embedding. http://caffe.berkeleyvision.org/, 2013.

[18] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *British Machine Vision Conference (BMVC)*, 2010.

[19] G. Kanizsa. *Organization in vision: Essays on Gestalt perception*. Praeger New York, 1979.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[21] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Poselet conditioned pictorial structures. In *Computer Vision and Pattern Recognition (CVPR)*, 2013.

[22] B. Sapp, C. Jordan, and B. Taskar. Adaptive pose priors for pictorial structures. In *Computer Vision and Pattern Recognition (CVPR)*, 2010.

[23] B. Sapp and B. Taskar. Modec: Multimodal decomposable models for human pose estimation. In *Computer Vision and Pattern Recognition (CVPR)*, 2013.

[24] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2014.

[25] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[26] Z. Tu and S.-C. Zhu. Parsing images into regions, curves, and curve groups. *International Journal of Computer Vision*, 2006.

[27] C. Wang, Y. Wang, and A. L. Yuille. An approach to pose-based action recognition. In *Computer Vision and Pattern Recognition (CVPR)*, 2013.

[28] Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2013.

[29] A. Yao, J. Gall, and L. Van Gool. Coupled action recognition and pose estimation from multiple views. *International journal of computer vision*, 2012.

[30] L. Zhu, Y. Chen, Y. Lu, C. Lin, and A. Yuille. Max margin and/or graph learning for parsing the human body. In *Computer Vision and Pattern Recognition (CVPR)*, 2008.

[31] L. Zhu, Y. Chen, A. Torralba, W. Freeman, and A. Yuille. Part and appearance sharing: Recursive compositional models for multi-view. In *Computer Vision and Pattern Recognition (CVPR)*, 2010.

[32] S.-C. Zhu and D. Mumford. A stochastic grammar of images. *Foundations and Trends in Computer Graphics and Vision*, 2006.