# Parsing Occluded People

Golnaz Ghiasi Yi Yang Deva Ramanan Charless C. Fowlkes
Dept. of Computer Science, University of California, Irvine
{gghiasi,yyang8,dramanan,fowlkes}@ics.uci.edu

## Abstract

*Occlusion poses a significant difficulty for object recognition due to the combinatorial diversity of possible occlusion patterns. We take a strongly supervised, non-parametric approach to modeling occlusion by learning deformable models with many local part mixture templates using large quantities of synthetically generated training data. This allows the model to learn the appearance of different occlusion patterns including figure-ground cues such as the shapes of occluding contours as well as the co-occurrence statistics of occlusion between neighboring parts. The underlying part mixture-structure also allows the model to capture coherence of object support masks between neighboring parts and make compelling predictions of figure-ground-occluder segmentations. We test the resulting model on human pose estimation under heavy occlusion and find it produces improved localization accuracy.*

## 1. Introduction

Occlusion poses a significant barrier to good recognition performance in complex cluttered scenes such as that shown in Fig. 1. Even when the type of occluder is known (e.g., other people in a crowded street) the relative layout of occluder and object is unconstrained resulting in a huge variety of possible appearances for a partially occluded object. Many approaches to detection and pose estimation treat occluders as outliers and simply ignore image evidence in hypothesized occluded regions. However, such an approach easily confuses occlusion with features that are simply hard to detect due to unusual appearance or weak discriminability.

Our goal is to develop appearance models that explain image features generated by occlusion rather than ignoring them, coupling pose estimation to segmentation. Figure-ground cues such as the presence and shape of occluding contours as well as prototypical appearances corresponding to self-occlusion serve as positive evidence for an occlusion event. To achieve this, we utilize part models in which local appearances are represented by a large library of discriminatively trained templates and their associated segmentations. Our system predicts the presence and pose of the object as well as detailed segmentation masks that contain figure, background, and occluder labels.

Figure 1. The image above depicts a scene where low-level feature descriptors are dominated by occlusions. We aim to model such appearances by training models with large numbers of local mixtures that capture these occlusion statistics, yielding improvements for the task of pose estimation and visibility prediction.

Unfortunately, full joint training of such high-dimensional models requires large amounts of hand-segmented training data that are representative of the huge variety of possible occlusion patterns. Since such training data is not readily available, we approach this difficulty through the extensive use of synthetically generated data. We generate tens of thousands of images of partially occluded objects which are then used to train deformable human templates. Each such generated example comes with a complete annotation of both the object and a segmentation of the occluder. We use occluder segmentation masks as a supervisory signal to group (cluster) the space of possible occlusions and infer occlusion by enumeration over such groups. Since we can generate an essentially infinite supply of training data, our ability to model the "long-tail" of rare occlusion patterns is only limited by computation.

Inference and learning in our model is based on now-standard approaches to discriminative training of pictorial structures. In particular, training our model is quite similar to flexible part model [26, 5] which also uses local mixtures.

However, to capture occlusion states requires a model with an order of magnitude more parameters and trained with several orders of magnitude more training data. This poses significant computational burden during learning and inference – using the off-the-shelf code of [26] would require 2 weeks of computation. We present several improvements to training that make such learning feasible.

Finally to demonstrate the value of modeling occlusion, we carry out an analysis of the effect of occlusion on model performance using images from the H3D [2] and "We are Family" datasets [7]. In addition to evaluating joint localization accuracy, we also evaluate occlusion/visibility prediction. We show that by modeling the appearance of occlusion, the model achieves improved accuracy over existing pose estimation techniques.

## 2. Related work

We posit that a shortfall of many proposed occlusion models for detection is that they don't model the visual appearance of occlusion. Instead the occluded portions of the object are described with the same model used for all background/non-object pixels. Algorithmically this means that a part is assumed occluded if it scores lower than some learned threshold. If this threshold is too high, unoccluded objects are predicted as being occluded. If this threshold is too low, occluded objects are easily confused with background. Instead we argue that occlusion should only be hypothesized if there is image evidence to support it.

**Occlusion Modeling:** One popular approach is to treat visibility as a binary variable that is inferred at test-time. Modeling part-level occlusion is a natural fit for models with an explicit representation of part detection. For example, the generative constellation models of Weber et. al. [23] and Fergus et. al. [10] exhaustively enumerated and scored all possible occlusion hypotheses. The supervised part models described in [1] includes templates for an occluded version of each part in the model but imposes no pairwise constraints on visibility of different parts in the model. The grammar-based model described by [12] also includes explicit occlusion part templates but enforces more structure in the pattern of occlusion, specifying a person detector that includes a variable number of parts arrayed in a vertical chain followed by an occluder part. While this grammar could be implemented with a local mixture model formally equivalent to our approach, the grammar provides an elegant and compact description of parameter sharing within the model. A major difference with our model is that [12] requires that allowed patterns of occlusion be specifically designed into the grammar. The idea we describe here sidesteps this structure learning problem, automatically learning valid occlusion patterns from data in a non-parametric way.

One drawback of part-level occlusion is that it doesn't capture the fine-scale pattern of occlusion within a part. An alternate family of techniques apply occlusion reasoning at the level of image feature maps or individual pixels that make up a template [21, 11, 22]. Spatial coherence is often enforced in such models by a Markov random field aligned with the pixel grid. However, in natural scenes the spatial statistics of occlusion patterns are not translation invariant and depend on the environment and imaging geometry (see e.g., the results in [13]). Our modeling approach describes occlusion within each part template but enforces consistency at the level of parts rather than pixels or HOG cells which allows the dependence structure between occlusions to adapt to articulated shapes. Our model thus implicitly learns the the spatial statistics of occlusion but with the benefit of a tree-structured distribution which makes hypothesis enumeration computationally efficient.

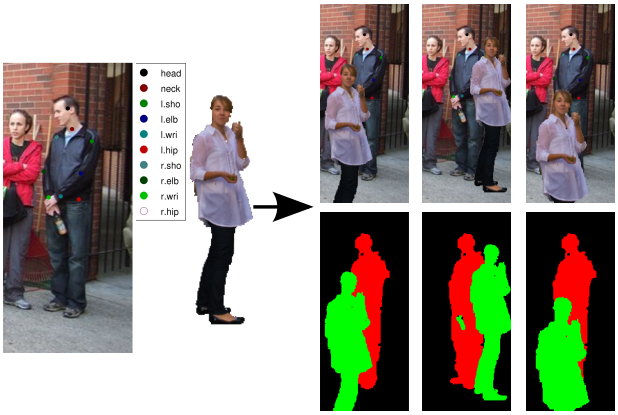



Figure 2. We synthesize a large corpus of training data by compositing segmented objects at random locations over a base training image. The position and scale of the occluder is tied to the occluded object by a weak ground-plane assumption. When combined with keypoint annotations and segmentations from the original dataset, this yields a limitless supply of strongly supervised training data that includes keypoint visibility and occlusion segmentation masks.

**Image Parsing:** An appealing alternative is to move from single object detection to whole-image parsing. The presence of occlusion can then be "explained away" by the presence of an occluding object. For example, [25] describe a layered segmentation model for reasoning about occlusions between detected objects at the pixel level. [11] enforce mutual exclusion in the assignment of HOG cells while [3] use competitive smoothing between shape masks associated with detectors. Our work is also closely inspired by a family of approaches that build templates for detecting the complicated appearances associated with typical object-object interactions. This includes multi-person [24, 7, 20, 16] or other multi-object [18, 5] models that implicitly capture occlusion interactions between objects.

An inherent difficulty with image parsing approaches is

that they require detecting the occluding object. In a real world setting where the occluder could be arbitrary, this involves training and scoring a huge bank of object detectors on every image. Learning explicit multi-object "visual phrase" detectors may be feasible for some small set of human-object interactions, but it seems unlikely to scale to occlusion where interactions are far less constrained. *There are a limited number of ways one can reasonably ride a horse but many ways to hide a horse.* The complexity of our modeling approach lies between that of single object models and multi-object models. We train a detector for a single object category which operates independently of other detections in the scene. However, we model the appearance of occluders in a generic manner, relying on a large corpus of synthetic training data to capture the generic statistics of occlusion appearance.

**Synthetic training data:** Several papers have explored the use of synthetic data in training systems for recognition and pose estimation. In [19] the authors use a large set of synthetically rendered poses spanning the space of articulations in order to perform nearest-neighbor (pose) regression. [15] use green-screening to augment training data with synthetic renderings of real objects on cluttered backgrounds and [14] generated a 3-million frame dataset of synthetic images of articulated models in real backgrounds. Our work differs in using an "image-based rendering" approach, cutting and pasting existing images to yield novel ones. This is most related to [17], who fit 3D articulated models to real images, and generate synthetic renderings by slightly perturbing joint angles.

## 3. Modeling Local Occlusion Patterns

We model the appearance of occluded people by a pictorial structure with local mixtures, similar to the flexible part model of [26]. In this section we describe how the local mixture labels for each part are derived. In the next section we describe how the appearance templates are learned and combined into a joint model.

**Generating synthetic images:** We generate a large corpus of synthetic occlusion data by compositing segmented objects over a base training data set that has been annotated with part locations and figure-ground masks. This process automatically produces examples of occluded appearance along with supervisory information including the pixel-level support of the occluding object. In our experiments we use the H3D dataset [2] which provides segmentation masks as well as joint locations for $\sim$1500 people. We scale occluders based on object annotations in the base image to produce realistic spatial distributions (e.g., people's heads are unlikely to be occluded by others feet). The bottom of the occluder is placed below the base object and scaled linearly as a function of relative y-offset. Fig. 2 shows examples of such synthetic training images.

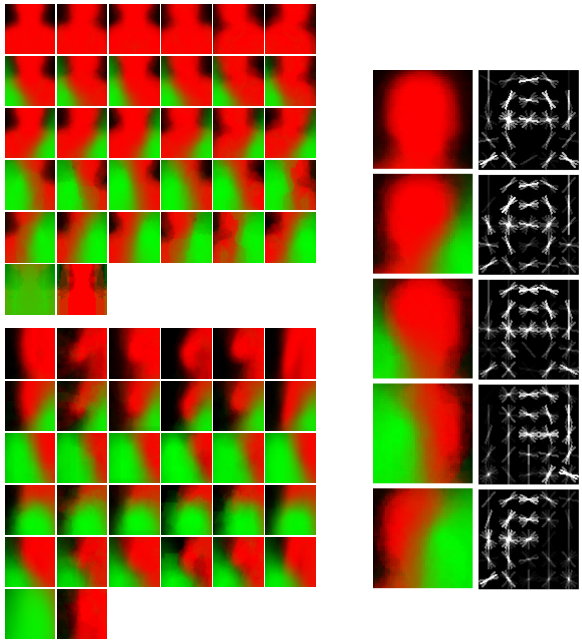
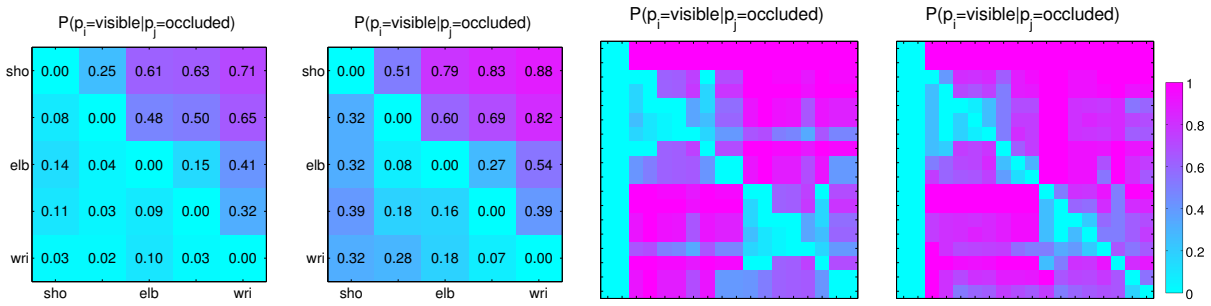


Figure 3. We cluster part appearances using a factored model that independently captures variation in pose (clustered on keypoint location) and occlusion patterns (clustering based on segmentation). Our model also includes separate fully-occluded and self-occluded components. Example factored clusterings for neck and elbow are shown on the left. Each row corresponds to a separate occlusion pattern and each column a separate pose. Colors show the average segmentation masks associated with each cluster. The right shows visualizations of part templates associated with different head occlusions. Relative to the unoccluded head, the partial occlusion templates include more edges oriented along the occluding contour, aiding detection.

**Learning part appearances:** We exploit this highly-annotated synthetic training data to find clusters of training examples that capture the appearance of each part under different pose and partial occlusion conditions. Generating a large number of quality clusters is a surprisingly hard problem; typical approaches of clustering image patches [6], clustering keypoint annotation [2], or even manual grouping [27] have shown only modest performance increases with increasing numbers of clusters. Some of these difficulties are due to insufficient data (given a finite dataset, the amount of training data per cluster decreases with more clusters) and poor metrics for clustering. We found simple appearance-based clustering performs poorly since the space of occlusion patterns is high-dimensional due to the arbitrary placement and appearance of the occluder.

Synthetic training data addresses these difficulties in two ways. First, synthetic training data generation allows us to increases the amount of training data per cluster. Second, synthetic data comes with supervisory information in the form of occluder-object-background segmentation masks which can provide stronger metrics for clustering.

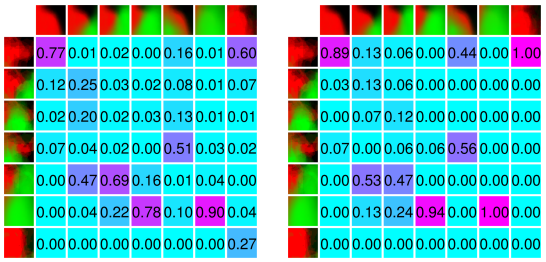



(a) training examples (b) test predictions (c) training examples (d) test predictions

Figure 4. Here we show co-occurrence structure between the occlusion state (visibility) of each part. The $j$th column contains the probability that a part $i$ is visible conditioned on $j$ being occluded. Panels (a) and (b) show the co-occurrence for a simple 5-part, 32-mixture model trained to localize arms. (a) gives the ground-truth visibility statistics for test data while (b) shows the statistics of the labels produced from running the model on test data. The discriminative SVM training produces a fairly good quantitative match with a slight bias towards increased visibility. Panels (c) and (d) show similar statistics for the whole upper-body model. The prominent block-structure corresponds to the head, left and right halves of the body respectively. Within each limb, occlusion at the top of the arm (e.g. shoulder) makes visibility of the lower arm unlikely while occlusion of the wrist does not strongly constrain visibility of the upper arm.

**Factored occlusion-pose clustering:** We separately cluster training patches for each keypoint. We label each training patch $i$ using both a geometric pose feature $g_i$ and a figure-background-occluder segmentation $o_i$. The pose feature vector describes the spatial offset of a part relative to its neighbors in the pictorial structure. The segmentation $o_i$ is a collection of three binary masks in a window surrounding the keypoint that indicate the local segmentation of the object (as in Fig. 2). A naive approach is to apply a standard clustering algorithm (e.g., K-means) on concatenated descriptors with some relative weighting between the two types of features. However, our training set is severely biased due to our synthetic training data, all of which contain significant occlusions. We do not want this to adversely affect our grouping. Furthermore, from a generative perspective, we expect that the pattern of occlusion and the object pose are largely independent (one very important exception being self-occlusion). For this reason, we use a factored clustering algorithm. We generate one clustering using geometric pose features into $K_g$ clusters with K-means. By construction, these clusters are not affected by the amount of synthetic training data. We also generate a second independent clustering of the occluder masks into $K_o$ clusters. We finally assign each training example to an element of the "cross-product" space of $K_g \times K_o$ clusters, or to fully or self-occluded mixtures.

Fig. 3 shows an example of such clusterings for several different object parts. Each row corresponds to different occlusion clusters while each column corresponds to pose clusters. We also include two additional clusters, a fully-occluded cluster and a self-occluded cluster. The appearance of a fully-occluded part is assumed to be independent of the pose since it only includes image features arising from the occluder. An example is assigned to the self-occluded cluster when the part is invisible in the image




(a) training examples (b) test predictions

Figure 5. Conditional probabilities for occlusion states of the elbow given the shoulder in (a) the ground-truth synthetic training data and (b) mixtures selected by the model on test data. The model learns coherency of the segmentation. E.g., if the left side of the shoulder is occluded (5th column) then the elbow tends to be visible (1st row) or also occluded on the left (4th row). The 7th row corresponds to self-occlusion.

even though there is no occluder present (e.g., hands clasped behind the head). Self-occlusion could presumably benefit from multiple pose clusters but this requires sufficient training data and our synthesis approach cannot automatically generate such self-occlusions

**Cluster statistics:** Occlusions of parts are not independent. If a person's elbow is occluded by the elbow of another person, then shoulder may also be occluded (by that occluders' shoulder). This implies that cluster labels across neighboring joints may have very specific co-occurrence statistics. We visualize examples of such statistics in Fig. 4 and Fig. 5.

## 4. Occlusion-aware Part Models

We now describe a method for training deformable pictorial structures with tens of thousands of images of human poses (under heavy occlusion).

**Deformable Templates:** Our model consists of a set of parts $V$ and pairwise relations $E$ which encode joint constraints on part locations and appearances. Let $I$ be an image, $p_i = (x, y)$ be the location for part $i \in V$ and $m_i$ be the mixture component of part $i$. Each local mixture $m_i$ corresponds to a occlusion-pose cluster learned for that part from the synthetic training data. If part $i$ corresponds to the left elbow, and $m_i$ selects the coarse elbow orientation along with a particular local pattern of occlusion (including full and self-occlusion). Each choice of $m_i$ is associated with an average figure-ground-occluder mask for the cluster which can be used to predict keypoint visibility and segmentation at test time.

Given an image, we score a collection of hypothesized part locations and local mixture selections with the following objective:

$$S(I, p, m) = \sum_{i \in V} \left[ \alpha_i^{m_i} \cdot \phi(I, p_i) \right] + \sum_{ij \in E} \left[ \beta_{ij}^{m_i, m_j} \cdot \psi(p_i - p_j) + \gamma_{ij}^{m_i, m_j} \right] \quad (1)$$

The first term scores the appearance evidence for placing a template $\alpha_i^{m_i}$ for part $i$, tuned for mixture $m_i$, at location $p_i$. We write $\phi(I, p_i)$ for the feature vector (e.g., HOG descriptor [4]) extracted from pixel location $p_i$ in image $I$. Note that we define a separate template for each mixture, even occluded states as such templates will capture visual features associated with occlusions.

The second term scores relational constraints between pairs of parts. The feature $\psi(p_i - p_j) = \begin{bmatrix} dx & dx^2 & dy & dy^2 \end{bmatrix}$ is a vector of relative offsets between part $i$ and part $j$ and the parameters $\beta_{ij}^{m_i, m_j}$ specify the relative rest location and quadratic spring penalty for deviating from that rest location. Both the spring and the bias, $\gamma_{ij}^{m_i, m_j}$, depend on the local mixtures $m_i$ and $m_j$ selected for parts $i$ and $j$. This allows the relational model to capture dependencies between visibility of neighboring parts in the model (as in Fig. 4) as well as providing a much richer, non-parametric description of pose and appearance than is possible with a single local template and spring.

**Learning and Inference:** Given a test image, we seek the maximum scoring part arrangement $p$ and mixture assignments $m$. When $E$ is tree-structured, this solution can be computed efficiently with dynamic programming [9, 26].

Let $(p^n, m^n)$ be the ground-truth part locations and mixture labels provided for the $n$th positive training example. We learn model parameters $w = (\alpha, \beta, \gamma)$ using a variant on the structured SVM.

$$\begin{aligned} \underset{w, \xi_i \geq 0}{\text{argmin}} \quad & \frac{1}{2}||w||^2 + C \sum_n \xi_n \quad (2) \\ \text{s.t.} \quad \forall n \in \text{pos} \quad & w \cdot \Phi(I^n, p^n, m^n) \geq 1 - \xi_n \\ \forall n \in \text{neg}, \forall p, m \quad & w \cdot \Phi(I^n, p, m) \leq -1 + \xi_n \end{aligned}$$

The above quadratic program (QP) attempts to learn a low-norm $w$ that scores positive examples above 1 (evaluated at ground-truth part locations and mixture labels) and scores negative examples below $-1$ (for any setting of part locations and mixtures). We use a standard cutting-plane approach to incrementally add negative constraints by running the detector on negative training images in order to find a subset of constraints that are active at the optimum.

**Mislabeled positives:** In order to improve localization and occlusion prediction, we also added incorrectly-labeled positive images as negative examples. To do so, we include the following negative constraints to our QP:

$$\forall p, n \in \text{pos}, m \nsim m^n \quad w \cdot \Phi(I^n, p, m) \leq -1 + \xi_n \quad (3)$$

We say that two sets of part-mixture assignments $m$ and $m'$ are in the same equivalence class iff they predict the same set of parts are visible. We use the notation $m \nsim m^n$ to mean that the mixture assignments $m$ are in a different equivalence class than the ground-truth. This constraint thus enforces that for a given positive example $n$, all poses associated with *incorrect* mixture assignments corresponding to an incorrect occlusion prediction score below $-1$. These constraints are similar to those found in traditional structured prediction (where the true label should outscore incorrect labels by a margin), but split into positive and negative constraints. We find that this splitting speeds up optimization without sacrificing performance.

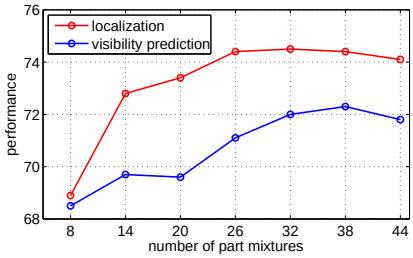
**Semi-latent learning:** Although the part locations and mixture labels are given in our positive training data, we found it was useful to perform re-estimation of part locations. This is particularly important for occluded parts where the discriminative appearance features being learned (the presence of an occluding contour) is more dependent on the position of the occluder than on the the part keypoint location.

We used a standard latent learning approach [8] to alternately train a model using convex optimization and then re-estimate the locations $p^n$ for the set of parts. For each positive example $n$, let $\Omega_n = \{p : |p_i - p_i^n| < r\}$ denote a set of possible part locations that lie near the ground-truth location $p^n$. We learn a model $w$ and update part locations $p^n$ with a coordinate descent algorithm:

1. *Model update:* Learn $w$ with a QP (2) using the inferred positive part locations $p^n$.
2. *Latent assignment:* Compute $p^n = \max_{p \in \Omega_n} w \cdot \phi(I^n, p, m^n)$ for $n \in$ pos.

Note that during this learning we do not update the mixture assignments, instead relying on the reliable ground-truth clustering.

**Computational bottleneck:** Typically, the computational bottleneck of latent SVM training[8] is Step 1, which




(a)

| | H3D | | H3D Occluded | | H3D Synthetic | |
|---|---|---|---|---|---|---|
| | pck | ocl | pck | ocl | pck | ocl |
| FMP6 [26] | 71.5 | 80.1 | 55.6 | 64.2 | 50.4 | 57.8 |
| FMP6.1+syn | 70.5 | 79.4 | 60.6 | 69.0 | 56.2 | 62.5 |
| OMP32 | 71.5 | 78.4 | 55.5 | 62.0 | 59.0 | 67.4 |
| OMP32+syn | 68.7 | 72.6 | 68.5 | 72.0 | 70.4 | 73.5 |
| OMP32+syn+struct | 70.0 | 74.4 | 68.5 | 72.8 | 71.1 | 74.5 |

(b)

Figure 6. (a) Localization and visibility prediction accuracy on synthetically occluded test data as a function of the number of mixture model components. We found that performance saturated above 32 mixtures. (b) Evaluation of performance on H3D test images, a subset containing heavy occlusion and a set of synthetically occluded examples. The benefits of modeling occlusion are more pronounced on the synthetic and heavily occluded subsets. FMP6.1 is a baseline model with a single mixture representing occlusion so it can exploit synthetic training data.

requires passing over a large set of negative images (typically on the order of a thousand) and performing "hard negative" mining. In our case, the computational bottleneck is Step 2, since we now have *hundreds of thousands* of positive examples in our synthetic dataset. A standard approach for latent updating of positive examples is to evaluate the model $w$ as a "detector" on each positive example. A simple but crucial observation is that in the semi-latent setting the mixture label $m^n$ is known and *not* latently updated, so only a *single* filter per part need be evaluated. Modifying the released code of [26] to allow for this efficient semi-latent update produced an order-of-magnitude speed up, reducing training time from over a week to under a day. We also note the latent assignment can be trivially parallelized.

## 5. Experimental results

**Dataset:** We use H3D as our primary source of training and testing data. Since H3D contains many challenging poses with different points of view and our baseline model [26] still struggles with non-frontal poses, we selected a subset of 668 images with frontal facing people. For our synthetic training experiments, each original training image was augmented with 100 different synthetic occlusions, yielding a training set of half a million positive images. We use negative training images from the INRIAPerson database [4] and evaluate models using 190 test images from H3D. Additionally, we also evaluate our model on the "We Are Family" dataset (WAF) and benchmark provided by [7].

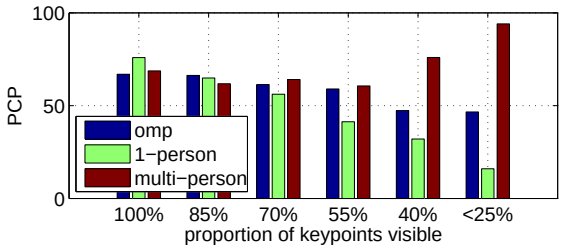
To better understand model performance, we considered variants of each test dataset which were enriched for occlusion. We evaluated on 3 different variants of the H3D dataset. The original 190 test images, a subset of 60 which were selected as heavily occluded (many invisible ground-truth keypoints), and an a synthetically occluded version of the 190 original test examples. For WAF, we considered six subsets of data based on the proportion of visible keypoints. In the WAF dataset, there are 6 parts or "sticks" (head, torso, upper arms and right arms) labeled in the data set, each with a visibility variable. We build six subsets according to the number of total visible sticks.

**Evaluation:** For evaluation on H3D we use the percentage of correctly localized keypoints (PCK) criteria used by [26]. A predicted keypoint is considered as correctly localized when it lies within a scale-normalized threshold distance (half the head height) of the ground-truth keypoint location. To extend the PCK criteria for occluded body parts, we require that any keypoint marked invisible in the ground-truth must correctly be predicted as occluded by the model. To separate out errors in localization from errors in visibility prediction, we also compute the accuracy of part visibility prediction as a binary classification task.

For WAF we used the percentage of correctly localized parts (PCP) criteria of [7] which is similar to PCK but measures the localization of the sticks rather than their endpoints. Similar to H3D we also measure stick occlusion prediction based on stick visibility marked in the ground-truth. We use the set of upper-body detections provided with the WAF dataset which is based on a combined face and body detector and achieves an 86% detection rate.

**Model Complexity:** In order to choose number of mixtures per part, we evaluated performance while varying $K_o$ to generate models with up to 44 mixtures per part. Fig. 6 shows test localization and visibility prediction accuracy as a function of the number of part mixtures. We chose $K_o = 5$ which yielded 32 mixture components and offered a good trade-off between performance and running time. Increasing the amount of synthetic training data did not seem to change the saturation point. In fact using $10x$ data gave similar performance to $100x$ data in many cases. This suggests that we may need more variety in the shapes of our synthetic occluders in order to usefully grow the number of part mixtures further.

**Synthetic Training Data:** Fig. 6 shows the performance of our 32-mixture model (OMP32) under a variety of training and test conditions. Training the model without syn-




| | WeAreFamily | |
|---|---|---|
| | pcp | ocl |
| FMP6 [26] | 58.0 | 74.5 |
| FMP6.1+syn | 60.4 | 74.2 |
| OMP32+syn | 61.9 | 75.2 |
| OMP32+syn+WAF$_{train}$ | 63.6 | 74.0 |
| 1-Person [7] | 58.6 | 73.9 |
| Multi-Person [7] | 69.4 | 80.0 |

Figure 7. Performance on subsets of the WeAreFamily dataset as a function of the amount of occlusion present. Our model (OMP) achieves a better PCP score than the 1-person model baseline in [7], primarily due to better handling of occluded examples. The much more complicated multi-person model of [7] outperforms our model for heavy occlusion ($< 50\%$ of keypoints visible) but does so at a loss of localization accuracy relative to the 1-person baseline. Table shows overall PCP and occlusion prediction accuracy.

thetic occlusion data (OMP32) yielded better performance on the original H3D data but including synthetic occlusions (OMP32+syn) gave very substantial improvements on the occlusion enriched datasets. Training the model with mislabeled positive examples (OMP32+syn+struct) gave significant improvements in part localization accuracy which boosted performance on the un-enriched original dataset.

**Comparative Evaluation:** Fig. 6 also compares the proposed model (OMP32+syn+struct) to several baselines. FMP6 is based on the code of [26]. We train a baseline version of the FMP on the H3D dataset which always predicts all parts visible at test time. Excluding occluded data during training produced a model which was slightly worse (PCK=69.6). To allow the FMP to predict visibility we also trained a version with the addition of a single occluded mixture component for each part (FMP6.1). This model achieved improved occlusion accuracy since it could predict occlusions at test time. With only 1 occlusion mixture, the addition of synthetic data offered some further improvement in occlusion prediction but at the expense of localization accuracy.

Fig. 7 compares the performance of several models on the "We are Family" dataset including the FMP baseline, the proposed OMP model, as well as a single and multi-person model proposed by [7]. By training the OMP model with synthetic data, we outperform the 1-Person model despite training on a completely different dataset. Including the WAF positive examples in our training set improved OMP performance further. We also examined performance as a function of occlusion level. For all but the most extreme occlusions, our model achieves a similar PCP to the Multi-Person model. This is particularly surprising since the Multi-Person detector performs joint inference over a collection of detector outputs and incorporates other image cues in order to find a depth ordering of detections.

## 6. Conclusions

We have presented a method for modeling occlusion that is aimed at explicitly learning the appearance and statistics of occlusion patterns. Our system produces models which are more robust to heavy occlusion than existing approaches. As an added benefit, our model explicitly represents part occlusions and hence can predict not only part locations but a local segmentation mask. The combination of synthetic training data and flexible models with many part appearance mixtures is in some sense "brute force" and perhaps less elegant that some more parametric approach. However, it has the distinct advantage of being amenable to discriminative learning and, as we have shown, capable of not only learning detailed occlusion statistics but also achieving competitive performance at the task of human pose estimation.

## 7. Acknowedgements

This work was supported by NSF IIS-1253538 and IIS-0954083

## References

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

[2] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *CVPR*, 2009. 2, 3

[3] T. Brox, L. Bourdev, S. Maji, and J. Malik. Object segmentation by alignment of poselet activations to image contours. In *CVPR*, 2011. 2

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005. 5, 6

[5] C. Desai and D. Ramanan. Detecting actions, poses, and objects with relational phraselets. *ECCV*, 2012. 1, 2

[6] S. K. Divvala, A. A. Efros, and M. Hebert. How important are deformable parts in the deformable parts model? In *ECCV Parts and Attributes Workshop*, 2012. 3

[7] M. Eichner and V. Ferrari. We are family: Joint pose estimation of multiple persons. In *ECCV*, pages 228–242. Springer, 2010. 2, 6, 7

[8] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE TPAMI*, 32(9):1627–45, Sept. 2010. 5

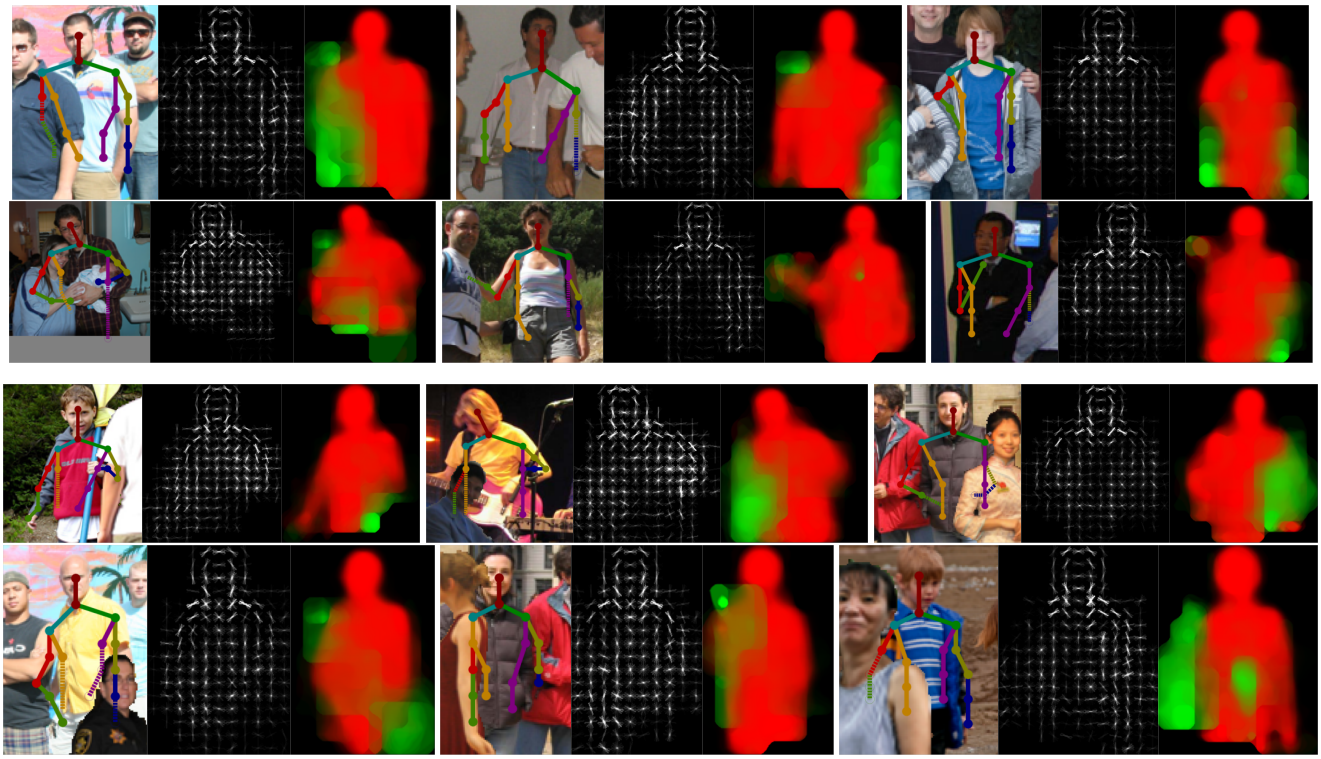


Figure 8. Examples of pose and occlusion estimation for images from the H3D test dataset. Each image shows the keypoint localization (dashed lines indicate occluded), a visualization of the deformed HOG template and a occluder-figure-ground segmentation estimated by compositing the predictions of individual part mixtures. Each of the 18 parts can take on one of 32 mixture (occlusion) states allowing for $32^{18}$ possible occlusion patterns. The top two rows show examples from H3D, the bottom two from H3D Synthetic.


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

[10] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, 2003. 2

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

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

[13] E. Hsiao and M. Hebert. Occlusion reasoning for object detection under arbitrary viewpoint. In *CVPR*, pages 3146–3153, 2012. 2

[14] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. Technical report, Institute of Mathematics of the Romanian Academy and University of Bonn, September 2012. 3

[15] Y. LeCun, F. J. Huang, and L. Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *CVPR*, volume 2, page 97, 2004. 3

[16] W. Ouyang and X. Wang. Single-pedestrian detection aided by multi-pedestrian detection. In *CVPR*, 2013. 2

[17] L. Pishchulin, A. Jain, M. Andriluka, T. Thormahlen, and B. Schiele. Articulated people detection and pose estimation: Reshaping the future. In *CVPR*, pages 3178–3185, 2012. 3

[18] M. A. Sadeghi and A. Farhadi. Recognition using visual phrases. In *CVPR*, pages 1745–1752, 2011. 2

[19] G. Shakhnarovich, P. Viola, and T. Darrell. Fast pose estimation with parameter-sensitive hashing. In *CVPR*, pages 750–757, 2003. 3

[20] S. Tang, M. Andriluka, and B. Schiele. Detection and tracking of occluded people. In *BMVC*, pages 1–11, 2012. 2

[21] A. Vedaldi and A. Zisserman. Structured output regression for detection with partial occulsion. In *NIPS*, 2009. 2

[22] X. Wang, T. X. Han, and S. Yan. An hog-lbp human detector with partial occlusion handling. In *CVPR*, pages 32–39, 2009. 2

[23] M. Weber, M. Welling, and P. Perona. Towards automatic discovery of object categories. In *CVPR*, volume 2, pages 101–108, 2000. 2

[24] Y. Yang, S. Baker, A. Kannan, and D. Ramanan. Recognizing proxemics in personal photos. In *CVPR*, pages 3522–3529, 2012. 2

[25] Y. Yang, S. Hallman, D. Ramanan, and C. C. Fowlkes. Layered object models for image segmentation. *IEEE TPAMI*, 34(9):1731–1743, 2012. 2

[26] Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures-of-parts. *IEEE TPAMI*, 2013. 1, 2, 3, 5, 6, 7

[27] X. Zhu, C. Vondrick, D. Ramanan, and C. Fowlkes. Do we need more training data or better models for object detection? In *BMVC*, 2012. 3