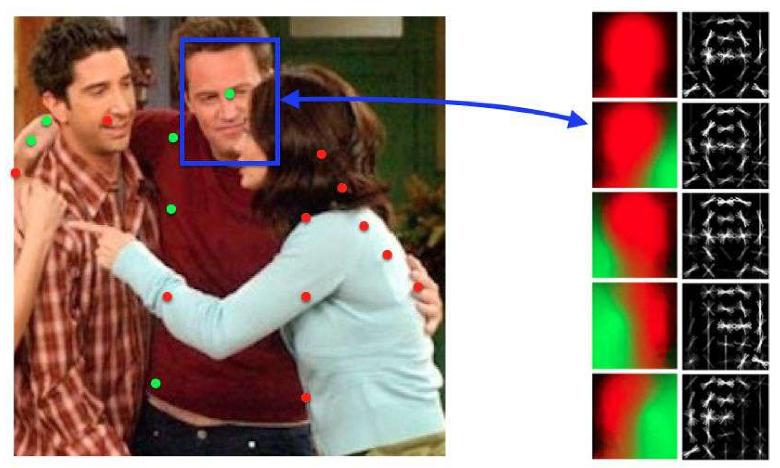
Parsing Occluded People

Golnaz Ghiasi, Yi Yang, Deva Ramanan, Charless Fowlkes

Department of Computer Science, University of California, Irvine

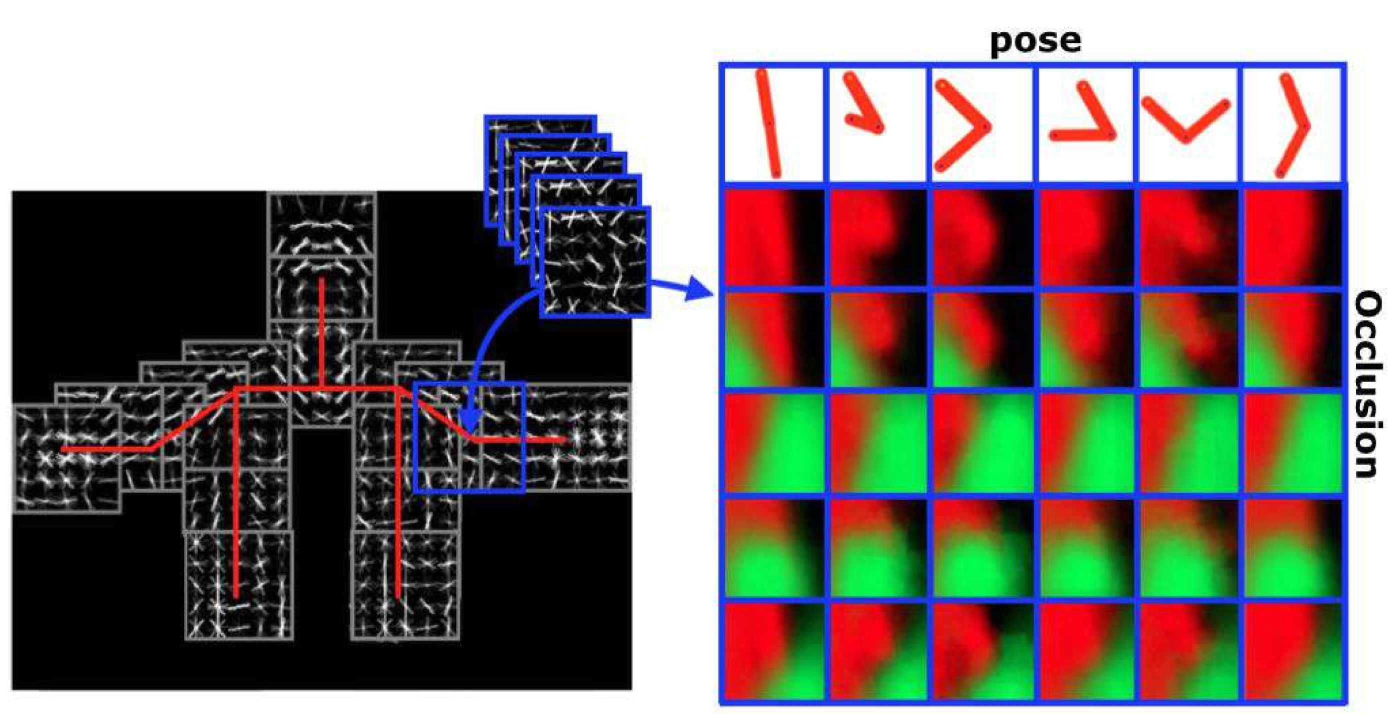
Introduction



Occlusion is common in real world images and poses a significant difficulty for pose estimation. Our goal is to develop appearance models that explain figure-ground cues generated by occlusion such as the presence and shape of occluding contours as well as prototypical appearances corresponding to self-occlusion.

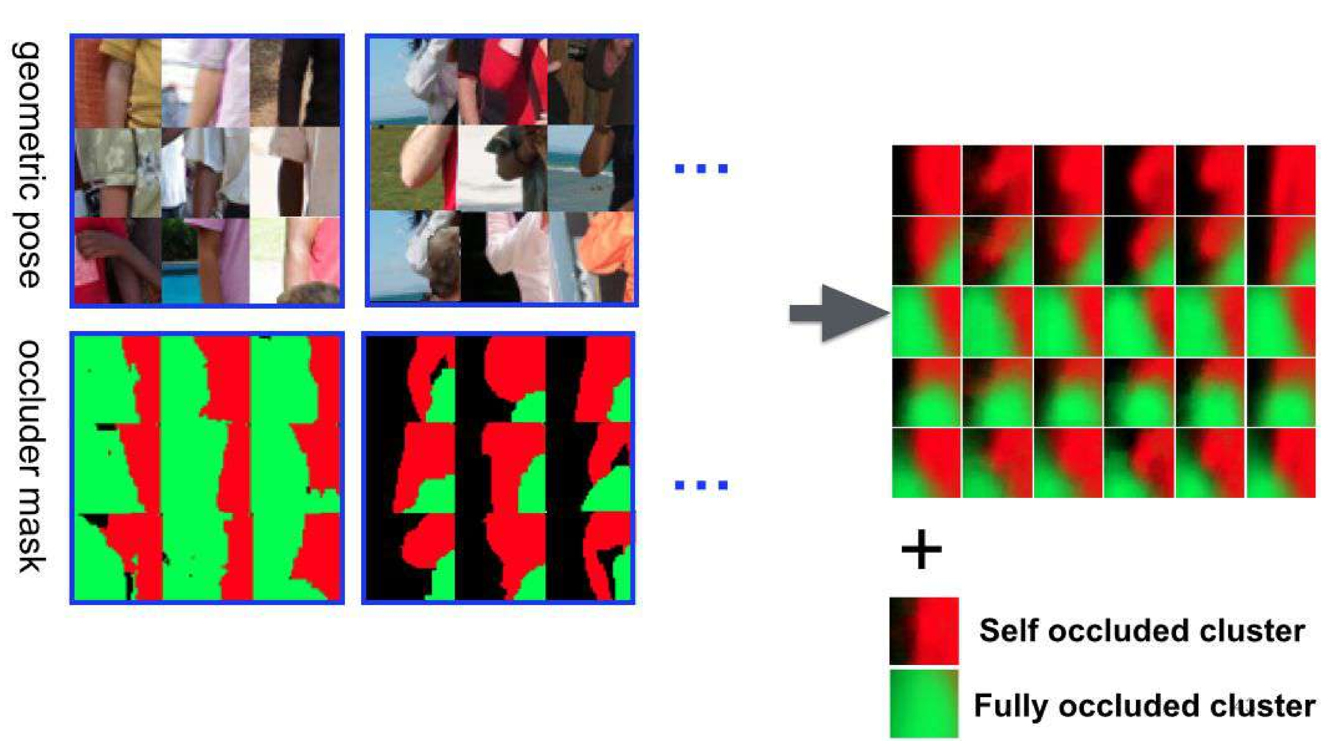
Model

- We model the appearance of occluded people by a pictorial structure with local mixtures, similar to the flexible part model of [1].
- Each local mixture corresponds to an occlusion-pose cluster.
- Each choice of local mixture 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.



Learning Part Mixtures

- We cluster part appearances using a factored occlusion-pose clustering.
- We generate one clustering using geometric pose features into K_g and a second independent clustering of the occluder masks into K_o clusters. Then, we 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.



Scoring Function

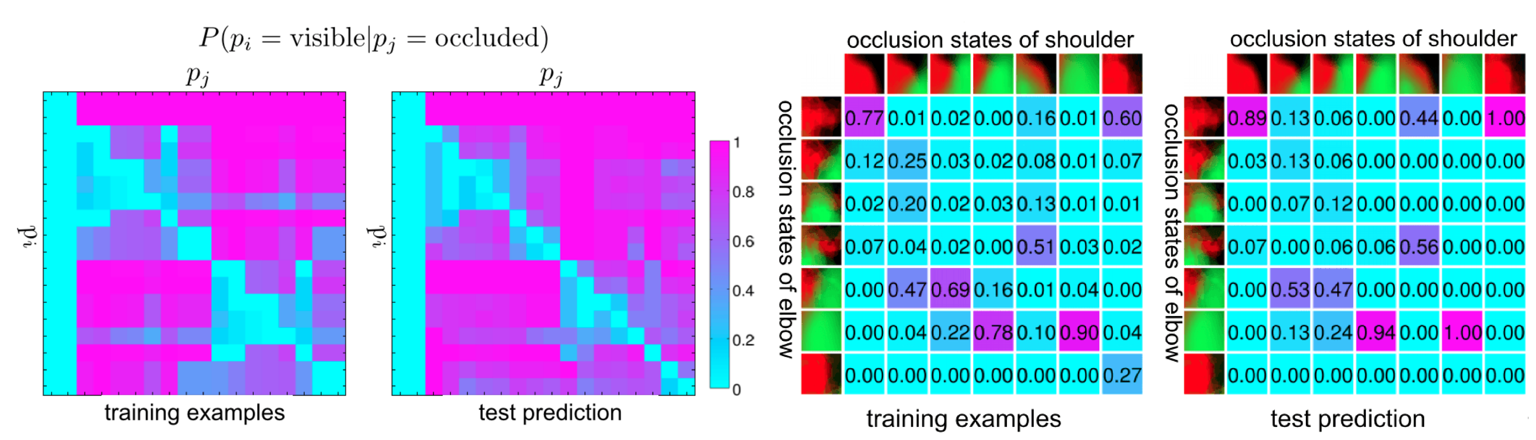
Given an image, we score a collection of hypothesized part locations and local mixture selections with the following objective:

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

- p_i: the pixel location of part i
- m_i: the type (mixture component) of part i
- unary term
  - $\phi(I,p_i)$: local appearance feature extracted from location p_i.
  - $\alpha_i^{m_i}$: local appearance template for shape mixture m_i of part i
- pairwise term
  - $\psi(p_i - p_j)$: spatial feature extracted for the relative location p_i and p_j
  - $\beta_{ij}^{m_i,m_j}$: spatial spring parameter for pair of types (m_i, m_j)
  - $\gamma_{ij}^{m_i,m_j}$: the bias for co-occurrences of pair of parts with types (m_i, m_j)

Cluster Statistics

Occlusions of parts are not independent and cluster labels across neighboring joints may have very specific co-occurrence statistics. Our model learns such statistics.

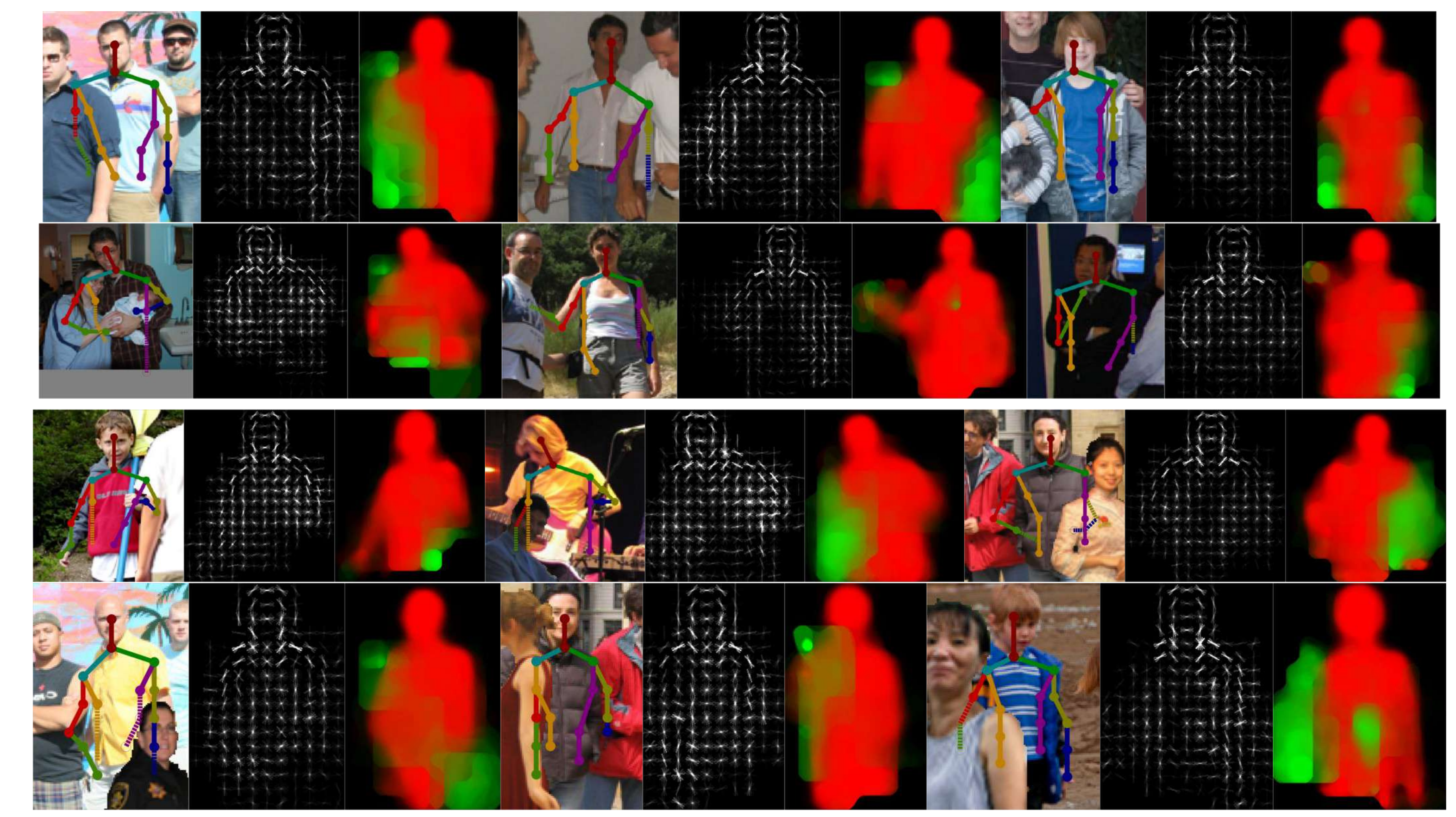


Left: co-occurrence structure between the occlusion state (visibility) of each part. The jth column contains the probability that a part i is visible conditioned on j being occluded. Right: Conditional probabilities for occlusion states of the elbow given the shoulder state.

Synthetic Training Data

- We use a subset of 668 images with frontal facing people from H3D [2] as our primary source of training. This dataset has some occluded examples.
- Training of our model requires large amounts of training data that are representative of the huge variety of possible occlusion patterns.
- Because such training data is not readily available, we generate synthetically occluded data by compositing segmented people over H3D training images.

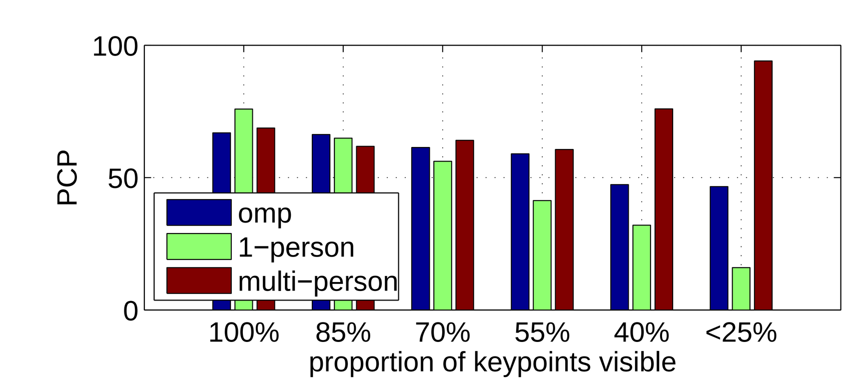
Results on the H3D Dataset



	H3D		H3D Occluded		H3D Synthetic	
	pck	ocl	pck	ocl	pck	ocl
FMP6 [1]	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	70.0	74.4	68.5	72.8	71.1	74.5

- Evaluation of performance on a subset of 190 front view images from H3D (H3D), a subset of 60 images containing heavy occlusion (H3D Occluded) and a set of 190 synthetically occluded images (H3D Synthetic).
- PCK (Percentage of Correctly Localized keypoints): A predicted key point is correctly localized when it lies within half the head height of the ground-truth keypoint
- OCL: accuracy of part visibility prediction as a binary classification task
- FMP6.1 is a baseline model with a single mixture representing occlusion so it can exploit synthetic training data.

Results on the We Are Family Dataset



	WeAreFamily	
	pcp	ocl
FMP6 [1]	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 [3]	58.6	73.9
Multi-Person [3]	69.4	80.0

- Performance on subsets of the WeAreFamily [3] dataset as a function of the amount of occlusion present (left). Right table shows overall PCP and occlusion prediction accuracy.
- Our model (OMP) achieves a better PCP score than the 1-person model baseline in [3].
- For all but the most extreme occlusions, our model achieves a similar PCP to the Multi-Person model.

References

[1] Yang, Y., Ramanan, D.: Articulated human detection with flexible mixtures-of-parts. IEEE TPAMI (2013)

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

[3] Eichner, M., Ferrari, V.: We are family: Joint pose estimation of multiple persons. In: ECCV. Springer (2010) 228–242