Parsing Occluded People by Flexible Compositions

Xianjie Chen, Alan Yuille
University of California, Los Angeles.

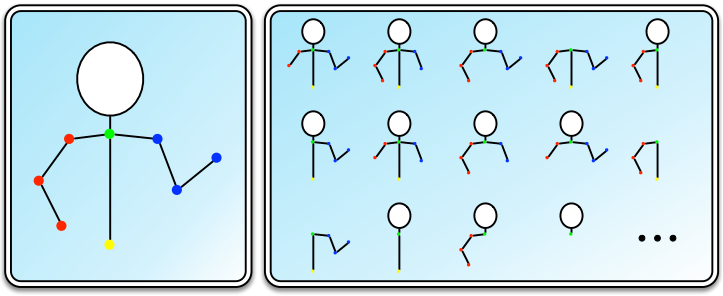

Figure 1: An illustration of the *flexible compositions*. Each connected subtree of the *full graph* (include the full graph itself) is a flexible composition. The flexible compositions that do not have certain parts are suitable for the people with those parts occluded.

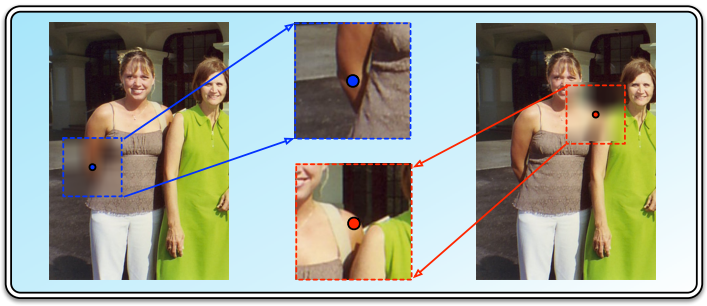

Figure 2: The absence of body parts evidence can help to predict occlusion. However, absence of evidence is not evidence of absence. It can fail in some challenging scenes. The local image measurements near the occlusion boundary (*i.e.*, around the right elbow and left shoulder) can reliably provide evidence of occlusion.

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 [1, 6] and exploit the *connectivity 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 [2] and obtain significant performance improvements over the best alternative algorithms.

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 [1, 6]. 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 [1] 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). In the terminology of graphical models, the visible (non-occluded) nodes form a connected subtree of the full graphical model (following current models, for simplicity, we assume that the graphical model is treelike). This connectivity prior is not always valid (*i.e.*, the visible parts of humans may form two or more connected subsets), but our analysis 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 [1], which is the state of the art on several benchmarked datasets [3, 4, 5], but is not designed for dealing with significant occlusion. We explicitly model occlusions using the connectivity prior above. This means that we have a mixture of models where the number of components equals the number of *all* the possible 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. 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 local cues such as T-junctions which can indicate local occlusions. But although these occlusion cues have been used by some models, 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 [2] for parsing humans when significant occlusion is present. We show strong performance with significant improvement over the best existing method [2] and also outperform our base model [1]. We perform diagnostic experiments to verify our connectivity 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.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.

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

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

[3] Vittorio Ferrari, Manual Marin-Jimenez, and Andrew Zisserman. Progressive search space reduction for human pose estimation. In *Computer Vision and Pattern Recognition (CVPR)*, 2008.

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

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

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