A Multi-layer Composite Model for Human Pose Estimation

Kun Duan[1]
kduan@indiana.edu
Dhruv Batra[2]
dbatra@ttic.edu
David Crandall[1]
djcran@indiana.edu

[1] Indiana University,
Bloomington, IN
[2] TTI-Chicago,
Chicago, IL

Detecting humans and recognizing their body poses is a key problem in understanding natural images, since people are the focus of many (if not most) consumer photographs. Pose recognition is a challenging problem due not only to the usual complications of object recognition—cluttered backgrounds, scale changes, illumination variations, etc.—but also because of the highly flexible nature of the human body. Traditional approaches that use deformable part-based models for human pose estimation typically assume a kinematic tree structure [2, 7], capturing the kinematic constraints between parts of the body (e.g. that the lower arm is connected to the upper arm, which is connected to the torso, etc.).

In this paper, we propose a new model that addresses these problems from a different perspective. We use a composition of multiple tree-structured models with different numbers of parts and resolution scales, allowing different degrees of structural flexibility at different levels, and connect these models through hierarchical decomposition links between body parts in adjacent levels.

A visualization of our model with three layers is shown in Figure 1. Even though the composite model is a loopy graph, it can be naturally decomposed into the constituent tree-structured sub-problems within each level and the cross-model constraint sub-problem across levels, which is also tree-structured as shown in Figure 1 (right). These tree-structured sub-problems are amenable to exact inference and thus joint inference on the composite model can be performed via dual-decomposition [1].

Our approach builds on the work of Yang and Ramanan [7], which has demonstrated state-of-art performance on recent pose estimation datasets. The key innovation in their deformable parts-based model is the use of a mixture of parts, which allows the appearance of each part to change discretely between different "part types." More formally, their model consists of a set $\mathcal{P}$ of parts in a tree-structured model having edges $\mathcal{E} \subseteq \binom{\mathcal{P}}{2}$, such that $\mathcal{E}$ is a tree. Let $\mathbf{y}$ be a vector that represents a particular configuration of the parts, *i.e.* the location and type of each part. They define a function $S(I,\mathbf{y})$ that scores the likelihood that a given configuration $\mathbf{y}$ corresponds to a person in the image,

$$S(I,\mathbf{y}) = \sum_{p\in\mathcal{P}} D(I,\mathbf{y}_p) + \sum_{(p,q)\in\mathcal{E}} \Big(L(\mathbf{y}_p,\mathbf{y}_q) + T(\mathbf{y}_p,\mathbf{y}_q)\Big), \tag{1}$$

where $D(I,\mathbf{y}_p)$ is the score for part p being in configuration $\mathbf{y}_p$ given local image data (the data term), $L(\mathbf{y}_p,\mathbf{y}_q)$ is the relative location term measuring agreement between locations of two connected parts, and $T(\mathbf{y}_p,\mathbf{y}_q) = \vec{\mathbf{B}}^{t(y_p),t(y_q)}$ measures the likelihood of the observing this pair of part-types. $L(\mathbf{y}_p,\mathbf{y}_q)$ is defined as the negative Mahalanobis distance between part locations, and $T(\mathbf{y}_p,\mathbf{y}_q)$ is a part concurrence table that is learned discriminatively in the training stage.

We generalize this model to include multiple layers, each layer being similar to the base model but with a different number of parts and a different (but still tree-structured) graph structure. In particular, let $\mathcal{M} = \{(\mathcal{P}_1,\mathcal{E}_1),...,(\mathcal{P}_K,\mathcal{E}_K)\}$ be a set of K tree-structured models, let $\mathbf{y}^k$ denote the configuration of the parts in the k-th model, and let $\mathbf{Y} = (\mathbf{y}^1,...,\mathbf{y}^K)$ be the configuration of the entire multi-layer composite model. We now define a joint scoring function,

$$\hat{S}(I,\mathbf{Y}) = \sum_{k=1}^{K} S_k(I,\mathbf{y}^k) + \sum_{k=1}^{K-1} \chi(\mathbf{y}^k,\mathbf{y}^{k+1}), \tag{2}$$

where $S_k(\cdot,\cdot)$ is the single-layer scoring function of equation (1) under the model $(\mathcal{P}_k,\mathcal{E}_k)$, and $\chi(\mathbf{y}^k,\mathbf{y}^{k+1})$ is the cross-model scoring function that measures the compatibility of the estimated configurations between different layers of the model.

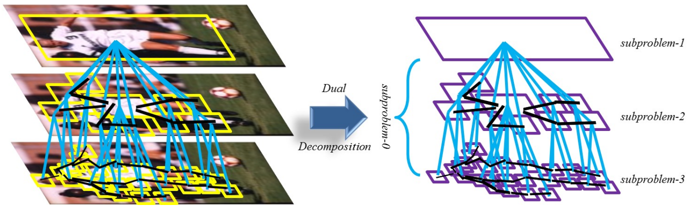

Figure 1: Illustration of our composite part-based models.

As Figure 1 shows, we impose a hierarchical structure on the composite model, such that each part at level k is decomposed into multiple parts at level $k+1$. We call these decomposed parts the child nodes. For a part $p \in \mathcal{P}_k$, let $C(p) \subseteq \mathcal{P}_{k+1}$ be the set of child nodes of p in layer $k+1$. The cross-model scoring function χ scores the relative location and part types of a node in one layer with respect to its children in the layer below,

$$\chi(\mathbf{y}^k,\mathbf{y}^{k+1}) = \sum_{p\in\mathcal{P}_k} \sum_{q\in C(p)} B(\mathbf{y}^k_p,\mathbf{y}^{k+1}_q), \tag{3}$$

where $B(\mathbf{y}^k_p,\mathbf{y}^{k+1}_q)$ is a measure of the likelihood of the relative configuration of a part and its child across the two submodels.

We exploit the natural decomposition of this composite model into tree-structured subproblems to perform inference using dual decomposition, where the key idea is to decompose a joint inference problem into easy sub-problems, solve each sub-problems, and then have the subproblems iteratively communicate with each other until they agree on variable values. To learn the composite model, we stack all of the features in all of the layers together along with the cross-model features into a single feature vector, and formulate this problem as a standard structural SVM problem [5].

We evaluate our composite models on two challenging datasets: Image Parse [4] and UIUC Sport [6]. We evaluate our results using the Percentage of Correct Parts (PCP) metric as defined in [3]. We show that our composite models perform substantially better than state-of-the-art methods on both datasets, which suggests that by combining evidence across submodels, our composite models can obtain better pose estimates of body limbs.

Our model is a general framework for combining different pose estimation models. In future work, we plan to study how to capture rich cross-model constraints inside our composite model (e.g. define relative location constraints between adjacent submodels). We also plan to apply our model to related tasks like human action recognition.

[1] Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 2nd edition, September 1999.

[2] Pedro F. Felzenzwalb and Daniel Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61 (1):55–79, 2005.

[3] Leonid Pishchulin, Arjun Jain, Mykhaylo Andriluka, Thorsten Thormaehlen, and Bernt Schiele. Articulated people detection and pose estimation: Reshaping the future. In *CVPR*, 2012.

[4] Deva Ramanan. Learning to parse images of articulated bodies. In *NIPS*, 2006.

[5] Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6: 1453–1484, 2005.

[6] Yang Wang, Duan Tran, and Zicheng Liao. Learning hierarchical poselets for human parsing. In *CVPR*, 2011.

[7] Yi Yang and Deva Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, 2011.