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

Abstract

We introduce a new approach for part-based human pose estimation using multi-layer composite models, in which each layer is a tree-structured pictorial structure that models pose at a different scale and with a different graphical structure. At the highest level, the submodel acts as a person detector, while at the lowest level, the body is decomposed into a collection of many local parts. Edges between adjacent layers of the composite model encode cross-model constraints. This multi-layer composite model is able to relax the independence assumptions of traditional tree-structured pictorial-structure models while permitting efficient inference using dual-decomposition. We propose an optimization procedure for joint learning of the entire composite model. Our approach outperforms the state-of-the-art on the challenging Parse and UIUC Sport datasets.

1 Introduction

Detecting humans and identifying body pose are key problems 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 to the highly flexible nature of the human body. To deal with this flexibility, deformable part-based models [6, 7] have emerged as a dominant approach in recognizing people and other articulated objects [6, 11, 19, 23, 24, 25]. These part-based models decompose an object into a set of parts, each of which is represented with a local appearance model, and a geometric model that constrains relative configurations of the parts. Recognition is then cast as an inference problem on an undirected graphical model, in which the parts are represented by vertices and the constraints between parts are represented as edges.

Many of these part-based models assume a tree structure [6, 7, 24], 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. Such tree structures allow exact inference to be performed efficiently on the underlying graphical model via dynamic programming. However, the tree structure makes conditional independence assumptions between unconnected parts, which can lead to pose estimates that obey kinematic constraints but are still

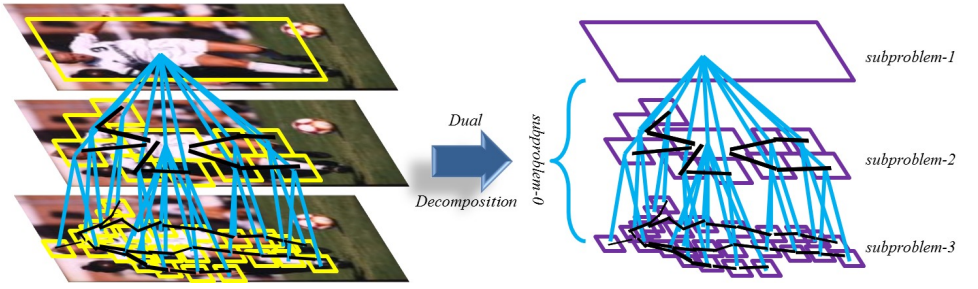


Figure 1: Illustration of our multi-layer composite part-based model.

not sensible; for example, a single image region might be recognized as two different body parts, or a pose might be estimated that defies constraints of gravity and human balance.

A variety of approaches have been proposed for dealing with these problems, including introducing a few cycles into a tree-structured graphical model [23, 25], adding common factor variables [11], or using a fully-connected graphical model [19] to capture more spatial constraints among the parts. Although effective, these approaches introduce cycles into the graphical model which generally makes exact inference intractable. How to model richer spatial constraints that still permit efficient inference is an important open question.

Overview and Contributions. In this paper, we propose a new model that addresses these problems from a different perspective. Instead of adding cycles to the original model, we build a multi-level model consisting of multiple tree-structured models with different resolution scales and numbers of parts, 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 (left). Even though the composite model is a loopy graph, it can be naturally decomposed into 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 [2].

We train these models jointly, and show that the composite models outperform state-of-the-art techniques on two challenging pose recognition datasets. We believe these composite models provide a principled way to trade off the competing goals of model expressiveness and ease of inference, by "stitching" together multiple tree-structured models into a richer composite model while keeping the complexity of joint inference in check.

2 Related Work

Felzenszwalb and Huttenlocher [7] introduced part-based pictorial structure models to the problem of human pose recognition, showing that exact inference on tree-structured graphical models could be performed efficiently via dynamic programming and distance transforms. Ramanan [15] used the same framework but improved the part appearance models and adopted an iterative inference approach. Andriluka *et al*. [1] achieved significantly better results using appearance models learned in a discriminative Adaboost-based framework.

Hierarchical Models. As mentioned in the previous section, various techniques for relaxing the part independence assumptions have been proposed [11, 19]. Particularly relevant to our

work are approaches using hierarchical models, such as Zhu *et al*. [25] and Wang *et al*. [23]. Our proposed composite models are also hierarchical, but differ in the structure of the hierarchy. In our ensemble, each submodel is a separate and complete tree-structured model of human pose, as opposed to simply being "larger" parts [23, 25]. This distinction is crucial since this unique graphical structure allows the use of principled and efficient inference based on dual-decomposition, while reusing existing algorithms developed for tree-structured models.

Multi-scale Models. Capturing visual features at multiple scales has been shown to be important. Sapp *et al*. [16] use cascaded models at different resolutions in order to speed up inference; Park *et al*. [12] use multi-resolution models to detect objects at different scales. Our models incorporate visual cues at multiple resolutions by building HOG feature pyramids as in [24]. We also follow recent work that has modeled the appearance of body joints in addition to body parts, by including joints as extra vertices in the pictorial structure [17, 24].

Mixture Models. To accurately model the highly flexible human form, mixture models for both appearance and geometry have been proposed. Singh *et al*. [18] use mixtures of heterogeneous part detectors, fusing evidence from different feature types. Wang and Mori [22] use mixtures of tree models to capture richer spatial constraints and explicitly model part occlusions. Johnson and Everingham [8] cluster human poses and then build mixtures of pictorial structure models using these clusters. Yang and Ramanan [24] assign a latent "type" variable to each part, allowing parts to select between several appearance models, and jointly learn the parameters in a discriminative structured learning framework. We use a similar approach based on latent part types, but in a framework featuring hierarchical, multi-scale models.

Dual-Decomposition. Some very recent work has applied dual-decomposition to pose recognition, but on different models and applications than ours. Wang and Koller [21] model pose estimation and segmentation jointly, and apply dual-decomposition for efficient inference. Sapp *et al*. [17] use dual-decomposition but their aim is articulated motion parsing in video and relies on motion features, and they do not consider hierarchical models as we do here.

3 Multi-layer Composite Models for Pose Recognition

We now describe our multi-layer composite model for pose recognition.

Base Model. Given an image I and a model of the human body, the goal of pose recognition is to find high-likelihood model configurations in the image. Our approach builds on the work of Yang and Ramanan [24] which has demonstrated state-of-art performance. 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. Moreover, $S(I,\mathbf{y})$ decomposes along the nodes and edges of the tree:

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

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 $T(\mathbf{y}_p,\mathbf{y}_q)$ measures the likelihood of observing this pair of part-types. $L(\mathbf{y}_p,\mathbf{y}_q)$ is defined as the negative Mahalanobis distance between part locations, and

$T(\mathbf{y}_p, \mathbf{y}_q) = \vec{\mathbf{B}}^{t(y_p),t(y_q)}$ is a part co-occurrence table that is learned discriminatively in the training stage, where $t(y_p)$ gives the part type of part p.

Proposed Generalization. We generalize this model to include multiple layers, with each layer like the base model but with a different number of parts and a different tree 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 adjacent layers of the model.

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,

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

where $B(\mathbf{y}_p^k, \mathbf{y}_q^{k+1})$ is a measure of the likelihood of the relative configuration of a part and its child across the two submodels. Next, we describe inference in this composite model and then discuss how to learn parameters of the composite model in Section 3.2.

3.1 Dual Decomposition for Efficient Inference

We have defined our multi-layer composite model as a collection of pose estimation models and a cross-model scoring function. As Figure 1 illustrates, each layer of the hierarchy is tree-structured, so exact inference within each layer can be performed efficiently via dynamic programming. The constraints between layers (blue lines in the figure) also form a tree-structured model, so they are also amenable to exact efficient inference. The overall graphical model has cycles, however, and thus exact inference on this model is not tractable. Fortunately, we can exploit the natural decomposition of this composite model into tree-structured subproblems to perform inference using dual-decomposition. Dual-decomposition is a classical technique [2] that has recently been introduced to the vision literature [10] for solving inference problems in loopy graphical models. The idea is to decompose a joint inference problem into easy sub-problems, solve each sub-problems, and then iteratively have the sub-problems communicate with each other until they agree on variable values.

The following steps are a straightforward adaptation from [10]. Let $\mathcal{C}_k$ denote the set of all feasible (discrete) values for $\mathbf{y}^k$ for each layer of the model. We make a copy of $\mathbf{Y}$, which we call $\mathbf{X} = (\mathbf{x}^1, ..., \mathbf{x}^K)$, and enforce equality constraints that require $\mathbf{Y} = \mathbf{X}$. With this notation, we can rewrite equation (2) as:

$$\max_{\mathbf{Y},\mathbf{X}} \sum_{k=1}^{K} S(I, \mathbf{y}^k) + \sum_{k=1}^{K-1} \chi(\mathbf{x}^k, \mathbf{x}^{k+1}), \quad s.t. \quad \mathbf{y}^k = \mathbf{x}^k, \mathbf{y}^k \in C_k, \mathbf{x}^k \in \mathcal{C}_k, \quad \forall k. \tag{4}$$

We then *dualize* the equality constraints, replacing the hard equality constraints between $\mathbf{Y}$ and $\mathbf{X}$ with a soft penalty term,

$$g(\lambda) = \max_{\mathbf{Y},\mathbf{X}} \sum_{k=1}^{K} S(I,\mathbf{y}^k) + \sum_{k=1}^{K-1} \chi(\mathbf{x}^k,\mathbf{x}^{k+1}) + \sum_{k=1}^{K} \lambda_k \cdot (\mathbf{y}^k - \mathbf{x}^k), \quad s.t. \quad \mathbf{y}^k \in \mathcal{C}_k, \mathbf{x}^k \in \mathcal{C}_k, \quad (5)$$

where λ_k is the Lagrangian multiplier that specifies the strength of the penalty, and $\cdot$ denotes inner product between two vectors. The effect of relaxing the hard equality constraint is that the maximization can now be decoupled into independent terms,

$$g(\lambda) = \sum_{k=1}^{K} \max_{\mathbf{y}^k} \left(S(I,\mathbf{y}^k) + \lambda_k^T \cdot \mathbf{y}^k \right) + \max_{\mathbf{X}} \left(\sum_{k=1}^{K-1} \chi(\mathbf{x}^k,\mathbf{x}^{k+1}) - \sum_{k=1}^{K} \lambda_k^T \cdot \mathbf{x}^k \right). \quad (6)$$

In this form, it is clear that $g(\lambda)$ can by evaluated for a given λ by solving a series of simpler sub-problems. The optimal $\mathbf{Y}$ is found by maximizing each term of the first summation, *i.e.* by performing inference on each individual layer of our composite model via dynamic programming. We can find the optimal $\mathbf{X}$ by solving the maximization in the second term of equation (6), which is also tree-structured and allows the use of dynamic programming.

It can be shown [2] that for each value of λ, the function $g(\lambda)$ provides an upper-bound on the original (constrained) maximization. Thus, we can set up a dual problem that achieves the tightest upper-bound as: $\min_\lambda g(\lambda)$. This dual problem is convex but non-smooth [2], and we use subgradient descent to perform the minimization. Subgradient descent is an iterative algorithm that updates the current setting of $\lambda_k^{(t)}$ at iteration t as:

$$\lambda_k^{(t+t)} \leftarrow \lambda_k^{(t)} - \alpha^{(t)} \left(\mathbf{y}^k(\lambda_k^{(t)}) - \mathbf{x}^k(\lambda_k^{(t)}) \right), \quad (7)$$

where $\mathbf{y}^k(\lambda_k^{(t)}), \mathbf{x}^k(\lambda_k^{(t)})$ are the optimal solutions in (6) for the current setting of $\lambda_k^{(t)}$; and $\alpha^{(t)}$ is the step size at iteration t. For a particular choice of step size, subgradient descent is guaranteed to converge to the optimum of the dual problem [2]. We discuss implementations details like the step size and stopping criteria in Section 4.

3.2 Learning with Structural SVMs

We now address the issue of learning the parameters of our composite model, including the submodel parameters for each layer and the parameters for the cross-model scoring function.

Features. Let $f(I_m,\mathbf{y}^k)$ denote the feature vector for image I_m under submodel k and $f^\chi(\mathbf{Y})$ be the feature vector for the cross-model scoring term. Submodel features $f(I_m,\mathbf{y}^k)$ are the same as those used by Yang and Ramanan [24], *i.e.* HOG features for each part filter, part type co-occurrence features, and deformation features (dx, dx^2, dy, dy^2), where (dx, dy) is the displacement between two parts. The cross-model scoring feature encodes part-type co-occurrences, and is defined as $f^\chi(\mathbf{Y}) = \vec{\delta}_{t(y_p),t(y_q)}$, where $\delta_{t(y_p),t(y_q)} = 1$ if $t(y_p) = t(y_q)$, otherwise $\delta_{t(y_p),t(y_q)} = 0$.

Parameters. To perform joint training for the entire composite model, we stack all features of all of the layers along with the cross-model features into a single feature vector $\Phi(I_m,\mathbf{Y})$,

$$\Phi(I_m,\mathbf{Y}) = \left[f(I_m,\mathbf{y}^1),\ f(I_m,\mathbf{y}^2),\ \ldots,\ f(I_m,\mathbf{y}^K),\ f^\chi(\mathbf{Y}) \right], \quad (8)$$

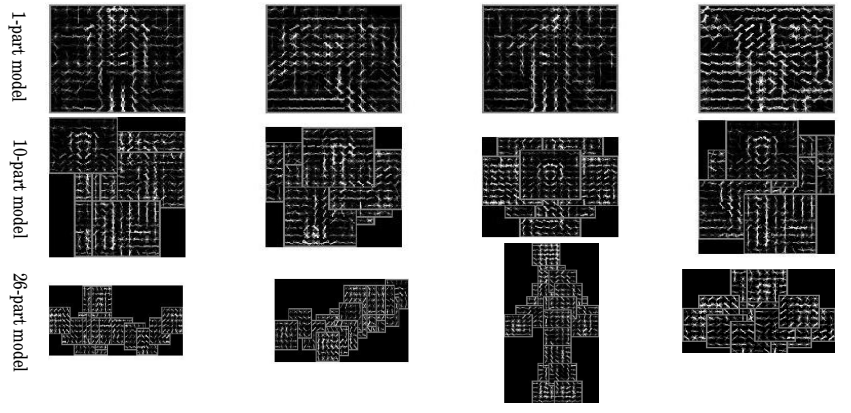


Figure 2: Part-based models used in our multi-layer composite model. For each layer (row) of the composite model, we show four randomly-chosen mixture components.

All parameters of the model are also placed into a single vector $\beta = (\beta^1, \ldots, \beta^K, \beta^\chi)$. The score of the entire composite model on a given image and configuration can then be written as a dot product between parameters and features, $\hat{S}(I, \mathbf{Y}) = \beta \cdot \Phi(I_m, \mathbf{Y})$.

Training. Given training data with labeled positive instances, *i.e.* images containing people with annotated part locations $\{\{I_m, \mathbf{Y}_m\} \mid m \in \text{pos}\}$, and negative instances, *i.e.* images not containing people $\{\{I_m, \emptyset\} \mid m \in \text{neg}\}$, we learn β with a structured SVM formulation [20],

$$\min_{\beta} \quad \frac{1}{2}\|\beta\|^2 + C\sum_m \xi_m \tag{9}$$

$$s.t. \quad \beta \cdot \Phi(I_m, \mathbf{Y}_m) \geq 1 - \xi_m \qquad \forall m \in \text{pos} \tag{10}$$

$$\beta \cdot \Phi(I_m, \mathbf{Y}) \leq -1 + \xi_m \qquad \forall m \in \text{neg}, \forall \mathbf{Y} \tag{11}$$

We optimize this objective function using the dual coordinate descent method of [24]. Note that this formulation forces all of the exponentially many configurations for negative instances to score lower than -1. In practice, we perform dual decomposition with our multi-layer composite model on each negative image to search for hard negative training examples. Implementation details are explained in Section 4.1.

4 Experiments

Datasets. We evaluate our composite models on two challenging datasets: Image Parse [15] and UIUC Sport [23]. Parse contains 100 training and 205 test images, while Sport contains 649 training and 650 test images. Both datasets have one person per image annotated with 14 body joints. We follow [24] and draw our negative images from the INRIA person dataset [4].

4.1 Implementation Details

Inference. For the part appearance models, we follow [24] and others by using HOG features [4] computed at multiple resolutions, yielding a feature pyramid for each image. We perform dual decomposition on each level of the feature pyramid independently, collect detections from all of the levels, and remove overlapping detections via non-maximal suppression. In our current implementation, we restrict our cross-modeling scoring function $B(\cdot,\cdot)$

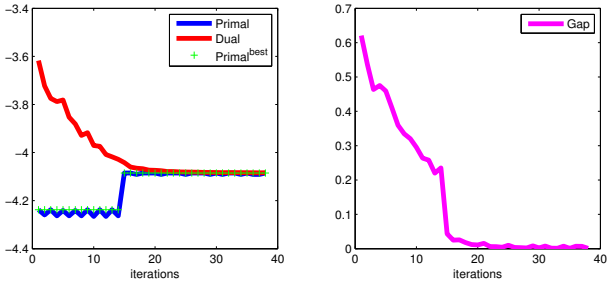


Figure 3: Primal objective and dual objective (left) and primal-dual gap (right) as a function of number of iterations during subgradient descent.

to capture only part type co-occurrence relations. This gives a relatively small label space, which allows efficient inference while obtaining good performance (although modeling relative location between parts across layers is an interesting direction for future work).

The subgradient descent step size in equation (7) is important in making inference work well in practice. We experimented with various strategies, finding that a modification of Polyak's step size rule [14], $\alpha_k^{(t)} = \frac{1+m}{\tau^t+m} \cdot \frac{(dual^t - primal^t_{best})}{\|\nabla g_t\|}$, worked best, where $dual^t$ is the objective value of the dual problem in equation (6) in iteration t, $primal^t_{best}$ is the *best* primal objective value in equation (4) observed so far in iterations up to t, $\|\nabla g_t\|$ is the norm of the subgradient at t, m is a scalar constant (we use $m = 10$), and τ^t is the number of times that the dual-objective has increased up to t. Using this step size rule, dual decomposition converges to a very small gap (< 0.001) quickly, as shown in Figure 3 for a sample image. The entire inference process takes about 20 seconds per Parse image on a 3.0GHz machine.

Learning. For each dataset, we train several variants of our composite models: i) a two-layer model consisting of a 1-part model and a 26-part model; ii) a two-layer model consisting of a 10-part model and a 26-part model; and iii) a three-layer model consisting of 1-part, 10-part, and 26-part models. The 26-part model is the same defined in [24], consisting of both body parts and joints. The 10-part model is defined using new body parts (head, torso, upper arms, lower arms, upper legs, lower legs), and the 1-part model is a simple whole-body template mixture model. The annotations for the 10 and 1 part models were derived from the existing annotations in the datasets. As in [24], the mixture types of each body part are obtained by k-means clustering over joint locations. For the 26-part model, we use the same number of part types per body part as in [24], while for the 10-part model we use 5 torso types, 5 head types, 5 arm types and 6 leg types. The 1-part model uses 9 types. To learn each composite model, we first train a separate model for each layer using the publicly-available code of [24], and then use these models as initialization for learning our composite model.

In practice, there are many more negative (non-person) instances available than positive instances. To reduce the set of negative exemplars that must be considered, we select hard negative exemplars for the next iteration of learning by looking for high-scoring non-person instances under the current multi-layer composite model. To construct negative training instances efficiently, we run the composite model on each negative image, select all detected poses having score above a threshold, sort the detections from each layer, and construct joint exemplars by matching them in the order of detection scores. To speed up training, we stopped subgradient descent after 50 iterations, since in practice the optimization algorithm has typically converged by that point (as in the example in Figure 3). A visualization of a

	Parse dataset							UIUC Sport dataset						
	Torso	UL	LL	UA	LA	Head	Total	Torso	UL	LL	UA	LA	Head	Total
Ramanan [15]	52.1	37.5	31.0	29.0	17.5	13.6	27.2	28.7	7.3	19.2	7.5	20.6	12.9	15.1
Wang [23]	–	–	–	–	–	–	–	75.3	49.2	39.5	25.2	11.2	47.5	37.3
Yang [24]	82.9	69.0	63.9	55.1	35.4	77.6	60.7	85.3	61.3	55.5	49.7	35.5	73.5	56.3
Ours (**26+10**)	82.0	**72.4**	**67.8**	55.6	**36.6**	79.0	62.6	85.4	61.6	**57.9**	49.1	34.8	72.9	56.4
Ours (**26+1**)	**85.6**	71.7	65.6	**57.1**	**36.6**	**80.4**	**62.8**	86.0	**62.2**	57.5	**51.0**	**36.3**	73.7	**57.3**
Ours (**26+10+1**)	81.0	71.7	67.6	55.9	36.3	79.5	62.3	**86.2**	61.2	55.7	49.9	35.9	**73.8**	56.5
Pishchulin [13]*	88.8	77.3	67.1	53.7	36.1	73.7	63.1	–	–	–	–	–	–	–
Johnson [9]*	87.6	74.7	67.1	67.3	45.8	76.8	67.4	–	–	–	–	–	–	–

*[13] and [9] are not directly comparable because they use additional training data with more annotations.

Table 1: Pose estimation results (PCP) on Parse (left) and UIUC Sport (right) datasets. PCP scores are shown for each of six body parts (torso, UL=upper legs, LL=lower legs, UA=upper arms, LA=lower arms, head) and the combined score for all parts. All PCP scores here use criterion 1A (see text for details); for consistency, we re-computed the results from [24] to use this criterion, and for [15] we use the re-computed statistics reported in [23].

sample multi-layer composite model learned using our technique is shown in Figure 2.

4.2 Results

Evaluation criteria. We evaluate our results using the Percentage of Correct Parts (PCP) metric, which counts the fraction of body parts that are correctly localized compared to the ground-truth (within some threshold). Unfortunately, as pointed out in [13], the PCP scoring metric has been implemented in slightly different ways in different papers, which has led to some confusion in the literature. These differences fall along two different dimensions. First, there are two subtly-different definitions of a correct part localization:

1. Part is correctly localized if the distance of *both* its endpoints from respective ground truth endpoints is less than a fraction of the part length; or
2. Part is correctly localized if the *mean* distance between estimated and ground truth endpoints is less than a fraction of the part length.

Second, there are two ways to compute the final aggregate PCP score across the dataset:

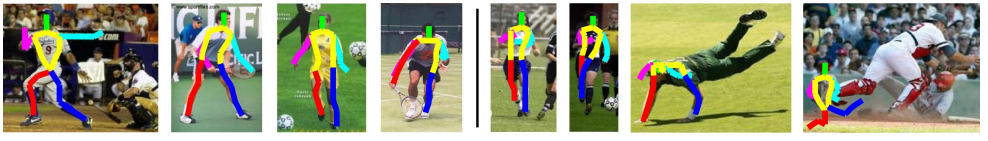
A. PCP is calculated for every image, and averaged across *all* images to produce an aggregate score; or
B. PCP is calculated *only* for images in which the human is correctly localized according to a ground truth bounding box, these scores are averaged together, and then multiplied by the detection rate.

According to our understanding, Eichner *et al*. [5] proposed variant 1B, but their publicly-released software toolkit implemented 2B which yields higher scores. Yang *et al*. [24] also used 2B, while both Pischulin *et al*. [13] and Wang *et al*. [23] use 1A. Unfortunately, these seemingly subtle variations lead to significant differences. We follow the two latter papers and also use 1A, which we hope will become the standard definition, but also report results under the other variants to illustrate the significant differences they create. Note that [13] does not report PCP numbers for individual parts, but rather combines right and left parts together. We do the same, and also average the PCP of the left and right limbs reported by [23] to convert their results into this metric as well.

Results. PCP results (using variant 1A) on Parse and UIUC Sport datasets are shown in Table 1. We see that our composite models outperform state-of-the-art methods on both datasets, beating [24] by about 2 percentage points for Parse and by 1 percentage point

	PCP (variant 1A)								PCP 1B	PCP 2B
Threshold	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.5	0.5
Yang [24]	33.4	47.2	56.0	60.7	64.4	67.2	69.7	71.5	56.0	74.9
Ours (**26+10**)	**34.5**	**49.2**	**57.6**	62.6	65.9	68.7	71.3	73.0	58.5	75.0
Ours (**26+1**)	**34.5**	48.3	56.5	**62.8**	**66.9**	**70.0**	**72.0**	**73.6**	59.3	75.8
Ours (**26+10+1**)	34.3	48.9	57.3	62.3	65.7	68.6	70.9	72.7	**59.5**	**75.9**

Table 2: Evaluation results on the Parse dataset under different definitions of Percentage of Correct Poses (PCP), using variants 1A, 1B and 2B which have all been used by different papers in the literature (see text for details). For variant 1A, we show results under different evaluation thresholds, where larger thresholds are more lenient in scoring part localizations.



(a) (b)

Figure 4: Sample results. *(a):* Examples in which [24] failed, but our 3-level model estimated poses correctly. *(b):* Some failure cases of our model.

for Sport. Among our composite models, the 2-layer model (26+1) achieves the best performance under PCP-1A, however the 3-layer model performs best under PCP-1B,2B. Our models also outperformed in terms of person detection rate, with 79.0%, 81.9%, and 82.4% for our 26+10, 26+1, and 26+10+1 models, respectively, compared to 76.6% for [24]. This suggests that much of our increase in PCP is due to more accurate detections. This is an intuitive result because our 1-part layer (consisting of a mixture of large HOG templates) can be considered a person detector. Our composite models with models at multiple scales thus combine the advantages of single-part models for person detection, with the highly flexible multi-part models needed for accurate part localization. Some qualitative results are presented in Figure 4, showing cases in which our method correctly estimated pose while [24] failed for one or more limbs, as well as some failure cases.

Table 2 presents experimental results under alternative definitions of PCP. For PCP criterion 1A, we present scores for different values of the part localization threshold (which specifies the percentage of body part length that part endpoints can be from the positions given in ground truth). The table also shows PCP results computed under two alternative definitions that have been used in the literature (1B and 2B). We see that seemingly subtle differences in PCP definition can yield very different conclusions. Our composite models beat [24] under all of the criteria, but which composite model performs best depends on the PCP metric. Moreover, variant 2B yields much higher absolute PCP scores, illustrating the importance of adopting a consistent metric to avoid further confusion in the literature.

5 Conclusion

In this paper we presented a multi-layer composite model for human pose estimation problems. By combining different cues from different submodels, our composite model outperforms state-of-the-art pose estimation methods on challenging datasets. Our model is a general framework for combining different pose estimation models. In future work, we plan to study how to capture richer cross-model constraints (e.g. define spatial constraints between

adjacent submodels), and to apply our model to related tasks like human action recognition.

Acknowledgements. We thank Devi Parikh for helpful comments and discussions. Part of this work was done while Kun Duan was an intern at TTI-Chicago. This work was supported in part by the Lilly Endowment and the Indiana University Data-to-Insight Center.

References

[1] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR*, 2009. 2

[2] Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 2nd edition, September 1999. 2, 4, 5

[3] D. Crandall, P. Felzenszwalb, and D. Huttenlocher. Spatial priors for part-based recognition using statistical models. In *CVPR*, 2005. 1

[4] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 6

[5] Martin Eichner, Manuel Marin-Jimenez, Andrew Zisserman, and Vittorio Ferrari. Articulated human pose estimation and search in (almost) unconstrained still images. Technical report, ETHZ, 2010. 8

[6] Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:1627–1645, 2010. ISSN 0162-8828. doi: http://doi.ieeecomputersociety.org/10.1109/TPAMI.2009.167. 1

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

[8] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010. 3

[9] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *CVPR*, 2011. 8

[10] Nikos Komodakis, Nikos Paragios, and Georgios Tziritas. Mrf energy minimization and beyond via dual decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(3):531–552, 2011. 4

[11] Xiangyang Lan and Daniel P. Huttenlocher. Beyond trees: Common-factor models for 2d human pose recovery. In *ICCV*, 2005. 1, 2

[12] Dennis Park, Deva Ramanan, and Charless Fowlkes. Multiresolution models for object detection. In *ECCV*, 2010. 3

[13] Leonid Pishchulin, Arjun Jain, Mykhaylo Andriluka, Thorsten Thormaehlen, and Bernt Schiele. Articulated people detection and pose estimation: Reshaping the future. In *CVPR*, 2012. 8

[14] Boris T. Polyak. A general method for solving extremum problems. *Soviet Math*, 8(3), 1967. 7

[15] Deva Ramanan. Learning to parse images of articulated bodies. In *NIPS*, 2006. 2, 6, 8

[16] Benjamin Sapp, Alexander Toshev, and Ben Taskar. Cascaded models for articulated pose estimation. In *ECCV*, 2010. 3

[17] Benjamin Sapp, David Weiss, and Ben Taskar. Parsing human motion with stretchable models. In *CVPR*, 2011. 3

[18] Vivek Kumar Singh, Ram Nevatia, and Chang Huang. Efficient inference with multiple heterogeneous part detectors for human pose estimation. In *ECCV*, 2010. 3

[19] Duan Tran and David Forsyth. Improved human parsing with a full relational model. In *ECCV*, 2010. 1, 2

[20] 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

[21] Huayan Wang and Daphne Koller. Multi-level inference by relaxed dual decomposition for human pose segmentation. In *CVPR*, 2011. 3

[22] Yang Wang and Greg Mori. Multiple tree models for occlusion and spatial constraints in human pose estimation. In *ECCV*, 2008. 3

[23] Yang Wang, Duan Tran, and Zicheng Liao. Learning hierarchical poselets for human parsing. In *CVPR*, 2011. 1, 2, 3, 6, 8

[24] Yi Yang and Deva Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, 2011. 1, 3, 5, 6, 7, 8, 9

[25] Long Zhu, Yuanhao Chen, Yifei Lu, Chenxi Lin, and Alan L. Yuille. Max margin AND/OR graph learning for parsing the human body. In *CVPR*, 2008. 1, 2, 3