Human Pose Estimation via Multi-layer Composite Models[☆]

Kun Duan[☆☆a], Dhruv Batra[b], David J. Crandall[a]

[a]*School of Informatics and Computing, Indiana University, 919 E. Tenth Street, Bloomington, IN 47408*
[b]*Bradley Department of Electrical and Computer Engineering, Virginia Tech, 302 Whittemore, Blacksburg, VA 24061*

Abstract

We introduce a hierarchical part-based approach for human pose estimation in static images. Our model is a multi-layer composite of tree-structured pictorial-structure models, each modeling human 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 in tree-structured pictorial-structures models (which can create problems like double-counting image evidence), while still 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 four challenging

[☆]This is an expanded version of a paper that appeared at the British Machine Vision Conference (Duan et al. (2012)).

[☆☆]Corresponding author. Email: kduan@indiana.edu.

datasets: Parse, UIUC Sport, Leeds Sport Pose and FLIC datasets.

Keywords: object detection, human pose estimation

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 and videos. Accurate pose modeling and recognition could enable a wide range of applications, including detecting suspicious actions in surveillance video, tracking user motion in interactive and immersive video games (Shotton et al. (2011)), organizing consumer photo collections automatically based on human activities, and even synthesizing realistic poses in graphics applications (Yu et al. (2012)). Many current commercial pose estimation systems like Microsoft Kinect use binocular cameras with additional depth sensors, but these techniques do not apply on consumer cameras that do not have these sensors. Detecting people and recognizing pose in 2D static images is significantly more difficult, 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.

Deformable part-based models have emerged as a dominant approach to deal with this flexibility in recognizing people and other articulated objects (Felzenszwalb and Huttenlocher (2005); Crandall et al. (2005); Lan and Huttenlocher (2005); Zhu et al. (2008); Tran and Forsyth (2010); Felzenszwalb

et al. (2010); Yang and Ramanan (2011); Wang et al. (2011)). These 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, 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. (Felzenszwalb and Huttenlocher (2005); Felzenszwalb et al. (2010); Yang and Ramanan (2011)). These tree structures allow exact inference to be performed efficiently on the underlying graphical model via dynamic programming. Tree-structured models may seem to be ideal for modeling the tree-structured human body, but it turns out that they make assumptions that are not realistic. In particular, the tree structure assumes disconnected parts are conditionally independent, which is not generally true due to constraints imposed by gravity and human balance. For instance, it is much more likely for a person to stand in a symmetrical pose than to stand with all limbs on one side of the torso, but there is no way of encoding this information in a tree model. Similarly, tree models may recognize a single image region as two different body parts, because there is no way to require mutual exclusivity between parts that are not directly connected in the model.

This problem is exacerbated as the number of parts grows: the more

parts we have, the more incorrect independence assumptions we are making. (More precisely, note that the number of possible pairs of parts increases quadratically with the number of parts in the model, whereas the number of constraints captured by a tree-structured model increases only linearly.) But recent work suggests that large numbers of parts are beneficial for accurate pose recognition. For instance, Yang and Ramanan (2011) found that many small parts with pairwise spatial models based only on translations can approximately model non-affine transformations of the whole object. Thus there is a trade-off between encoding more flexibility in the model by adding parts, and accurately modeling the structural constraints on human pose.

A variety of approaches have been proposed for overcoming the assumptions made by tree-structured models, including introducing a few cycles into an otherwise tree-structured graphical model (Zhu et al. (2008); Wang et al. (2011)), adding common factor variables (Lan and Huttenlocher (2005)), or even using a fully-connected graphical model (Tran and Forsyth (2010)). 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.

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 we connect these models through

cross-level links between body parts in adjacent levels. The set of cross-level links also forms a tree. A visualization of our model with three layers is shown in Figure 2, with cross-level links shown in blue and intra-level links shown in black. The combined graph is no longer a tree, but can be decomposed into tree-structured sub-problems within each level and a cross-model constraint sub-problem across levels. These tree-structured sub-problems are amenable to exact inference, and thus joint inference on the composite model can be performed via dual-decomposition (Bertsekas (1999)). To learn parameters for our models, we train the cross-layer and intra-layer models jointly, and show that the composite models outperform state-of-the-art techniques on challenging pose recognition datasets. We believe these composite models provide a principled way to trade off between model expressiveness and ease of inference, by "stitching" together multiple tree-structured models into a richer model while keeping the complexity of joint inference in check.

Contributions. To summarize, our contributions include: (1) a novel multi-layer model for human pose estimation, (2) an efficient inference algorithm using dual-decomposition, (3) an algorithm for jointly learning the parameters of our models using structural SVMs, and (4) experimental results showing that the models outperform existing work on modern benchmarks.

Outline. The remainder of this paper presents our method in detail. We discuss related work in Section 2 and describe the model and inference and learning algorithms in Section 3. We evaluate our approach again strong

baseline methods in Section 4, before concluding in Section 5.

2. Related Work

We begin our summary of related work by describing in Section 2.1 the pictorial structures model, which forms the basis for most work in statistical pose recognition. We then describe several recently-proposed improvements to these models. We then give background on dual decomposition, the technique that forms the basis of our inference algorithm, in Section 2.2. In Section 2.3 we briefly highlight other recent work in pose recognition that are less related to our work but still relevant to the general problem domain.

2.1. Statistical pose recognition

Felzenszwalb and Huttenlocher (2005) introduced part-based tree-structured deformable models to human pose recognition, calling these models "pictorial structures." In a part-based model, an object is represented as a collection of "parts" with constraints on the spatial relationship between them. Each part's appearance model captures how it "looks" locally, while the spatial models prefer some configurations but allow some deformation. Pictorial structures can be modeled as probabilistic graphical models, with a latent variable for each part denoting that part's pose, and connections between some variables that encode constraints on relative part position. Each variable also observes the image data. Pose recognition can then be thought of as an inference problem, with the goal of finding the most-likely values for the pose variables given the image data.

The original work of Felzenszwalb and Huttenlocher (2005) uses 4-d latent variables that parameterized part pose as 2-d position, orientation, and scale, and encodes the pairwise configuration priors as Gaussian distributions. They show that exact inference on the tree-structured graphical models could be performed efficiently via dynamic programming and distance transforms, requiring only $O(ph)$ time, where p is the number of parts in the model and h is the number of possible pose configurations for each part. They use very simple part appearance models, essentially looking for rectangular blobs in binary segmentations produced by background subtraction, but later work built on their technique to create state-of-the-art pose recognition systems.

For instance, Ramanan (2006) uses the same framework but improves the part appearance models and adopted an iterative inference approach. An edge-based deformable model is first applied on the image to obtain a soft (and noisy) estimate of body part locations, then a region-based model is used to look for body parts based on learned "part-specific" appearance models. The resulting soft estimates are then used to build new models, and this process is repeated iteratively. Later work, Ramanan et al. (2007), extends this to track pose across time in video, where the appearance models become customized to a particular person (*e.g.* based on clothing). Andriluka et al. (2009) achieve better results by learning appearance models in a discriminative Adaboost-based framework. Their representation is based on dense shape context descriptors, while the kinematic pose prior is modeled as a tree learned separately on a multi-view and multi-articulation dataset.

More recent work has explored other strategies for enhancing the pictorial structures framework, and we borrow several of these innovations. We discuss three specific enhancements in the following sections: hierarchical models, multi-scale feature representations, and mixture models.

Hierarchical Models. As mentioned above, various techniques for relaxing the part independence assumptions have been proposed (*e.g.* Lan and Huttenlocher (2005); Tran and Forsyth (2010)). Particularly relevant to our work are approaches using hierarchical models, such as Zhu et al. (2008) and Wang et al. (2011). Zhu et al. (2008) propose a Max Margin AND/OR graph for parsing the human body into parts. The model is a multi-level mixture of Markov Random Fields where each node represents a human body part at a certain level in the hierarchy. Wang et al. (2011) propose using "hierarchical poselets," which are multi-scale body parts of various sizes. The body parts are defined in a hierarchy at different scales, and are able to cover human poses at various levels of granularity. 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. This distinction is crucial since this unique graphical structure allows principled and efficient inference with dual-decomposition, applying existing algorithms developed for tree-structured models to solve the sub-problems.

Multi-scale Models. In addition to hierarchical spatial models, recent work has shown that capturing visual features at multiple image scales is

important. Sapp et al. (2010) use cascaded models at different resolutions, because most non-human image windows can be rejected at a coarse scale. Their cascaded models progressively filter the pose state space, reducing computation while maintaining state-of-art performance. Park et al. (2010) use multi-resolution models to detect objects at different scales. For each object, they introduce a binary variable corresponding to two visual scales (coarse and fine). During detection, this variable and object locations are jointly inferred. Yang and Ramanan (2011) use visual cues at multiple resolutions with HOG feature pyramids, which we also do here.

Mixture Models. To accurately model the flexible human form, mixture models for both appearance and geometry have been proposed. Singh et al. (2010) use a linear weighted combination of heterogeneous part detectors, fusing evidence from different feature types. A branch-and-bound approach makes inference on the graphical model faster. Wang and Mori (2008) use mixtures of tree models to capture richer spatial constraints and explicitly model part occlusions. A boosting procedure is used to combine different tree structures during inference. Johnson and Everingham (2010) cluster human poses and then build mixtures of pictorial structure models using these clusters. The appearance models are "pose specific" and capture the correlation between part appearances. Yang and Ramanan (2011) introduce mixture models for each human body part. Specifically, their model assigns a latent "type" variable to each part, allowing parts to select between several

appearance models, and they 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.

2.2. Dual Decomposition

We apply an inference technique called dual decomposition (also called) Lagrangian relaxation), which has been shown to be useful for solving optimization problems with discrete variables (Komodakis et al. (2011)). It decomposes the original optimization objective into small, independent, and easy-to-solve subproblems, and then combines the solutions to the subproblems into a global solution. Some recent work has also applied dual decomposition to pose recognition, but on different models and applications than ours. Wang and Koller (2011) model pose estimation and segmentation jointly, and apply dual-decomposition for efficient inference. The human pose estimation guides the foreground pose segmentation at a high level, and the segmentation cues also improve pose estimation by explaining low-level pixels. Sapp et al. (2011) use dual-decomposition for pose parsing in video. A tree model is used to capture the human pose structures, and motion cues connect joints across consecutive frames. Dual-decomposition decouples the problem into several small subproblems, each of which tracks a single joint and connects two tree structures. Since all subproblems are still tree-structured, one can perform exact and efficient inference on them. However, that paper does not consider hierarchical models as we do here.

2.3. Other Relevant Work

Other papers have addressed pose estimation from other perspectives. Work in human motion analysis, such as Zhou et al. (2009), has used gait to achieve efficient tracking. Hara and Chellappa (2013) propose to use a dependency graphs for modeling relations between body parts, with independently trained discriminative part regressors. Dantone et al. (2013) jointly learn non-linear part regressors using two-layered random forests, and show significant improvement over the start-of-art methods. Sapp and Taskar (2013a) capture multiple pose models for full and half (above the waist) body models, where each pose mode is defined as a cluster of possible joint configurations, and is trained using a discriminative structured linear model. Finally, since we use sparse representations for human pose features (i.e. the deformation and part-type co-occurrence features), our work is related with efficient model training using sparse representations (Cheng et al. (2013)).

3. Multi-layer Composite Models for Pose Recognition

We now describe our technique for pose recognition using multi-layer composite models. We begin by describing our baseline model, which is very similar to Yang and Ramanan (2011). We show how to generalize this to a multiscale model in Section 3.2, and then discuss how to perform inference efficiently using dual-decomposition in Section 3.3. We then describe how to jointly learn the parameters of the model from labeled training data in Section 3.4. All of the notation used in this section is summarized in Table 3.

3.1. 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 (2011) which has demonstrated state-of-art performance. The key innovation in their deformable part-based model is to use a mixture of parts, which allows the appearance of each part to change discretely between different "part types." A part type is a specific configuration of the pose for a specific part. For example, a leg might have part types like "lying down" or "standing up," corresponding to different orientations and articulations of the leg. Also, instead of using parts that correspond with natural body (arms, torso, hands, *etc.*), they use small square part templates for each body *joint* (*e.g.* ankles, elbows, chin, top of head, *etc.*). This gives additional invariance to pose changes, since the appearance of a joint varies less dramatically than the part appearances themselves.

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. Specifically, $D(I, \mathbf{y}_p)$ is the template matching score for part p at location $\mathbf{y}_p$, $L(\mathbf{y}_p, \mathbf{y}_q)$ is defined as the negative Mahalanobis distance between part locations, and $T(\mathbf{y}_p, \mathbf{y}_q) = \vec{\mathbf{B}}^{t(\mathbf{y}_p), t(\mathbf{y}_q)}$ is a part co-occurrence table that is learned discriminatively in the training stage, where $t(\mathbf{y}_p)$ gives the part type of part p.

3.2. 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 2(a) 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 also a look-up table, and measures 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 3.4.

3.3. 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 2(a) illustrates, each layer of the hierarchy is tree-structured, so exact inference within each layer can be performed efficiently. The constraints between layers (blue lines in the figure) also are tree-structured and thus also amenable to exact inference. The overall graphical model has cycles, however. 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 (Bertsekas (1999)) that has more recently been used in vision (Komodakis et al. (2011)) for performing infer-

ence in loopy graphical models. The idea is to decompose a joint inference problem into easy sub-problems, solve each sub-problem, and then have the sub-problems iterate until they agree on variable values.

The following is a straightforward adaptation of Komodakis et al. (2011). 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 $\mathbf{Y} = \mathbf{X}$. Then we rewrite equation (2) as:

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

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

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

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,

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

again with $\mathbf{y}^k \in \mathcal{C}_k$, $\mathbf{x}^k \in \mathcal{C}_k$. In this form, it is clear that $g(\lambda)$ can by evaluated for a given λ by solving several 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, which is efficient because each layer is tree-structured. We find the optimal $\mathbf{X}$ by maximizing the second term of equation (6), which is also tree-structured.

A conceptual illustration of dual decomposition is shown in Figure 2(b), while Figure 2(c) shows how we apply it to pose detection. Each of the slave problems are tree-structured inference tasks and can be performed efficiently with dynamic programming. The master's job is to encourage agreement between the solutions found by the slave tasks.

It can be shown that for each value of λ, the function $g(\lambda)$ provides an upper-bound on the original (constrained) maximization (Bertsekas (1999)). 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, so 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,

$$\lambda_k^{(t+t)} \leftarrow \lambda_k^{(t)} - \alpha^{(t)}\Big(\mathbf{y}^k(\lambda_k^{(t)}) - \mathbf{x}^k(\lambda_k^{(t)})\Big), \tag{7}$$

where $\mathbf{y}^k(\lambda_k^{(t)}), \mathbf{x}^k(\lambda_k^{(t)})$ are the optimal solutions in equation (6) for the current setting of $\lambda_k^{(t)}$, and $\alpha^{(t)}$ is the step size at iteration t. For a good choice of step size, subgradient descent is guaranteed to converge to the optimum of the dual problem (Bertsekas (1999)). We discuss implementation details

like the step size and stopping criteria in Section 4.

3.4. Learning with Structural SVMs

We now address how to learn the parameters of our composite model, including the parameters for each layer and the cross-model scoring function.

Features. We use four kinds of features: part appearance features that characterize local image evidence, deformation features that capture spatial relationships between parts, part type co-occurrence features within layers, and part type co-occurrence features across layers. We combine the first three into a feature vector called $f(I_m, \mathbf{y}^k)$ for image I_m under submodel k (Yang and Ramanan (2011)). In particular, $f(I_m, \mathbf{y}^k)$ consists of 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. We denote the fourth feature type, the cross-model part type co-occurrence feature as $f^{\chi}(\mathbf{Y})$ by converting the 2D look up table $\delta_{t(\mathbf{y}_p),t(\mathbf{y}_q)}$ to a 1D vector, where $\delta_{t(\mathbf{y}_p),t(\mathbf{y}_q)} = 1$ if $t(\mathbf{y}_p) = t(\mathbf{y}_q)$, and otherwise $\delta_{t(\mathbf{y}_p),t(\mathbf{y}_q)} = 0$.

Parameters. To jointly train the composite model, we stack features from all layers along with the cross-model features into a single vector $\Phi(I_m, \mathbf{Y})$,

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

and 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 $\big\{\{I_m, \mathbf{Y}_m\} \mid m \in \text{pos}\big\}$, and negative instances, *i.e.* images not containing people $\big\{\{I_m, \emptyset\} \mid m \in \text{neg}\big\}$, we learn β with a structured SVM (Tsochantaridis et al. (2005)). Structured SVMs are a generalization of standard binary SVMs that can generate structured outputs (instead of simple positive or negative outputs) and can use loss functions that are defined based on this structured information. In our context, the structured learning problem involves finding β such that,

$$\begin{aligned} \min_{\beta} \quad & \frac{1}{2}\|\beta\|^2 + C\sum_m \xi_m \\ s.t. \quad & \beta \cdot \Phi(I_m, \mathbf{Y}_m) \geq 1 - \xi_m \qquad \forall m \in \text{pos} \\ & \beta \cdot \Phi(I_m, \mathbf{Y}) \leq -1 + \xi_m \qquad \forall m \in \text{neg}, \forall\, \mathbf{Y}. \end{aligned} \tag{9}$$

We optimize this objective function using dual coordinate descent (Yang and Ramanan (2011)). This formulation forces all of the exponentially many configurations of negative instances to score lower than -1. For efficiency, in practice we enforce this constraint only on hard negative examples, as explained in Section 4.1.2.

4. Experiments

We evaluate our composite models on four challenging datasets: Image Parse (Ramanan (2006)), UIUC Sport (Wang et al. (2011)), Leeds Sport Pose (Johnson and Everingham (2010)) and FLIC (Sapp and Taskar (2013b)). Parse contains 100 training and 205 test images, UIUC Sport includes 649 training and 650 test images, Leeds Sport has 1,000 training and 1,000 test images, while FLIC has 3,987 training and 1,016 test images The first three datasets have one person per image annotated with 14 body joints, while FLIC includes only upper body annotations. We follow Yang and Ramanan (2011) and draw our negative images from the INRIA person dataset (Dalal and Triggs (2005)).

4.1. Implementation

We implemented inference and learning methods described in Section 3. Here we give some implementation details that are important in practice.

4.1.1. Inference

For part appearance models, we follow Yang and Ramanan (2011) by using HOG features (Dalal and Triggs (2005)) computed at multiple resolutions, yielding a feature pyramid for each image. We perform dual decomposition on each level of the pyramid independently, collect detections from all of the levels, and remove overlapping detections via non-maximal suppression. We currently restrict our cross-modeling scoring function $B(\cdot, \cdot)$ in equation (3) to only part type co-occurrence relations; this gives a relatively small label

space which allows more efficient inference, 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 (Polyak (1967)),

$$\alpha_k^{(t)} = \frac{1+m}{\tau^{(t)}+m} \cdot \frac{(dual^{(t)} - primal_{best}^{(t)})}{\|\nabla g_t\|},$$

works best, where $dual^{(t)}$ is the objective value of the dual problem in equation (6) in iteration t, $primal_{best}^{(t)}$ 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 $\tau^{(t)}$ 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. Since each submodel is a pictorial structure, inference at each level takes linear time in the number of parts using dynamic programming Felzenszwalb and Huttenlocher (2005). The entire inference takes about 20 seconds per Parse image on a 3.0GHz machine.

4.1.2. Learning

For Parse, UIUC Sport, Leeds Sport dataset, we trained several variants of our composite models: 1) a two-layer model consisting of a 1-part and a 26-part model; 2) a two-layer model consisting of a 10-part and a 26-part model; 3) a three-layer model consisting of 1-part, 10-part, and 26-part models. For

FLIC, we use 11-part, 6-part and 1-part models.

In the models for the first three datasets, the 26-part model is the same defined in Yang and Ramanan (2011), 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 Yang and Ramanan (2011), the mixture types of each body part were obtained by k-means clustering over joint locations. For the 26-part model, we use the same number of part types per part as in Yang and Ramanan (2011), *i.e.* a variable number of 5 or 6 mixtures for each part, 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 model for each layer using the public code of Yang and Ramanan (2011), and then use these models as initialization for learning our composite model. For FLIC, we re-train a 11-part model using the method in Yang and Ramanan (2011), while the 6-part and 1-part models are defined as above.

In practice, there are many more negative (non-person) instances available than positive ones. To reduce the number of negative exemplars to be considered in equation (9), we select hard negative exemplars for the next iteration of learning by finding high-scoring non-person instances under the current multi-layer composite model. To construct negative instances efficiently, we run the composite model on each negative image, select all de-

tected poses having score above a threshold, sort the detections in 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 the optimization typically converges by then (as in Figure 3). A multi-layer composite model learned by our technique is shown in Figure 4.

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 according to the ground-truth. One needs to define what a correct localization is, since small discrepancies in part pose are probably not noticable in most applications. Unfortunately, as noted in Pishchulin et al. (2012), the PCP metric has been defined differently across papers, leading to confusion in the literature. These differences fall along two 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. This difference is illustrated in Figure 5. 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 lo-

calized according to a ground truth bounding box, these scores are averaged together, and then multiplied by the detection rate.

The cross-product of these two possibilities yields four possible evaluation criteria. According to our understanding, Eichner et al. (2010) proposed variant 1B, but their publicly-released software toolkit implemented 2B which yields higher scores. Yang and Ramanan (2011) also used 2B, while both Pishchulin et al. (2012) and Wang et al. (2011) used 1A. We follow the two latter papers and use 1A, which we hope will become the standard, but also report results under the other variants to illustrate the significant differences they create. Note that Pishchulin et al. (2012) do not report PCP numbers for individual parts, but rather combine right and left parts together. We do the same, and also average the PCP of the left and right limbs reported by Wang et al. (2011) to convert their results into this metric as well.[1]

Results. Results on Parse, UIUC Sport, Leeds Sport Pose and FLIC datasets are shown in Table 1 for our technique and several other recent methods. To make results compatible, we converted all numbers into metric 1A (by re-computing results from Yang and Ramanan (2011), and for Ramanan (2006) by using re-computed numbers reported in Wang et al. (2011)). Our models outperform the baselines on all four datasets, beating Yang and Ramanan

[1]Recently, Yang and Ramanan (2013) propose new metrics called *Percentage of Correct Keypoints* and *Average Precision of Keypoints*, which compare predicted and ground truth bounding boxes surrounding each keypoint of the body. Since no other work has reported performance under these criteria, we stick with the better-known PCP metric here.

(2011) by about 2 percentage points for Parse and by 1.0-1.5 percentage point for the other datasets. We also show results from two recent techniques, Pishchulin et al. (2012) and Johnson and Everingham (2011), that are not comparable because they used richer training data.[2] Our technique is still competitive considering that we use vastly less training information.

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 (*i.e.* the maximum allowable distance that endpoints can be from ground truth positions, as a fraction of part length). The table also compares results using two other PCP definitions that have been used in the literature (1B and 2B). We see that subtle differences in PCP definition can yield very different conclusions. Our composite models beat Yang and Ramanan (2011) under all PCP metrics, but which composite model performs best varies. For instance, the 2-layer model (26+1) achieves the best performance under 1A, but the 3-layer model performs best under 1B and 2B. Moreover, 2B yields much higher PCP scores, showing the importance of adopting a consistent metric to avoid further confusion in the literature.

Some qualitative pose recognition results are presented in Figure 6, showing cases in which we correctly estimated pose while Yang and Ramanan (2011) failed, as well as some images on which our technique failed.

[2]In particular, Johnson and Everingham (2011) annotated a training dataset of $10,800$ images downloaded from Flickr using Amazon Mechanical Turk. Pishchulin et al. (2012) fit a 3D human body shape model to each training image with annotated 2D body keypoints, and then vary the 3D shape parameters in order to create additional 2D training poses.

We also evaluated 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 Yang and Ramanan (2011). This suggests that much of our increase in PCP is due to more accurate detections. This makes sense because our 1-part model (a mixture of large HOG templates) is like the person detector of Dalal and Triggs (2005). Our composite models with multiple scales combine the advantages of single-part person detection models with the flexible multi-part models needed for accurate part localization.

5. Conclusion

We presented a multi-layer composite model for human pose estimation. By combining cues from different submodels, our composite model outperforms state-of-the-art pose estimation methods on challenging datasets. These results show that hierarchical structures and mixture models for parsing the human body are important, and that dual decomposition for such composite model is effective in practice. Our model is a general framework for combining different pose estimation models. In future work, we plan to study improvements including richer cross-model constraints by defining spatial constraints between adjacent submodels, and learning the composite model in a weakly supervised mode when annotations are not available for all key points on the training images. Our model could also be applied to interesting tasks like action recognition and tracking.

Acknowledgements. This work was supported in part by an NSF CAREER award (IIS-1253549) and by the IU Office of the Vice Provost for

Research through the Faculty Research Support Program. Part of this work was done when Kun Duan was an intern at TTI-Chicago.

Andriluka, M., Roth, S., Schiele, B., 2009. Pictorial structures revisited: People detection and articulated pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition.

Bertsekas, D. P., September 1999. Nonlinear Programming, 2nd Edition. Athena Scientific.

Cheng, H., Liu, Z., Yang, L., Chen, X., 2013. Sparse representation and learning in visual recognition: Theory and applications. Signal Processing 93 (6), 1408–1425.

Crandall, D., Felzenszwalb, P., Huttenlocher, D., 2005. Spatial priors for part-based recognition using statistical models. In: IEEE Conference on Computer Vision and Pattern Recognition.

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. In: IEEE Conference on Computer Vision and Pattern Recognition.

Dantone, M., Gall, J., Leistner, C., Gool, L. J. V., 2013. Human pose estimation using body parts dependent joint regressors. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3041–3048.

Duan, K., Batra, D., Crandall, D. J., 2012. A multi-layer composite model for human pose estimation. In: British Machine Vision Conference.

Eichner, M., Marin-Jimenez, M., Zisserman, A., Ferrari, V., 2010. Articulated human pose estimation and search in (almost) unconstrained still images. Tech. rep., ETH Zurich.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., Ramanan, D., 2010. Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence 32, 1627–1645.

Felzenszwalb, P. F., Huttenlocher, D., 2005. Pictorial structures for object recognition. International Journal of Computer Vision 61 (1), 55–79.

Hara, K., Chellappa, R., 2013. Computationally efficient regression on a dependency graph for human pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3390–3397.

Johnson, S., Everingham, M., 2010. Clustered pose and nonlinear appearance models for human pose estimation. In: British Machine Vision Conference.

Johnson, S., Everingham, M., 2011. Learning effective human pose estimation from inaccurate annotation. In: IEEE Conference on Computer Vision and Pattern Recognition.

Komodakis, N., Paragios, N., Tziritas, G., 2011. Mrf energy minimization and beyond via dual decomposition. IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (3), 531–552.

Lan, X., Huttenlocher, D., 2005. Beyond trees: Common-factor models for 2d human pose recovery. In: International Conference on Computer Vision.

Park, D., Ramanan, D., Fowlkes, C., 2010. Multiresolution models for object detection. In: European Conference in Computer Vision.

Pishchulin, L., Jain, A., Andriluka, M., Thormaehlen, T., Schiele, B., 2012. Articulated people detection and pose estimation: Reshaping the future. In: Computer Vision and Pattern Recognition.

Polyak, B. T., 1967. A general method for solving extremum problems. Soviet Math 8 (3).

Ramanan, D., 2006. Learning to parse images of articulated bodies. In: Neural and Information Processing Systems.

Ramanan, D., Forsyth, D., Zisserman, A., 2007. Tracking people by learning their appearance. IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (1), 65–81.

Sapp, B., Taskar, B., 2013a. Modec: Multimodal decomposable models for human pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3674–3681.

Sapp, B., Taskar, B., 2013b. Modec: Multimodal decomposable models for human pose estimation. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. IEEE, pp. 3674–3681.

Sapp, B., Toshev, A., Taskar, B., 2010. Cascaded models for articulated pose estimation. In: European Conference in Computer Vision.

Sapp, B., Weiss, D., Taskar, B., 2011. Parsing human motion with stretchable models. In: IEEE Conference on Computer Vision and Pattern Recognition.

Shotton, J., Fitzgibbon, A. W., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A., 2011. Real-time human pose recognition in parts from single depth images. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1297–1304.

Singh, V. K., Nevatia, R., Huang, C., 2010. Efficient inference with multiple heterogeneous part detectors for human pose estimation. In: European Conference in Computer Vision.

Tran, D., Forsyth, D., 2010. Improved human parsing with a full relational model. In: European Conference on Computer Vision.

Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y., 2005. Large margin methods for structured and interdependent output variables. Journal of Machine Learning Research 6, 1453–1484.

Wang, H., Koller, D., 2011. Multi-level inference by relaxed dual decomposition for human pose segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition.

Wang, Y., Mori, G., 2008. Multiple tree models for occlusion and spatial constraints in human pose estimation. In: European Conference in Computer Vision.

Wang, Y., Tran, D., Liao, Z., 2011. Learning hierarchical poselets for human parsing. In: IEEE Conference on Computer Vision and Pattern Recognition.

Yang, Y., Ramanan, D., 2011. Articulated pose estimation with flexible mixtures-of-parts. In: IEEE Conference on Computer Vision and Pattern Recognition.

Yang, Y., Ramanan, D., 2013. Articulated human detection with flexible mixtures of parts. IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (12), 2878–2890.

Yu, J., Wang, M., Tao, D., 2012. Semisupervised multiview distance metric learning for cartoon synthesis. IEEE Transactions on Image Processing 21 (11), 4636–4648.

Zhou, H., Wallace, A. M., Green, P. R., 2009. Efficient tracking and ego-motion recovery using gait analysis. Signal Processing 89 (12), 2367–2384.

Zhu, L., Chen, Y., Lu, Y., Lin, C., Yuille, A. L., 2008. Max margin AND/OR graph learning for parsing the human body. In: IEEE Conference on Computer Vision and Pattern Recognition.

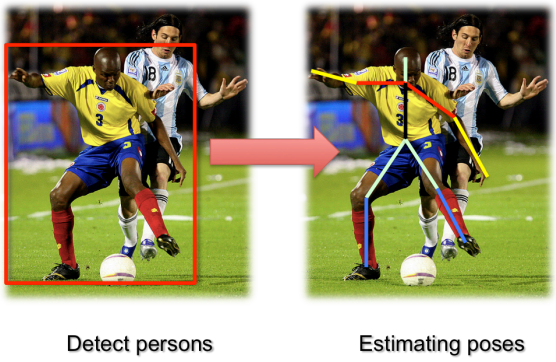

Figure 1: We consider the problem of detecting humans and estimating their body poses in 2D static images.

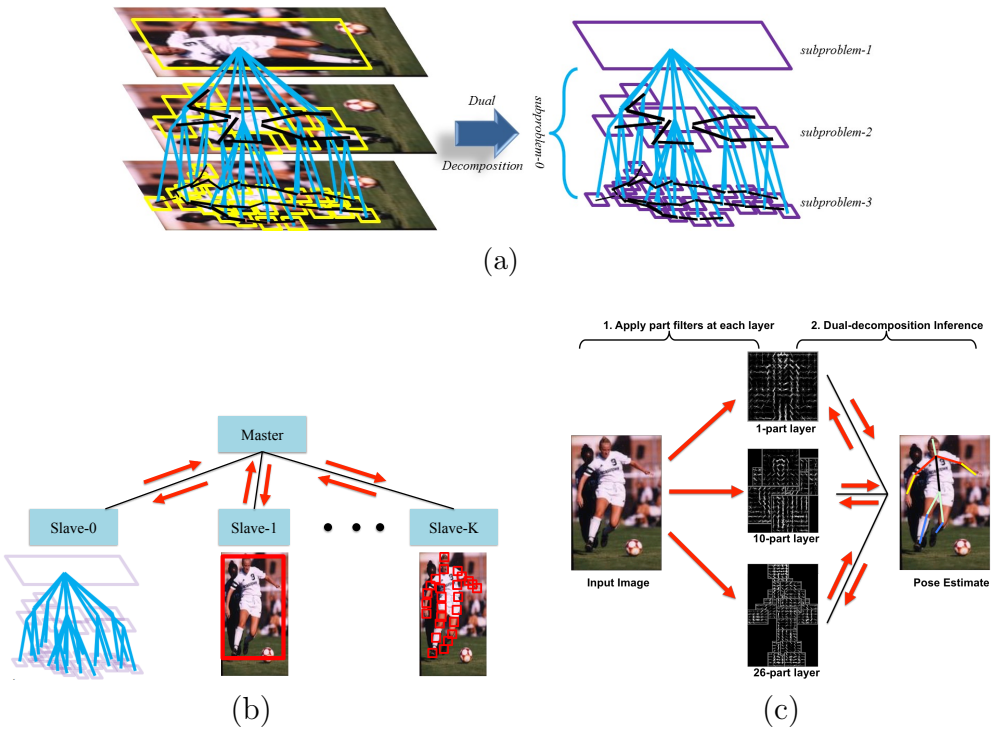

Figure 2: Illustrations of our multi-layer part-based model for human pose recognition and our efficient inference algorithm based on dual decomposition. *(a):* The model consists of multiple layers of tree-structured models (black lines), each with a different structure and a different resolution. The layers are connected together with a cross-layer tree model (blue lines). *(b):* Inference based on dual decomposition consists of iteratively passing messages between the Master problem (primal objective) and slave problems (decomposed dual objectives) until convergence. *(c):* We apply dual decomposition to human pose recognition by running HOG-based part appearance models at multiple scales on the image, and then performing inference on the tree-structured model for each layer as well as the cross-layer tree-structured model. These inference steps are repeated iteratively until the pose variable values converge.

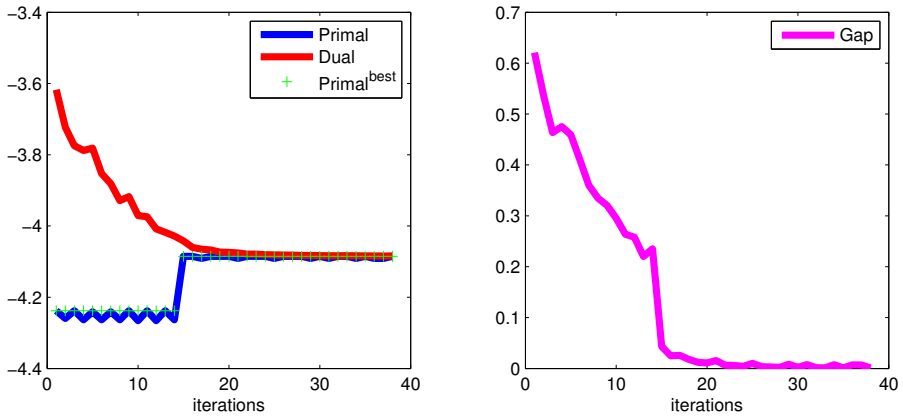

Figure 3: Primal objective and dual objective (left) and primal-dual gap (right) as a function of number of iterations during subgradient descent.

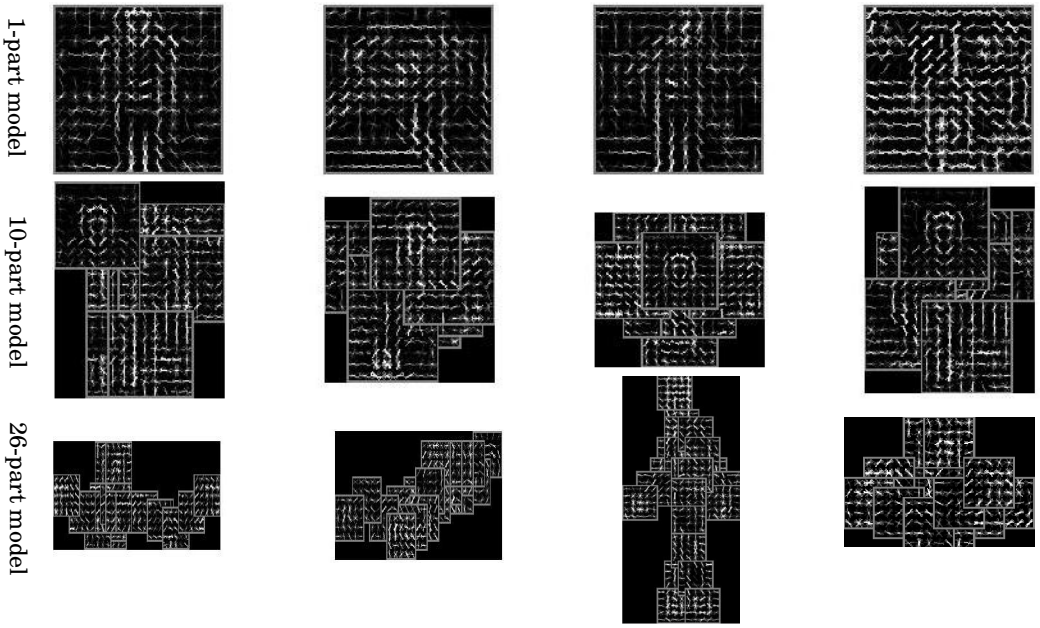

Figure 4: Part-based models used in our multi-layer composite model. For each layer (row) of the composite model, we show four randomly-chosen mixture components.

Figure 5: Illustration of the two measurements involved in evaluating body part localizations (see text). The first measure of PCP considers a part correctly localized if both distances are below a threshold, whereas the second measure considers a part correctly localized if their *mean* is below a threshold.

Figure 6: Sample results. **(Top):** Examples in which Yang and Ramanan (2011) failed, but our 3-level model estimated poses correctly. **(Bottom):** Some failure cases of our model.

	Torso	Upper legs	Lower legs	Upper arms	Lower arms	Head	Total
				Parse dataset			
Ramanan *et al.*	52.1	37.5	31.0	29.0	17.5	13.6	27.2
Yang *et al.*	82.9	69.0	63.9	55.1	35.4	77.6	60.7
Ours (**26+10**)	82.0	**72.4**	**67.8**	55.6	**36.6**	79.0	62.6
Ours (**26+1**)	**85.6**	71.7	65.6	**57.1**	**36.6**	**80.4**	**62.8**
Ours (**26+10+1**)	81.0	71.7	67.6	55.9	36.3	79.5	62.3
Pishchulin *et al.* *	88.8	77.3	67.1	53.7	36.1	73.7	63.1
Johnson *et al.* (2011)*	87.6	74.7	67.1	67.3	45.8	76.8	67.4
				UIUC Sport dataset			
Ramanan *et al.*	28.7	7.3	19.2	7.5	20.6	12.9	15.1
Wang *et al.*	75.3	49.2	39.5	25.2	11.2	47.5	37.3
Yang *et al.*	85.3	61.3	55.5	49.7	35.5	73.5	56.3
Ours (**26+10**)	85.4	61.6	**57.9**	49.1	34.8	72.9	56.4
Ours (**26+1**)	86.0	**62.2**	57.5	**51.0**	**36.3**	73.7	**57.3**
Ours (**26+10+1**)	**86.2**	61.2	55.7	49.9	35.9	**73.8**	56.5
				Leeds Sport Pose dataset			
Johnson *et al.* (2010)	78.1	**65.75**	**58.8**	47.4	**32.85**	62.9	55.1
Yang *et al.*	86.0	62.4	55.25	48.6	31.65	80.0	56.2
Ours (**26+10**)	86.2	64.0	57.3	47.6	31.9	79.8	56.6
Ours (**26+1**)	**86.9**	65.3	58.3	**48.9**	32.3	**80.5**	**57.7**
Ours (**26+10+1**)	86.2	64.1	57.4	47.9	31.9	80.0	56.9
Johnson *et al.* (2011)*	88.1	74.5	66.5	53.7	38.9	74.6	62.7
				FLIC dataset			
Yang *et al.*	48.61	–	–	13.59	14.98	1.98	17.9
Ours (**11+6**)	50.71	–	–	13.39	14.02	2.24	17.93
Ours (**11+1**)	53.56	–	–	13.64	15.05	2.71	19.44
Ours (**11+6+1**)	**56.34**	–	–	**14.57**	**16.87**	**3.80**	**21.1**

Table 1: Pose estimation results (PCP) on Parse, UIUC Sport, Leeds Sport, and FLIC datasets. We test three variants of our models: **26+1** is a 2-layer model consisting of 26-part and 1-part submodels; **26+10** is a 2-layer model consisting of 26-part and 10-part submodels; **26+10+1** is a 3-layer model consisting of 26-, 10-, and 1-part submodels. All PCP scores here use criterion 1A (see text for details); for consistency, we re-computed the results from Yang and Ramanan (2011) to use this criterion, and for Ramanan (2006) we use the re-computed statistics reported in Wang et al. (2011). *Pishchulin et al. (2012) and Johnson and Everingham (2011) are not directly comparable because they use additional training data with more annotations.

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

Name	Notes
$\mathcal{P}$	the set of all parts of a human body
$\mathcal{E}$	the set of all edges that connect different body parts in $\mathcal{P}$
$\mathbf{y}_p$	a vector that represents the configuration of the part p
$\mathbf{y}$	a vector that represents the configuration of all body parts
$S(I, \mathbf{y})$	the scoring function given the configuration $\mathbf{y}$ of the pose model and the input image I
$D(I, \mathbf{y}_p)$	the score of the part template for $\mathbf{y}_p$ applied on a particular image I
$L(\mathbf{y}_p, \mathbf{y}_q)$	a relative location term measuring agreement between two parts p and q
$T(\mathbf{y}_p, \mathbf{y}_q)$	a table for part type co-occurrence between any two parts
$\mathcal{M}$	defined as $\{(\mathcal{P}_1, \mathcal{E}_1), ..., (\mathcal{P}_K, \mathcal{E}_K)\}$, a set of K tree-structured models
$\mathbf{y}^k$	the configuration of the parts in the k-th model
$\mathbf{Y}$	defined as $(\mathbf{y}^1, ..., \mathbf{y}^K)$, the configuration of the entire multi-layer composite model
$S_k(\cdot, \cdot)$	scoring function of a single layer model $(\mathcal{P}_k, \mathcal{E}_k)$
$\chi(\mathbf{y}^k, \mathbf{y}^{k+1})$	the cross-model scoring function that measures the compatibility of the estimated configurations between adjacent layers
$B(\mathbf{y}_p^k, \mathbf{y}_q^{k+1})$	a look-up table that measures the likelihood of the relative configuration of a part and its child across two adjacent layers
$\mathcal{C}_k$	the set of all feasible (discrete) values for $\mathbf{y}^k$ for each layer of the composite model.
λ_k	the Lagrangian multiplier that specifies the strength of the penalty for submodel k in dual-decomposition
$\mathbf{X}$	duplicate or copy variables of $\mathbf{Y}$ (used in dual-decomposition)

Table 3: List of variables and their explanations.