Exploring the Spatial Hierarchy of Mixture Models for Human Pose Estimation

Yuandong Tian[1], C. Lawrence Zitnick[2], and Srinivasa G. Narasimhan[1]

[1] Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213
{yuandong, srinivas}@cs.cmu.edu
[2] Microsoft Research, One Microsoft Way, Redmond WA 98052
larryz@microsoft.com

Abstract. Human pose estimation requires a versatile yet well-constrained spatial model for grouping locally ambiguous parts together to produce a globally consistent hypothesis. Previous works either use local deformable models deviating from a certain template, or use a global mixture representation in the pose space. In this paper, we propose a new hierarchical spatial model that can capture an exponential number of poses with a compact mixture representation on each part. Using latent nodes, it can represent high-order spatial relationship among parts with exact inference. Different from recent hierarchical models that associate each latent node to a mixture of appearance templates (like HoG), we use the hierarchical structure as a pure spatial prior avoiding the large and often confounding appearance space. We verify the effectiveness of this model in three ways. First, samples representing human-like poses can be drawn from our model, showing its ability to capture high-order dependencies of parts. Second, our model achieves accurate reconstruction of unseen poses compared to a nearest neighbor pose representation. Finally, our model achieves state-of-art performance on three challenging datasets, and substantially outperforms recent hierarchical models.

1 Introduction

Human pose estimation is a challenging task in computer vision with numerous practical applications. For a scenario using a single image, the goal is to estimate the location of each human part. To avoid the curse of dimensionality, one popular approach is to build individual detectors for each human part. Spatial reasoning between parts is then utilized to filter the often noisy responses of the individual detectors. It is critical to design the spatial model so that it captures a versatile yet plausible set of poses.

The influential work of Felzenszwalb et al. [5] uses pictorial structures (PS) [7] to efficiently capture the pairwise spatial relationships between nearby parts. The resulting structure forms a tree allowing for efficient inference. A disadvantage of [5] is that it only allows for small deformations from a fixed template. To solve this problem, [8] used a (global) mixture of pictorial structures to capture greater variations in pose. However, since the number of plausible human poses is exponential, the number of parameters that need to be estimated is prohibitive without a large dataset or part sharing mechanism.

Recently, [17] treats each part rather than the entire body as a mixture of templates while modeling their pairwise relationships. As a result, it offers a compact way to represent exponentially many poses with shared parameters. Unfortunately, their use of pairwise relationships fails to capture the complex characteristics of pose space. Other works [16, 2, 14] model the high-order relationship among parts with high-order cliques. However, they either use approximate inference or heuristic search with worse-case exponential time complexity.

As first proposed by David Marr [11], one way to introduce high-order relationships without losing the benefit of efficient tree-based inference is to build a hierarchical structure with latent nodes. In this paper, we incorporate both the hierarchical structure with latent nodes and part mixtures from [17]. This enables us to model both the high-order spatial relationship among parts, and to capture an exponential number of plausible poses.

We train the hierarchical model shown in Fig. 1 with a max-margin framework and explore the obtained model in two ways: the visual quality of pose samples and the reconstruction error of recovered poses from labeled testing datasets. As far as we know, we make the first attempt to explore the pose space specified by the trained model, and conduct a more thorough analysis compared to previous work [1] that only shows a few pose samples. For model exploration, *first*, we sample the pose mixture (type) from the learned model and reconstruct the poses to judge their realism. Compared to the samples from pairwise tree models [17], our samples appear more human-like and natural. This demonstrates the capability of our model at capturing global and high-order relationships of poses rather than just local information. In addition, our model is able to build a smooth pose trajectory between two pre-defined poses, showing its potential for human motion synthesis. *Second*, the reconstruction error of a given pose using our method is lower than nearest neighbor, the lower error bound for methods that model pose space as a mixture. This shows that our method captures a large variation of poses compactly and generalizes well.

Many recent works [16, 13] also use hierarchical models for pose estimation. In particular, [13] propose part types that can be shared among different configurations. In their settings, each mixture component of a latent node also corresponds to one HoG template, modeling the image appearance covered by the descendants of that node. While their main focus is detection, we argue that this is not the ideal strategy for pose estimation, since appearance could vary exponentially with respect to the number of parts, especially for a latent top node (e.g. root node) and cannot be captured by a few mixtures. In our model, only the leaf nodes receive image evidence. The latent nodes handle only geometric deformation and compatibility between parent/child types. Therefore, the number of poses that our model can capture is not the number of mixtures of the root node, but instead the product of mixtures of all nodes.

In terms of numerical evaluation, the performance of our model is on par with the state-of-art on three benchmark datasets (PARSE [12], Leeds Sports Dataset [8] and UIUC people [15]). Besides, we also show substantial improvement ($\sim 45\%$) over recent work [16] that builds a hierarchical loopy model.

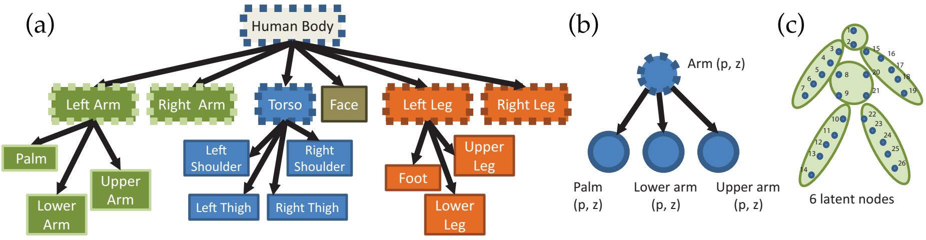


Fig. 1. **(a)** The hierarchical model for human pose estimation. All nodes with dashed boundaries are latent variables that are on top of body parts (leaves). **(b)** The latent tree model. For any part j, we want to estimate its location $\mathbf{p}_j = (x_j, y_j)$ in the image, and its type z_j. Each type of the object specifies a certain configuration and appearance of that object, as shown in Fig. 2(a). **(c)** The 6 latent and 26 leaf nodes in our three-layered hierarchy.

2 Related Work

While several works [4, 1] focus on designing better part detectors, in this work, we focus solely on the spatial relationships between parts obtained using the standard HoG features. Following the work of Felzenszwalb et al. [5], there have been many extensions. An iterative approach was taken in [12] to capture high-order color information of the human body to the model. Poselets [3] model strongly correlated appearance across nearby parts, and uses a discriminative Hough voting to localize the object. A hierarchical loopy model with poselets is built in [16]. Going further, a complete graph for part spatial relationship is proposed in [2]. To model poses with large variations, [8] builds a cluster of pictorial structures, with each cluster component responsible for a particular pose. Similarly, [18] uses a set of hierarchical models for each object view for better objection detection. Recently, [17] regarded each part as a mixture of templates and models the pairwise spatial relationship between mixtures on a tree structure. This encodes exponentially many poses with a few mixtures at each node. In this work, we build a hierarchical tree model with a mixture built on both leaves and intermediate (latent) parts. This allows us to handle both large variation of poses, model high-order relations among parts, and performs efficient inference. Our work differs from [18] whose hidden nodes of hierarchy do not have a mixture and also differs from [16] which contains loops and where each latent part is also associated with the appearance of the image.

Computational efficiency plays a critical role in designing the pictorial structure. A tree model [5, 8, 17] yields fast and exact inference, but it may not be able to encode sufficient constraints to avoid problems such as double-counting arms and legs. High-order models can be solved only with approximate inference [16] or Branch-and-Bound search [14] with no guarantee on the quality of solution or the time complexity. To model high-order interaction with tractable computational time, recently [10] proposes to use one (global) latent node to model a mixture of different poses. Following this thought, our work can be regarded as an extension that uses a hierarchy of latent nodes.

3 The Hierarchical Model

3.1 Overview

We use a hierarchical tree to represent the articulation of human pose, as shown in Fig. 1. In this hierarchy, there are two sets of nodes, the *leaf nodes* (with solid outlines) and the *latent nodes* (with dashed outlines). Each leaf node is the primitive body part (i.e., lower arm, upper arm, palm) that has been manually labeled. Each latent node covers a subset of primitive parts that are *spatially nearby* (i.e. left arm is a latent node that covers lower and upper arms and palm). Finally, the root node represents the entire human body. All nodes follow a tree structure and standard dynamic programming leads to efficient inference.

Different from many previous works [1, 5] and similar to [17], each node V_j has a **type variable** z_j in addition to the location variable $\mathbf{p}_j = (x_j, y_j)$. Here (x_j, y_j) is the center location of each part and z_j is a discrete variable that represents the mixture nature of each node.

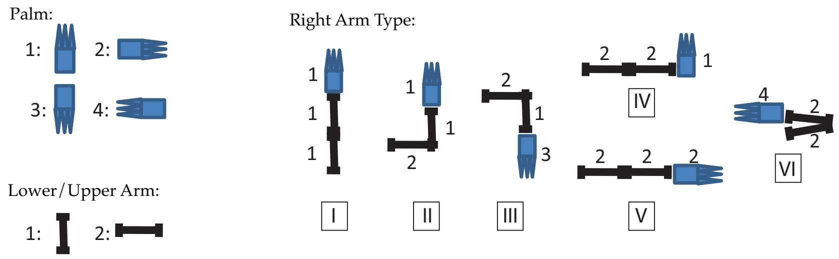
For a *leaf node* such as the palm, its type variable z_j means the appearance of that node could be different, i.e., open/close palm, vertical/horizontal arm (Fig. 2(a)). We use a discrete type variable rather than a continuous transformation of a simple template (e.g. rotation/scaling transformation as in [1]), since these transformations usually cannot capture complicated appearance changes, and it is not always necessary to enumerate all rotation/scales that are rare in both training and test sets.

For a *latent node* like arm, its type variable z_j means that both the *spatial configuration* and the *preferred types* of its children could be different for different hidden state specified by z_j (Fig. 2(b)). Furthermore, for each type z_j of the parent, one could specify how often each child type z_k appears using the *compability* term $\theta^c_{jk}(z_j, z_k)$. As we will show in Section 3.2, this modeling offers two benefits, **(a)** enabling sharing of child apperance models and **(b)** introducing high-order relationships among nearby parts.

3.2 Compatibility between Parent and Child Nodes

Appearance Sharing: In our model, appearance can be shared among different latent node (parent) types. For example, templates of open/close hand can be shared in both upright and side-way straight arms. Furthermore, our model can specify what kind of sharing is allowed and what is prohibited (e.g. upright arm cannot have a horizontal lower arm). Such information is encoded in the compability term $\theta^c_{jk}(z_j, z_k)$, which is a function between parent type z_j and child type z_k.

Several illustrative examples are shown in Fig. 2. Our model may represent the constraint “an upright arm cannot contain a horizontal lower arm” by setting $\theta^c_{jk}(z_j = \text{upright}, z_k = \text{horizontal lower arm}) = -\infty$, meaning the child and the parent states are incompatible. Similarly, our model may represent “the hand of an upright arm could be open or closed, facing the camera or not”, by



(a) Leaf nodes

(b) Latent nodes

Fig. 2. (a) The model of hand and arm segments and their associated types. **(b)** An illustration of six potential latent configurations (types) of an arm. Note that for different type of arms, the relative locations of the child nodes are different, while the appearance model of upper/lower arms and palms can be shared.

setting $\theta^c_{jk}(z_j = \text{upright}, z_k = \text{open hand}) = \theta^c_{jk}(z_j = \text{upright}, z_k = \text{close hand})$, meaning they are equally probable.

On the other hand, appearance sharing in [17] is more restricted. In their model, an open palm pointing up in an upright arm cannot be shared with an open palm in a side-way straight arm (Fig. 2(b), type I vs. type IV). This is because by Eqn. 4 in [17], the predicted relative location of the lower arm is determined by the type of palm. If these two open palms are shared (assigned the same type), then the relative location of the lower arm is forced to be the same, which is not the case in general.

Modeling high-order part relationships: With the compatibility term $\theta^c_{jk}(z_j, z_k)$, one can also model high-order relationships between multiple children. Indeed, by having a common latent parent with large number of mixtures, it is possible to model any joint distribution of children's types and locations.

3.3 Objective Function

We formulate the objective function for the hierarchical model (Fig. 1) as follows. For convenience, we still use the parent part node V_j "arm" and its child node V_k "palm" as an example.

Pairwise term. Part V_j has variables $(\mathbf{p}_j, z_j)$ and its child has variables $(\mathbf{p}_k, z_k)$. The type z_j can take a few discrete values, including "upright", "bent", "straight" and so on. Given the type z_j and the location $\mathbf{p}_j$ of "arm", we can predict the location $\tilde{\mathbf{p}}_k$ of its child V_k as:

$$\tilde{\mathbf{p}}_k = \mathbf{p}_j + \delta \mathbf{p}_{jk}(z_j) \tag{1}$$

Then, given a configuration (location and type) of V_j and V_k, we define the following pairwise score $\theta_{jk}(V_j, V_k)$:

$$\theta_{jk}(V_j, V_k) = \theta^d_{jk}(\tilde{\mathbf{p}}_k, \mathbf{p}_k, z_k) + \theta^c_{jk}(z_j, z_k) \tag{2}$$

where, $\theta^c_{jk}(z_j, z_k)$ is the compatibility term as discussed in Section 3.2, and

$$\theta^d_{jk}(\tilde{\mathbf{p}}_k, \mathbf{p}_k, z_k) = -a_k(z_k) dist(\tilde{\mathbf{p}}_k, \mathbf{p}_k) \tag{3}$$

is the *deformation* term that computes the *negative* distance between $\tilde{\mathbf{p}}_k$ predicted using parent location and type, and the location $\mathbf{p}_k$ of V_k. Here $dist(\cdot, \cdot)$ is simply the squared Euclidean distance (or any distance that enables distance transform), and $a_k(z_k)$ is the type-specific weight to be learned.

Unary term. For each leaf V_k, denote $\phi(I, \mathbf{p}_k)$ as the HoG feature extracted from location $\mathbf{p}_k$ of image I, then the unary term $\theta_k(V_k)$ is defined as $w_k(z_k)^T \phi(I, \mathbf{p}_k)$, the inner product betwen the feature and a type-dependent mask. Note that w_k is a function of z_k, meaning that there is a different template for a different type of the part.

Finally, the entire objective function to be maximized is the following:

$$J(V) = \sum_j \sum_{k \in ch(j)} \theta_{jk}(V_j, V_k) + \sum_{k \in leaves} \theta_k(V_k) \tag{4}$$

For an image I, we estimate the best location and type $(\mathbf{p}_k, z_k)$ for all parts.

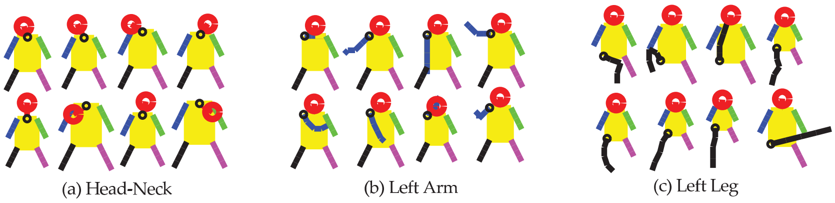


Fig. 3. Different spatial configurations (type z) of each hidden node (head-neck, left arm, left leg) in our hierarchical model learnt from Leed Sports dataset. The black circle indicates the child parts of that hidden node.

4 Training and Inference

The tree structure of our three-layered hierarchy is shown in Fig. 1(c). Given training images with groundtruth location $\mathbf{p}_k$ for leaf node k, for each latent part j (e.g. arm), we first pick one of its child's location as its location $\mathbf{p}_j$ (e.g. shoulder). Then we concatenate the relative spatial configuration $\mathbf{p}_k - \mathbf{p}_j$ of all its children in a vector, cluster them into groups using k-mean, and estimate the parent-child offsets $\delta\mathbf{p}_{jk}(\cdot)$ accordingly. Example clusters are shown in Fig. 3. Similarly, we can also build a four-layered hierarchy by subdividing the set of leaf nodes into 2 further subsets.

Given the parent-child offsets, we follow the standard max-margin paradigm and use latent SVM to discriminatively and jointly train the hierarchical structure, for both the part detector and the weights for the hierarchical model. Denote $\{I_i\}$ as a set of training images, $\{y_i\}$ as labels indicating whether an image contains a human ($y_i = 1$) or not ($y_i = -1$). The formulation then follows:

$$\begin{aligned} \min_{\mathbf{w},\xi_i} \quad & \tfrac{1}{2}\mathbf{w}^T\mathbf{w} + C\textstyle\sum_i \xi_i \\ \text{s.t. } \max\nolimits_{V^i} & \; y_i \mathbf{w}^T \varPhi(I^i, V^i) \geq 1 - \xi_i \\ & \xi_i \geq 0, a_k(\cdot) \geq 0, \end{aligned} \tag{5}$$

where, $\mathbf{w}$ is the concatenation of the parameters $w_k(\cdot)$, $a_k(\cdot)$ and $\theta^c_{jk}(\cdot,\cdot)$, and $\varPhi(I^i, V^i)$ are the overall features extracted from training sample i at given part locations and types specified by V^i, including the HoG templates of leaf parts. $\delta\mathbf{p}_{jk}(\cdot)$ is fixed during training. Note the weights $a_k(\cdot)$ for the distance have to be non-negative, which can be easily enforced in the primal-dual optimization procedure as borrowed from [17].

For latent SVM, initialization is important to achieve good performance. Following the implementation of [17], firstly the HoG weights for each part are independently trained given the labeled locations of each part, and concatenated as the initialization of $\mathbf{w}$. Compatibility weights $\theta^c_{jk}(\cdot,\cdot)$ are set to uniform and deformation weights $a_k(\cdot)$ are set to 0.01. Then, Eqn. 5 is optimized. For all our experiments, we set the regularization constant $C = 0.02$. Like [17], we use 1218 images without human in INRIA Person dataset as negative samples. See our project website for more information.

Since the entire hierarchical model is a tree structure, exact inference follows with the standard message passing algorithm. For node V_j, the incoming message $m_{k\to j}(\mathbf{p}_j, z_j)$ and the outgoing message $m_j(\mathbf{p}_j, z_j)$ are computed as:

$$m_j(\mathbf{p}_j, z_j) = \sum_{k\in ch(j)} m_{k\to j}(\mathbf{p}_j, z_j) \tag{6}$$

$$m_{k\to j}(\mathbf{p}_j, z_j) = \max_{z_k}\left[\max_{\mathbf{p}_k} m_k(\mathbf{p}_k, z_k) + \theta^d_{jk}(\tilde{\mathbf{p}}_k(z_j), \mathbf{p}_k, z_k)\right] + \theta^c_{jk}(z_j, z_k) \tag{7}$$

where, $\theta^d_{jk}(\tilde{\mathbf{p}}_k(z_j), \mathbf{p}_k, z_k) = -a_k(z_k)dist(\tilde{\mathbf{p}}_k, \mathbf{p}_k)$ ($\tilde{\mathbf{p}}_k(z_j)$ defined in Eqn. 1). Note that if the parameter $a_k(z_k)$ is only dependent on z_k, the message $m^{dt}_k(\cdot, z_k)$ can be shared during inference among different z_j after applying the distance transform making the inference procedure much faster,

$$m^{dt}_k(\cdot, z_k) = \max_{\mathbf{p}_k} m_k(\mathbf{p}_k, z_k) - a_k(z_k)dist(\cdot, \mathbf{p}_k). \tag{8}$$

5 Experiments

Datasets: We use three benchmark datasets for evaluation: PARSE dataset [12], Leeds Sports Dataset (LSP) [8] and UIUC people [15]. PARSE dataset contains 305 images with 100 for trainings and 205 for testing. Leeds Sports Dataset has 1000 training images and 1000 test images and show a large variation of pose changes. UIUC people dataset contains 346 for training and 247 for testing. Similar to previous works, we use the criterion proposed in [6] for performance evaluation, i.e., a part is regarded as correctly identified if both its end-points are within 50% of the labeled segment length from their true locations.

5.1 Exploring the Hierarchical Model

We first study how well our hierarchical model, once trained on a dataset, can represent the spatial configuration of the human body. For this, we (1) sample spatial configurations from the model and observe whether the samples are human-like, and (2) reconstruct a given spatial configuration using the learned model, and check the reconstruction error.

These two operations show complementary effects of a spatial model. A weakly constrained spatial model, due to its flexibility, may perform extremely well on the reconstruction task, but once sampled, may generate poses that are not human-like. On the other hand, a model with strong spatial prior (e.g., tranditional pictorial structure [5]) will generate human-like poses with small variations. But it may fail to reconstruct rare poses accurately. As we shall show, our model achieves a balance between the two criteria.

Sampling. To sample the model, we omit the unary potentials (image evidence) and the deformation score $\theta^d_{jk}(\tilde{\mathbf{p}}_k(z_j), \mathbf{p}_k, z_k)$, and only use the binary potentials $\theta^c_{jk}(z_j, z_k)$ between parents and children. The truncated score function becomes:

$$J_{\text{trunc}}(\mathbf{Z}) = \sum_j \sum_{k \in ch(j)} \theta^c_{jk}(z_j, z_k) \tag{9}$$

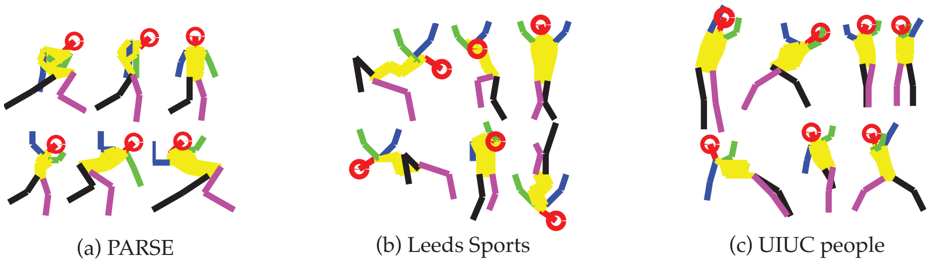
where $\mathbf{Z}$ is the collection of all type variables of all nodes. Once samples of $\mathbf{Z}$ are obtained, Eqn. 1 is used to deterministically generate the human poses in a top-down manner.

To sample the type variables so that those with high score will be drawn more often, we assign a probablistic distribution $p(Z|T) \propto \exp(J(Z)/T)$, where T is the temperature that controls how scattered the samples are. The case $T \rightarrow +\infty$ wipes out the score variations in the pose space and outputs a uniform distribution, while the case $T \rightarrow 0$ samples only the best poses.

Since the hierarchical model follows a tree structure, exact sampling is possible. We start with all the leaf nodes and compute the marginal distribution of their parents, all the way up to the root node. Then we sample the root node, and then conditionally sample its children and grand children, until leaves are sampled. We also sample conditionally given the type of an arbitrary node, by treating the fixed node as the root.

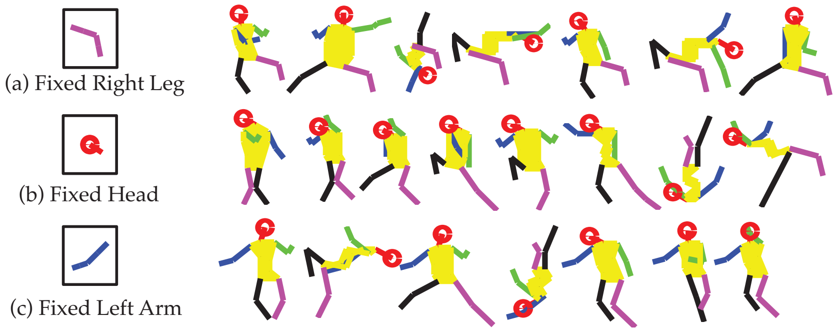
For two different models, their temperatures T_1 and T_2 have to be calibrated to show approximately the same variations in pose space. This is achieved by finding a pair of T_1 and T_2 such that their pose variations, estimated by sampling, are the same.

Fig. 6 shows that our model can generate reasonable human-like poses with large variations. In comparison, sampling from [17] often results in weird-looking poses, since they only model the local relationships between neighboring parts. See supplementary material for a video of samples along a path in the pose-space. In particular, fixing the root type and setting $T \rightarrow 0$ gives the most likely pose of that type, as shown in Fig. 4. This roughly corresponds to the pose clusters the hierarchical model can capture. However, within each cluster, significant variation is still allowed by changing the type variables below the root



(a) PARSE (b) Leeds Sports (c) UIUC people

Fig. 4. The most probable poses of our model for given root types (roughly this corresponds to a given cluster in the pose space). We can see it encodes very different poses. In particular, Leeds Sports Dataset contain more pose changes than PARSE or UIUC sports dataset.



(a) Fixed Right Leg

(b) Fixed Head

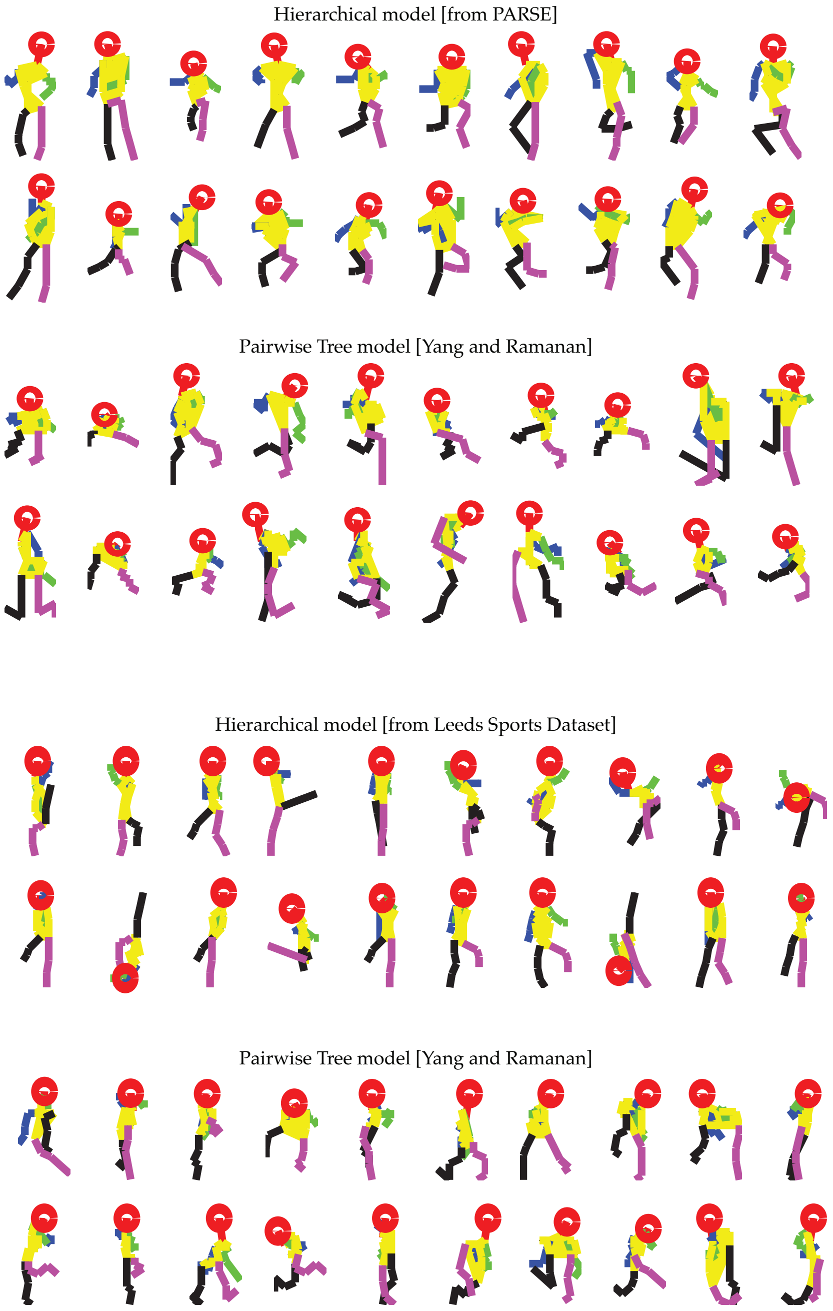
(c) Fixed Left Arm

Fig. 5. Pose sampling from the hierarchical model learned from Leeds Sports Dataset with one latent node fixed. We can see our model is able to sample human-like poses.

node. In contrast, previous works on hierarchical structure [16, 13] all associate a type variable with a template restricting the possible pose variation each node can handle. Alternatively, we can also fix one type of a latent node (e.g. leg, head, arm) and sample the remaining nodes. As shown in Fig. 5, our method gives human-like extrapolation of poses, which is not shown in previous works.

Reconstruction. For reconstruction, we take one pose from the test sample, allowing the part detectors to fire only at the groundtruth location, and run the detection procedure. Once the best detection is obtained, we can reconstruct the pose with just the type variables and remove all deformation. Such a reconstructed pose has zero deformation score. The closer the reconstructed pose is to the groundtruth pose, the better the spatial model can fit to the given pose.

Fig. 7 shows our model, once trained on PARSE, can reconstruct poses in the test set of PARSE, better than nearest-neighbor. The average root-mean-square error for our three-layered model is 7.19, and for our four-layered model it reduces to 6.55. Our error is lower than 14.63 computed from the Nearest-neighbor approach that searches for the best globally matching pose in the training set. A more flexible model [17] achieves slightly lower error (5.39) with 10 mixture com-



Hierarchical model [from PARSE]

Pairwise Tree model [Yang and Ramanan]

Hierarchical model [from Leeds Sports Dataset]

Pairwise Tree model [Yang and Ramanan]

Fig. 6. Comparison between samples drawn from our hierarchical tree model and from the pairwise tree model [17], both learned on PARSE dataset (top) and on Leeds Sports dataset (bottom). Our model captures large variation of poses (especially for Leeds Sports Dataset) and generates reasonable human-like poses. See supplementary material for more samples.

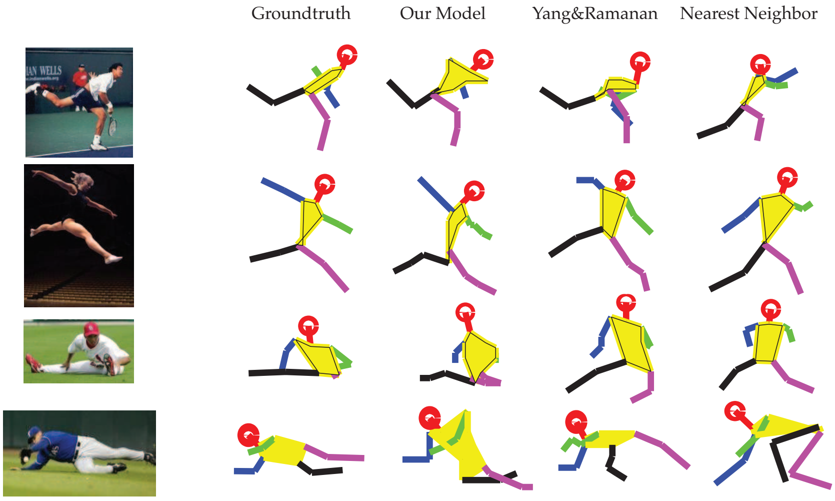


Fig. 7. Reconstruction of poses in the test set with hierarchical model. Compared with nearest neighbor and [17] trained on the same dataset, the hierarchical model can recover the pose more accurately.

ponents for each part. For their original less-flexible model with 5-6 mixtures per part, the reconstruction error is 7.47. This demonstrates that our model achieves a good balance between creating human-like poses and reconstructability.

5.2 Performance on Benchmark Datasets

For the PARSE dataset, we use 5-6 mixtures for each leaf node, and 10 mixtures for each hidden node. Our performance (in terms of PCP, the percentage of parts being correctly detected) on the test set (205 images) is 74.4%, while [17] achieves 74.9%. Fig. 8 shows that our hierarchical model can handle large pose variations and also often tackle the double-counting problem with high-order relationships among parts. Table 1 shows correctly detected ratio per part.

For the Leeds Sports dataset that contains large variations of pose, we use 15 mixtures for both leaf and hidden nodes. With only the first 200 training images, we achieve 58.8%. For 1000 training images, they are evenly partitioned into 5 disjoint training sets, on which 5 separate models are trained. We simply put the candidate detections of 5 models together to achieve 61.3%. Intuitively, this means we take the maximum score of 5 models. In comparison, previous work [8] achieves 55.2% with the same number of training samples, and [9] achieves 62.7% with 11000 training samples. Table 1 shows the per-part performance and Fig. 9 shows examplar comparisons.

For the UIUC dataset, we use 5-6 mixtures for leaf nodes and 10 mixtures for hidden nodes, which is the same as in the PARSE dataset. As a result, our hierarchical model outperforms hierarchical poselets [16] by 45% (Table 1). The

Dataset	Method	Torso	Head	U. Leg	L. Leg	U. Arm	L. Arm	Total
PARSE	JE [8]	85.4	76.1	73.4	65.4	64.7	46.9	66.2
	YR [17]	**97.6**	**93.2**	83.9	75.1	**72.0**	**48.3**	**74.9**
	Ours, 3-layer	97.1	92.2	**85.1**	**76.1**	71.0	45.1	74.4
	Ours 4-layer	96.1	92.7	81.2	71.0	69.5	39.0	71.0
Leeds	JE [8]	78.1	62.9	65.8	58.8	47.4	32.9	55.1
	Ours (first 200 training)	93.7	86.5	68.0	57.8	49.0	29.2	58.8
	Ours (1000 training, 5 models)	**95.8**	**87.8**	69.9	60.0	51.9	32.9	61.3
	JE [9] (11000 training)	88.1	74.6	**74.5**	**66.5**	**53.7**	**37.5**	**62.7**
UIUC	Wang et.al [16]	86.6	68.8	56.3	50.2	30.8	20.3	47.0
	Our method	**98.8**	**96.8**	**78.7**	**64.2**	**62.2**	**39.5**	**68.5**

Table 1. Performance in PARSE, Leeds Sports and UIUC people dataset. For each dataset, the performance of our method is on par with the state-of-art.

	PARSE [12]	Leeds Sports [8]	UIUC people [15]
PARSE [12]	74.4	53.5	64.5
Leeds Sports Dataset [8]	67.0	61.3	64.3
UIUC people [15]	63.5	53.6	68.5

Table 2. Cross performance between 3 different datasets. Each row shows a model that is trained on a single dataset (shown in left-most cell), and tested on the other datasets. We can see our model generalizes well. Note that for Leeds Sports Dataset, we train 5 models, each taking 200 training as input, and combine the candidate detections.

loopy structure and the HoG poselet over a large image region may have lead to their worse performance. See Table 1 for the per-part performance.

As shown in Table 1, a deeper hierarchy may hurt the performance since more parameters need to be trained (estimated), resulting in overfitting. However, a shallow one may fail to generalize well since one latent node will be required to encode a large number of joint configurations of children.

We also train our model on one dataset using its training set, and evaluate it on the test set of the other datasets. As shown in Table 2, our method generalizes well across different datasets. In addition, the model trained on Leeds Sports performs equally well on other datasets, showing Leeds Sports covers a variety of poses and is a better dataset for a model to train on without overfitting. PARSE is a specific dataset. Only a model trained on it performs well in test.

From our experiments, we conclude that our methods are on par with the state of the art reported for two datasets and improve significantly over other hierarchical methods. In the future, we will attempt to learn the hierarchical structure automatically from the training data. In terms of applications, we will apply this approach to tracking of human actions in videos.

Acknowledgement. This research was supported in parts by an ONR Grant N00014-11-1-0295 and a Samsung Advanced Institute of Technology grant. Yuandong Tian is supported by a Microsoft PhD fellowship.

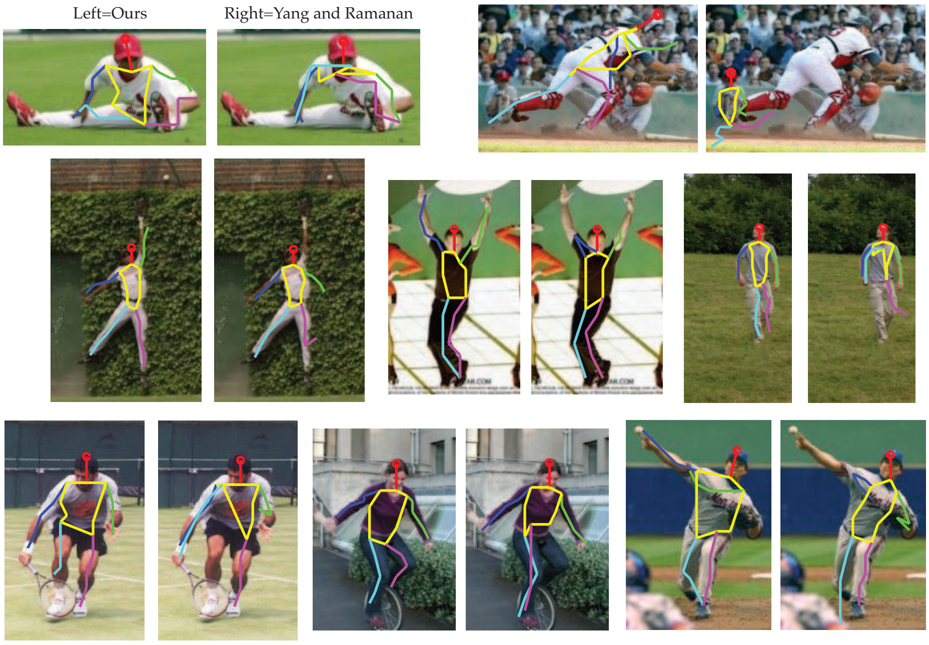


Fig. 8. Comparison between our method (Left) and [17] (Right) on some test images of PARSE dataset. Note that our model can handle large pose variation and occasionally the double counting of arms and legs.

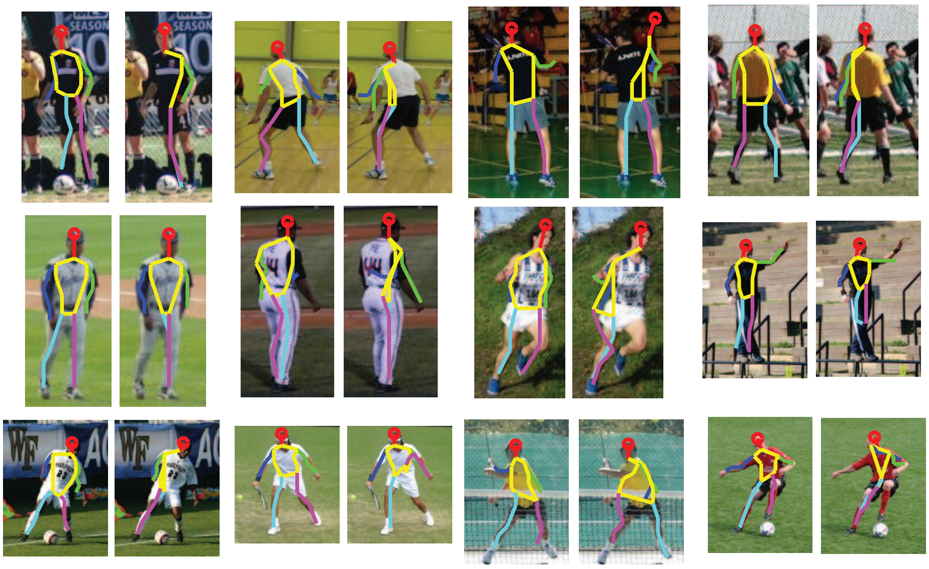


Fig. 9. Comparison between our method (Left) and [17] (Right) on Leeds Sports dataset.

References

1. M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *ter Vision and Pattern Recognition*, pages 1014–1021. IEEE, 2009.
2. M. Bergtholdt, J. Kappes, S. Schmidt, and C. Schnörr. A study of parts-based object class detection using complete graphs. *International journal of computer vision*, 87(1):93–117, 2010.
3. Lubomir Bourdev and Jitendra Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *International Conference on Computer Vision (ICCV)*, 2009.
4. M. Eichner and V. Ferrari. Better appearance models for pictorial structures. In *BMVC*, 2009.
5. P.F. Felzenszwalb and D.P. Huttenlocher. Efficient matching of pictorial structures. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 2, pages 66–73. IEEE, 2000.
6. V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.
7. M.A. Fischler and R.A. Elschlager. The representation and matching of pictorial structures. *Computers, IEEE Transactions on*, 100(1):67–92, 1973.
8. S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010.
9. S. Johnson and M. Everingham. Learning effective human pose estimation from inaccurate annotation. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1465–1472. IEEE, 2011.
10. X. Lan and D.P. Huttenlocher. Beyond trees: Common-factor models for 2d human pose recovery. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 470–477. IEEE, 2005.
11. D. Marr and H.K. Nishihara. Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 200(1140):269–294, 1978.
12. D. Ramanan. Learning to parse images of articulated bodies. In *NIPS*, volume 19, page 1129, 2007.
13. M. Sun and S Savarese. Articulated part-based model for joint object detection and pose estimation. In *ICCV*, 2011.
14. T.P. Tian and S. Sclaroff. Fast globally optimal 2d human detection with loopy graph models. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 81–88. IEEE, 2010.
15. D. Tran and D. Forsyth. Improved human parsing with a full relational model. *Computer Vision–ECCV 2010*, pages 227–240, 2010.
16. Y. Wang, D. Tran, and Z. Liao. Learning hierarchical poselets for human parsing. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1705–1712. IEEE, 2011.
17. Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1385–1392. IEEE, 2011.
18. L.L. Zhu, Y. Chen, A. Yuille, and W. Freeman. Latent hierarchical structural learning for object detection. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1062–1069. IEEE, 2010.