Learning Hierarchical Poselets for Human Parsing

Yang Wang Duan Tran Zicheng Liao
Department of Computer Science, University of Illinois at Urbana-Champaign
{yangwang, ddtran2, liao17}@uiuc.edu

Abstract

We consider the problem of human parsing with part-based models. Most previous work in part-based models only considers rigid parts (e.g. torso, head, half limbs) guided by human anatomy. We argue that this representation of parts is not necessarily appropriate for human parsing. In this paper, we introduce hierarchical poselets – a new representation for human parsing. Hierarchical poselets can be rigid parts, but they can also be parts that cover large portions of human bodies (e.g. torso + left arm). In the extreme case, they can be the whole bodies. We develop a structured model to organize poselets in a hierarchical way and learn the model parameters in a max-margin framework. We demonstrate the superior performance of our proposed approach on two datasets with aggressive pose variations.

1. Introduction

Part-based representations, such as cardboard people [11] or pictorial structure [7], provide an elegant framework for human parsing. A part-based model represents the human body as a constellation of a set of rigid parts (e.g. torso, head, half limbs) constrained in some fashion. The typical constraints used are tree-structured kinematic constraints between adjacent body parts, e.g. torso-upper half-limb connection, or upper-lower half-limb connection. Part-based models consist of two important components: (1) part appearances specifying what each body part should look like in the image; (2) configuration priors specifying how parts should be arranged relative to each other.

Considerable progress has been made in the past few years in human parsing. Most of the progress can be attributed to improvements on one of these two components. A representative example of building better part appearance models is the work by Ramanan [16], which learns color histograms of parts from an initial edge-based model. Ferrari *et al*. [8] and Eichner *et al*. [5] further improve the part appearance models by reducing the search space using various tricks, e.g. the relative locations of part locations with respect to a person detection, the relationship between different part appearances (e.g. upper-arm and torso tend to have the same color), etc. Andriluka *et al*. [1] build better edge-based appearance models using shape contexts. Sapp *et al*. [20] develop efficient inference algorithm to allow the use of more expensive features. There is also work [15, 13, 23] on using segmentation for pose estimation.

Most work on improving configuration priors focuses on developing representations and fast inference algorithms that by-pass the limitations of kinematic tree-structured spatial priors in standard pictorial structure models. Examples include common-factor models [12], loopy graphs [9, 18, 25, 26], mixtures of trees [28]. There is also work on building spatial priors that adapt to testing examples [19].

Despite of the success, there is one important issue overlooked by previous work – the basic representation of "parts". An implicit assumption made by almost all the previous approaches is that a "part" corresponds to a rigid piece of the human body that is meaningful in an anatomical sense, e.g. torso, head, half limbs. In this paper, we challenge this "rigid part assumption" and argue that it is not necessarily a good representations for human parsing. Rigid parts, usually represented as rectangles (e.g. [1, 7, 16, 22, 28]) or parallel lines (e.g. [18]), are inherently difficult to detect. They can be easily confused with rectangular shapes often found on buildings, windows, etc.

Some recent work [6] has shown tremendous success of part-based models in object detection. Our work builds on two important lessons from [6]. First, "parts" do not have to be semantically meaningful. Second, "parts" should have multi-level hierarchy to capture different granularity of details. Similar observations have been made by recent work on poselet representations for human detection [3, 2]. In this paper, we extend this line of ideas for human parsing. We use a part-based model, but our notion of "parts" can range from basic rigid parts (e.g. torso, head, half-limb), to large pieces of bodies covering more than one rigid part (e.g. torso + left arm). In the extreme case, we have "parts" corresponding to the whole body. We learn a model defined on this hierarchy of "parts". For a new image, we infer the human pose using this hierarchical representation.

Our work can be seen as bridging the gap between two popular schools of approaches for human parsing: part-based methods, and exemplar-based methods. Part-based methods, as explained above, model the human body as a collection of rigid parts. They use local part appearances to search for those parts in an image, and use configuration priors to put these pieces together in some plausible way. But since the configuration priors in these methods are typically defined as pairwise constraints between parts, these methods usually lack any notion that captures what a

person should look like as a whole. In contrast, exemplar-based methods (e.g. [14, 21, 24]) search for images with similar whole body configurations, and transfer the poses of those well-matched training images to a new image. The limitation of exemplar-based approaches is that they require good matching of the entire body. They cannot handle test images of which the legs are similar to some training images, while the arms are similar to other training images. Our work combines the benefits of both schools. On one hand, we capture the large-scale information of human pose via large parts. On the other hand, we have the flexibility to compose new poses from different parts.

2. Hierarchical Poselet - A New Representation for Human Parsing

Our pose representation is based on the concept of "poselet" introduced by Bourdev and Malik [3]. In a nutshell, poselets refer to pieces of human poses that are tightly clustered in both appearance and configuration spaces. Poselets have been shown to be effective at person detection [3, 2]. Variants of poselets have also been developed for solving other vision problems, e.g. action recognition in static images [29].

In this paper, we propose a new representation called *hierarchical poselets* for parsing human poses. Hierarchical poselets extend the original poselets in several important directions to make them more appropriate for human parsing. We start by highlighting the important properties of our representation.

Beyond rigid "parts": Most of the previous work in human parsing are based on the notion that the human body can be modeled as a set of rigid parts connected in some way. Almost all of them use a natural definition of parts (e.g. torso, head, upper/lower limbs) corresponding to body segments, and model those parts as rectangles, parallel lines, or other primitive shapes.

As pointed out in [3], this natural definition of "parts" fails to acknowledge the fact that rigid parts are not necessarily the most salient features for visual recognition. For example, rectangles and parallel lines can be found as limbs, but they can also be easily confused with windows, buildings, and other objects in the background. So it is inherently difficult to build reliable detectors for those parts. On the other hand, certain visual patterns covering large portions of human bodies, e.g. "a torso with the left arm raising up" or "legs in lateral pose", are much more visually distinctive and easier to identify. This phenomenon was observed even prior to the work of poselet and was exploited to detect stylized human poses and build appearance models for kinematic tracking [17].

Multiscale hierarchy of "parts": Another important property of our representation is that we define "parts" at different levels of hierarchy to cover pieces of human poses at various granularity, ranging from the configuration of the whole body, to small rigid parts. In particular, we define 20 parts to represent the human pose and organize them in a hierarchy shown in Fig. 1. To avoid terminological confusion, we will use "part" to denote one of the 20 parts in Fig. 1 and use "primitive part" to denote rigid body parts (i.e. torso, head, half limbs) from now on.

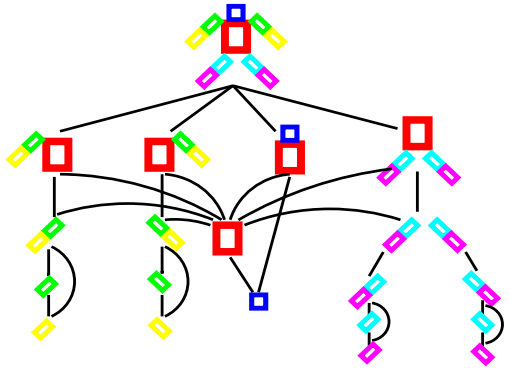

Figure 1: An illustration of the hierarchical pose representation. The black edges indicate the connectivity among different parts. The structure layout of the graph reflects the message passing scheme in Sec. 4.

We use a procedure similar to [29] to select poselets for each part. First, we cluster the joints on each part into several clusters based on their relative x and y coordinates with respect to some reference joint of that part. For example, for the part "torso", we choose the middle-top joint as the reference and compute the relative coordinates of all the other joints on the torso with respect to this reference joint. The concatenation of all those coordinates will be the vector used for clustering. We run K-means clustering on the vectors collected from all training images and remove clusters that are too small. Similarly, we obtain the clusters for all the other parts. In the end, we obtain 5 to 20 clusters for each part. Based on the clustering, we crop the corresponding patches from the images and form a set of poselets for that part. Figure 2 shows examples of two different poselets for the part "legs".

Our focus is the new representation, so we use standard HOG descriptors [4] to keep the feature engineering to the minimum. For each poselet, we construct HOG features from patches in the corresponding cluster and from random negative patches. Inspired by the success of multiscale HOG features [6], we use different cell sizes when computing HOG features for different parts. For example, we use cells of 12×12 pixel regions for poselets of the whole body, and cells of 2×2 for poselets of the upper/lower arm. This is motivated by the fact that large body parts (e.g. whole body) are typically well-represented by coarse shape information, while small body parts (e.g. half limb) are better represented by more detailed information. We then train a linear SVM classifier for detecting the presence of each poselet. The learned SVM weights can be thought as a template for the poselet. Examples of several HOG templates for the "legs" poselets are shown as the last columns of Fig. 2. Examples of poselets and their corresponding HOG templates for other body parts are shown in Fig. 3.

A poselet of a primitive part contains two endpoints. For example, for a poselet of upper-left leg, one endpoint cor-

Figure 2: Examples of two poselets for the part "legs". Each row corresponds to a poselet. We show several patches from the poselet cluster. The last column shows the HOG template of the poselet.

responds to the joint between torso and upper-left leg, the other one corresponds to the joint between upper/lower left leg. We record the mean location (with respect to the center of the poselet image patch) of each endpoint. This information will be used later when we need to infer the endpoints of a primitive part for a test image.

3. Model Formulation

We denote the complete configuration of a human pose as $L = \{l_i\}_{i=1}^{K}$, where K is the total number of parts (i.e. $K = 20$ in our case). The configuration of each part l_i is parametrized by $l_i = (x_i, y_i, z_i)$. Here (x_i, y_i) defines the image location, and z_i is the index of the corresponding poselet for this part, i.e. $z_i \in \{1, 2, ..., \mathcal{P}_i\}$, where $\mathcal{P}_i$ is the number of poselets for the i-th part. In this paper, we assume the scale of the person is fixed and do not search over multiple scales. It is straightforward to augment l_i with other information, e.g. scale, foreshortening, etc.

The complete pose L can be represented by a graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where a vertex $i \in \mathcal{V}$ denotes a part and an edge $(i, j) \in \mathcal{E}$ captures the constraint between parts i and j. The structure of $\mathcal{G}$ is shown in Fig. 1. We define the score of labeling an image I with the pose L as:

$$F(L, I) = \sum_{i \in \mathcal{V}} \phi(l_i; I) + \sum_{(i,j) \in \mathcal{E}} \psi(l_i, l_j) \quad (1)$$

The details of the potential functions in Eq. 1 are as follows.

Spatial prior $\psi(l_i, l_j)$: This potential function captures the compatibility of configurations of part i and part j. It is parametrized as:

$$\psi(l_i, l_j) = \alpha_{i;j;z_i;z_j}^{\top} \text{bin}(x_i - x_j, y_i - y_j) \quad (2a)$$

$$= \sum_{a=1}^{\mathcal{P}_i} \sum_{b=1}^{\mathcal{P}_j} \mathbb{1}_a(z_i) \mathbb{1}_b(z_j) \alpha_{i;j;a;b}^{\top} \text{bin}(x_i - x_j, y_i - y_j) \quad (2b)$$

Similar to Ramanan [16], the function bin$(\cdot)$ is a vectorized count of spatial histogram bins. We use $\mathbb{1}_a(\cdot)$ to denote the function that takes 1 if its argument equals a, and 0 otherwise. Here $\alpha_{i;j;z_i;z_j}$ is a model parameter that favors certain relative spatial bins when poselets z_i and z_j are chosen for parts i and j, respectively. Overall, this potential function models the (relative) spatial arrangement and poselet assignment of a pair (i, j) of parts.

Local appearance $\phi(l_i; I)$: This potential function captures the compatibility of placing the poselet z_i at the location (x_i, y_i) of an image I. It is parametrized as:

$$\phi(l_i; I) = \beta_{i;z_i}^{\top} f(I(l_i)) = \sum_{a=1}^{\mathcal{P}_i} \beta_{i;a}^{\top} f(I(l_i)) \cdot \mathbb{1}_a(z_i) \quad (3)$$

where $\beta_{i;z_i}$ is a vector of model parameters corresponding to the poselet z_i and $f(I(l_i))$ is a feature vector corresponding to the image patch defined by l_i. We define $f(I(l_i))$ as a length $\mathcal{P}_i + 1$ vector as:

$$f(I(l_i)) = [f_1(I(l_i)), f_2(I(l_i)), ..., f_{\mathcal{P}_i}(I(l_i)), 1] \quad (4)$$

Each element $f_r(I(l_i))$ is the score of placing poselet z_r at image location (x_i, y_i). The constant 1 appended at the end of vector allows us to learn the model with a bias term. In other words, the score of placing the poselet z_i at image location (x_i, y_i) is a linear combination (with bias term) of the responses all the poselet templates at (x_i, y_i) for part i. We have found that this feature vector works better than the one used in [29], which defines $f(I(l_i))$ as a scalar of a single poselet template response. This is because the poselet templates learned for a particular part are usually not independent of each other. So it helps to combine their responses as the local appearance model.

We summarize and highlight the important properties of our model and contextualize our research by comparing with related work.

Discriminative "parts": Our model is based on a new concept of "parts" which goes beyond the traditional rigid parts. Rigid parts are inherently difficult to detect. We instead consider parts covering a wide range of portions of human bodies. We use poselets to capture distinctive appearance patterns of various parts. These poselets have better discriminative powers than traditional rigid part detectors.

Coarse-to-fine granularity: Different parts in our model are represented by features at varying levels of details (i.e. cell sizes in HOG descriptors). Conceptually, this multi-level granularity can be seen as providing an efficient coarse-to-fine search strategy. However, it is very different from the coarse-to-fine cascade pruning in Sapp *et al*. [20]. The method in [20] prunes the search space of small parts (e.g. right lower arm) at the coarse level using simple features and apply more sophisticated features in the pruned search space. However, we would like to argue that at the coarse level, one should not even consider small parts, since they are inherently difficult to detect or prune at this level. Instead, we should focus on large body parts since they are easy to find at the coarse level. The configurations of large pieces of human bodies will guide the search of smaller parts. For example, an upright torso with arms raising up (coarse-level information) is a very good indicator of where the arms (fine-level details) might be.

Structured hierarchical model: A final important property of our model is that we combine information across different parts in a structured hierarchical way. The original work on poselets [3, 2] uses a simple Hough voting

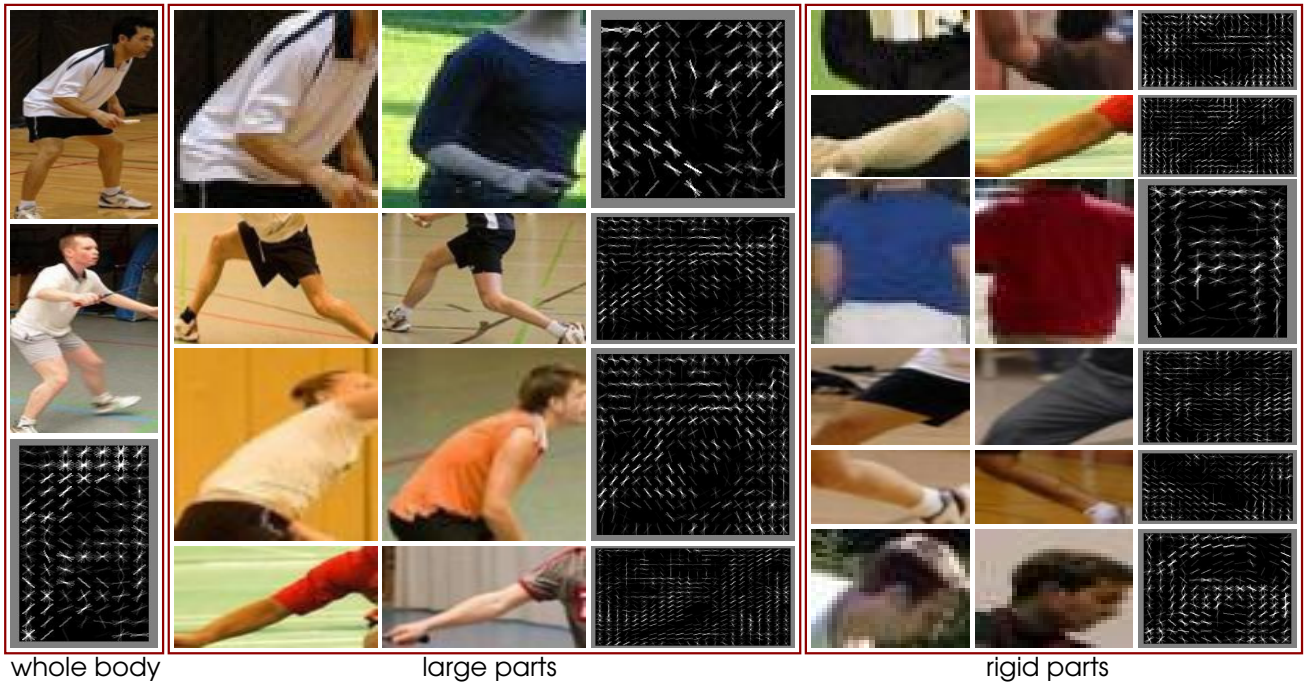

Figure 3: Visualization of some poselets learned from different body parts on the UIUC people dataset, including whole body, large parts (top to bottom: torso+left arm, legs, torso+head, left arm), and rigid parts (top to bottom: upper/lower left arm, torso, upper/lower left leg, head). For each poselet, we show two image patches from the corresponding cluster and the learned SVM HOG template.

scheme for person detection, i.e. each poselet votes for the center of the person, and the votes are combined together. This Hough voting might be appropriate for person detection, but it is not enough for human parsing which involves highly complex and structured outputs. Instead, we have developed a structured model that organize information about different parts in a hierarchical fashion. Another work that uses hierarchical models for human parsing is the AND-OR graph in [30]. But there are two important differences. First, the appearance models used in [30] are only defined on sub-parts of body segments. Their hierarchical model is only used to put all the small pieces together. As mentioned earlier, appearance models based on body segments are inherently unreliable. In contrast, we use appearance models associated with parts of varying sizes. Second, the OR-nodes in [30] are conceptually similar to poselets in our case. But the OR-nodes in [30] are defined manually, while our poselets are learned.

4. Inference

Given an image I, the inference problem is to find the optimal pose labeling L^* that maximize the score $F(L, I)$, i.e. $L^* = \arg\max_L F(L, I)$. We use the max-product version of belief propagation to solve this problem. We pick the vertex corresponding to part "whole body" as the root and pass messages upwards towards this root. The message from part i to its parent j is computed as:

$$m_i(l_j) = \max_{l_i}(u(l_j) + \psi(l_i, l_j)) \tag{5a}$$

$$u(l_j) = \phi(l_j) + \sum_{k \in \text{kids}_j} m_k(l_j) \tag{5b}$$

Afterwards, we pass messages downward from the root to other vertices in a similar fashion. This message passing scheme is repeated several times until it converges. If we temporarily ignore the poselet indices z_i and z_j and think of $l_i = (x_i, y_i)$, we can represent the messages as 2D images and pass messages using techniques similar to those in [16]. The image $u(l_j)$ is obtained by summing together response images from its child parts $m_k(l_j)$ and its local response image $\phi(l_j)$. $\phi(l_j)$ can be computed in linear time by convolving the HOG feature map with the template of z_j. The maximization in Eq. 5a can also be calculated in time linear to the size of $u(l_j)$. In practice, we compute messages on each fixed (z_i, z_j) and enumerate all the possible assignments of (z_i, z_j) to obtain the final message. Note that since the graph structure is not a tree, this message passing scheme does not guarantee to find the globally optimal solution. But empirically, we have found this approximate inference scheme to be sufficient for our application.

The inference gives us the image locations and poselet indices of all the 20 parts (both primitive and non-primitive). To obtain the final parsing result, we need to compute the locations of the two endpoints for each primitive part. These can be obtained from the mean endpoint locations recorded for each primitive part poselet (see Sec. 2).

Figure 4 shows a graphical illustration of applying our model on a test image. For each part in the hierarchy, we show two sample patches and the SVM HOG template corresponding to the poselet chosen for that part.

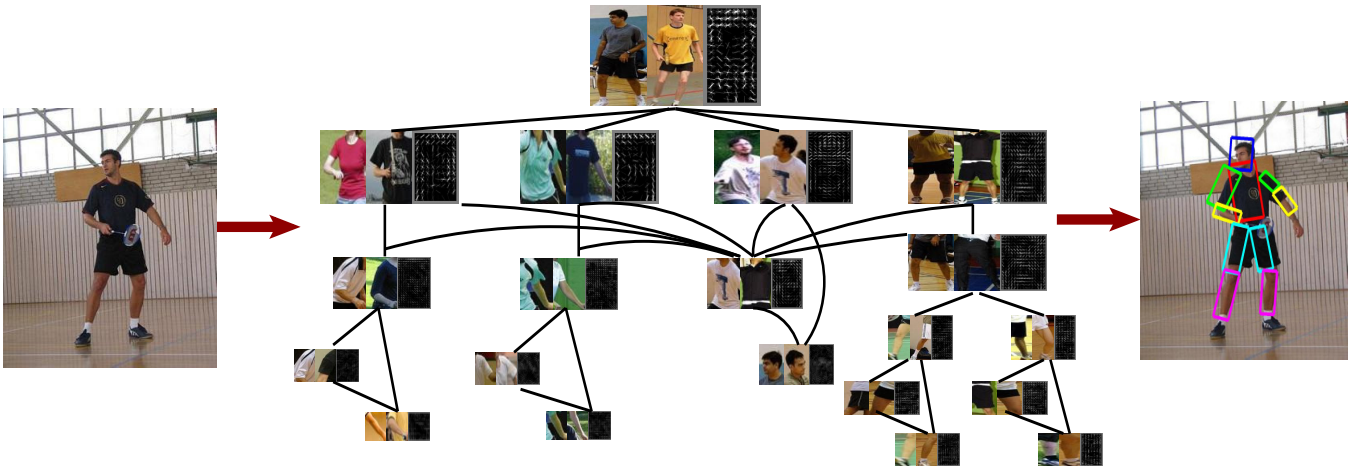

Figure 4: A graphical illustration of applying our model on a test image. For each part (please refer to Fig. 1), we show the inferred poselet by visualizing two sample patches from the corresponding poselet cluster and the SVM HOG template.

5. Learning

In order to describe the learning algorithm, we first write Eq. 1 as a linear function of a single parameter vector w which is a concatenation of all the model parameters, i.e.:

$$F(L,I) = w^\top \Phi(I,L), \quad \text{where} \tag{6a}$$
$$w = [\alpha_{i;j;a;b}; \beta_{i;a}], \quad \forall i,j,a,b \tag{6b}$$
$$\Phi(I,L) = [\mathbb{1}_a(z_i)\mathbb{1}_b(z_j)\text{bin}(x_i - x_j, y_i - y_j); \tag{6c}$$
$$f(I(l_i))\mathbb{1}_a(z_i)], \quad \forall i,j,a,b \tag{6d}$$

The inference scheme in Sec 4 solves $L^* = \arg\max_L w^\top \Phi(I,L)$. Given a set of training images in the form of $\{I^n, L^n\}_{n=1}^N$, we learn the model parameters w using a form of structural SVM [27] as follows:

$$\min_{w,\xi} \ \frac{1}{2}||w||^2 + C\sum_n \xi^n, \quad \text{s.t. } \forall n, \ \forall L \tag{7a}$$
$$w^\top \Phi(I^n, L^n) - w^\top \Phi(I^n, L) \geq \Delta(L, L^n) - \xi^n \tag{7b}$$

Consider a training image I^n, the constraint in Eq. 7b enforces the score of the true label L^n to be larger than the score of any other hypothesis label L by some margin. The loss function $\Delta(L, L^n)$ measures how incorrect L is compared with L^n. Similar to regular SVMs, ξ_n are slack variables used to handle soft margins. This formulation is often called margin-rescaling in the SVM-struct literature [27].

We use a loss function that decomposes into a sum of local losses defined on each part $\Delta(L, L^n) = \sum_{i=1}^K \Delta_i(L_i, L_i^n)$. If the i-th part is a primitive part, we define the local loss $\Delta_i(L_i, L_i^n)$ as:

$$\Delta_i(L_i, L_i^n) = \lambda \cdot \mathbb{1}(z_i \neq z_i^n) + d((x_i, y_i), (x_i^n, y_i^n)) \tag{8}$$

where $\mathbb{1}(\cdot)$ is an indicator function that takes 1 if its argument is true, and 0 otherwise. The intuition of Eq. 8 is as follows. If the hypothesized poselet z_i is the same as the ground-truth poselet z_i^n for the i-th part, the first term of Eq. 8 will be zero. Otherwise it will incur a loss λ (we choose $\lambda = 10$ in our experiments). The second term in Eq. 8, $d((x_i, y_i), (x_i^n, y_i^n))$, measures the distance (we use l_1 distance) between two image locations (x_i, y_i) and (x_i^n, y_i^n). If the hypothesized image location (x_i, y_i) is the same as the ground-truth image location (x_i^n, y_i^n) for the i-th part, no loss is added. Otherwise a loss proportional to the l_1 distance of these two locations will be incurred.

If the i-th part is not a primitive part, we simply set $\Delta(L_i, L_i^n)$ to be zero. This choice is based on the following observation. In our framework, non-primitive parts only serve as some intermediate representations that help us to search for and disambiguate small primitive parts. The final human parsing results are still obtained from configurations l_i of primitive parts. Even if a particular hypothesized L gets one of its non-primitive part labeling wrong, it should not be penalized as long as the labellings of primitive parts are correct.

The optimization problem in Eq. 7 is convex and can be solved using the cutting plane method implemented in the SVM-struct package [10]. However we opt to use a simpler stochastic subgradient descent method to allow greater flexibility in terms of implementation.

First, it is easy to show that Eq. 7 can be equivalently written as:

$$\min_w \frac{1}{2}||w||^2 + C\sum_n \mathcal{R}^n(L), \text{where } \mathcal{R}^n(L) = \tag{9a}$$
$$\max_L \Big(\Delta(L, L^n) + w^\top \Phi(I^n, L) - w^\top \Phi(I^n, L^n)\Big) \tag{9b}$$

In order to do gradient descent, we need to calculate the subgradient $\partial_w \mathcal{R}^n(L)$ at a particular w. Let us define:

$$L^\star = \arg\max_L \Big(\Delta(L, L^n) + w^\top \Phi(I^n, L)\Big) \tag{10}$$

Eq. 10 is called loss-augmented inference [10]. It can be shown that the subgradient $\partial_w \mathcal{R}^n(L)$ can be computed as

$\partial_w \mathcal{R}(L) = \Phi(I^n, L^\star) - \Phi(I^n, L^n)$. Since the loss function $\Delta(L, L^n)$ can be decomposed into a sum over local losses on each individual part, the loss-augmented inference in Eq. 10 can be solved in a similar way to the inference problem in Sec. 4. The only difference is that the local appearance model $\phi(l_i; I)$ needs to be augmented with the local loss function $\Delta(L_i, L_i^n)$. Interested readers are referred to [10] for more details.

6. Experiments

There are several datasets popular in the human parsing community, e.g. Buffy dataset [8], PASCAL stickmen dataset [5]. But these datasets are not suitable for us. First of all, they only contain upper-bodies, but we are interested in full-body parsing. Second, as pointed out by Tran and Forsyth [26], there are very few pose variations in those datasets. In fact, previous work has exploited this property of these datasets by pruning search spaces using upper-body detection and segmentation [8], or by building appearance model using location priors [5]. Third, the contrast of image frames of the Buffy dataset is relatively low. This issue suggests that better performance can be achieved by engineering detectors to overcome the contrast difficulties. Please refer to the discussion in [26] for more details. In our work, we choose to use two datasets[1] containing very aggressive pose variations. The first one is the UIUC people dataset introduced in [26]. The second one is a new sport image dataset we have collected from the Internet.

6.1. UIUC people dataset

The UIUC people dataset [26] contains 593 images (346 for training, 247 for testing). Most of them are images of people playing badminton. Some are images of people playing Frisbee, walking, jogging, standing, etc. Sample images and their parsing results are shown in the first three rows of Fig. 5. We compare with two other state-of-the-art approaches that do full-body parsing (with published codes): the improved pictorial structure by Andriluka *et al*. [1], and the iterative parsing method by Ramanan [16]. The results are also shown in Fig. 5.

To quantitatively evaluate different methods, we measure the percentage of correctly localized body parts. Following the convention proposed in [8], a body part is considered correctly localized if the endpoints of its segment lies within 50% of the ground-truth segment length from their true locations. The comparative results are shown in Table 1(a). Our method outperforms other approaches in localizing most of body parts.

Detection and parsing: An interesting aspect of our approach is that it produces not only the configurations of primitive parts, but also the configurations of other larger body parts. These pieces of information can potentially be used for applications (e.g. gesture-based HCI) that do not require precise localizations of body segments. In Fig. 6, we visualize the configurations of four larger parts on some examples. Interestingly, the configuration of the whole body directly gives us a person detector. So our model can be seen as a principled way of unifying human pose estimation, person detection, and many other areas related to understanding humans. In the first row of Table 2, we show the results of person detection on the UIUC people dataset by running our human parsing model, then picking the bounding box corresponding to the part "whole body" as the detection. We compare with the state-of-the-art person detectors in [1, 2, 6]. Since most images contain one person, we only consider the detection with the highest score on an image for all the methods. We use the metric defined in the PASCAL VOC challenge to measure the performance. A detection is considered correct if the intersection over union with respect to the ground truth bounding box is at least 50%. It is interesting to see that our method outperforms other approaches, even though it is not designed for person detection.

[1] Available at `http://vision.cs.uiuc.edu/humanparse/`

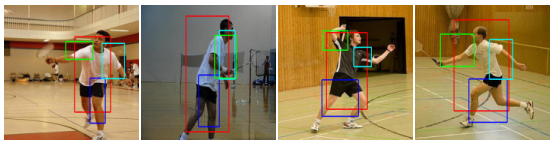

Figure 6: Examples of other information produced by our model. On each image, we show bounding boxes corresponding to the whole body, left arm, right arm and legs. The size of each bounding box is estimated from its corresponding poselet cluster.

	Our method	[1]	[6]	[2]
UIUC people	**66.8**	50.61	48.58	45.75
Sport image	**63.94**	59.94	45.61	39.75

Table 2: Comparison of accuracies of person detection on both datasets. In our method, the configuration of the poselet corresponding to the whole body can be directly used for person detection.

6.2. Sport image dataset

The UIUC people dataset is attractive because it has very aggressive pose and spatial variations. But one limitation of that dataset is that it mainly contains images of people playing badminton. One might ask what happens if the images are more diverse. To answer this question, we have collected a new sport image dataset from about 20 sport categories, including acrobatics, American football, croquet, cycling, hockey, figure skating, soccer, golf, horseback riding, etc. There are in total 1299 images. We randomly choose 649 of them for training and the rest for testing. The last three rows of Fig. 5 show examples of human parsing results, together with results of [1] and [16] on this dataset. The quantitative comparison is shown in Table 1(b). We can see that our approach outperforms the other two on the majority of body parts.

Similarly, we perform person detection using the poselet corresponding to the whole body. The results are shown in the second row of Table 2. Again, our method outperforms other approaches.

Figure 5: Examples of human body parsing on the UIUC people dataset (rows 1-3) and the sport image dataset (rows 4-6). We compare our method with the pictorial structure (PS) by Andriluka *et al*. [1] and the iterative image parsing (IIP) by Ramanan [16]. Notice the large pose variations, cluttered background, self-occlusions, and many other challenging aspects of these two datasets.

Method	Torso	Upper leg		Lower leg		Upper arm		Forearm		Head
Ramanan [16]	44.1	11.7	7.3	25.5	25.1	11.3	10.9	**25.9**	**25**	30.8
Andriluka *et al*. [1]	70.9	37.3	35.6	23.1	22.7	22.3	30.0	9.7	10.5	59.1
Our method	**86.6**	**58.3**	**54.3**	**53.8**	**46.6**	**28.3**	**33.2**	23.1	17.4	**68.8**

(a) UIUC people dataset

Method	Torso	Upper leg		Lower leg		Upper arm		Forearm		Head
Ramanan [16]	28.7	7.4	7.2	17.6	20.8	8.3	6.6	**20.2**	**21**	12.9
Andriluka *et al*. [1]	71.5	44.2	43.1	30.7	31	**28**	**29.6**	17.3	15.3	**63.3**
Our method	**75.3**	**50.1**	**48.2**	**42.5**	**36.5**	23.3	27.1	12.2	10.2	47.5

(b) Sport image dataset

Table 1: Human parsing results by our method and two comparison methods on two datasets. The percentage of correctly localized parts is shown for each primitive part. If two numbers are shown in one cell, they indicate the left/right body parts.

6.3. Kinematic tracking

To further illustrate our method, we apply the model learned from the UIUC people dataset for kinematic tracking by independently parsing the human figure in each frame. In Fig. 7, we show our results compared with applying the method in [16]. It is clear from the results that kinematic tracking is still a very challenging problem. Both methods make mistakes. Interestingly, when our method makes mistakes(e.g. figures with blue arrows), the output still looks like a valid body configuration. But when the method in [16] makes mistakes (e.g. figures with red arrows), the errors can be very wild. We believe this can be explained by the very different representations used in these two methods. In [16], a human body is represented by the set of primitive parts. Kinematic constraints are used

Figure 7: Examples of kinematic tracking on the baseball and figure skating datasets. The 1st and 3rd rows are our results. The 2rd and 4th rows are results of Ramanan [16]. Notice how mistakes of our method (blue arrows) still look like valid human poses, while those of [16] (red arrows) can be wild.

to enforce the connectivity of those parts. But these kinematic constraints have no idea what a person looks like as a whole. In the incorrect results of [16], all the primitive parts are perfectly connected. The problem is their connectivity does not form a reasonable human pose as a whole.

In contrast, our model uses representations that capture a spectrum of both large and small body parts. Even in situations where the small primitive parts are hard to detect, our method can still reason about the plausible pose configuration by pulling information from large pieces of the human bodies.

7. Conclusion

We have presented hierarchical poselets, a new representation for human parsing. Different poselets in our representation capture human poses at various levels of granularity. Some poselets correspond to the rigid parts typically used in previous work. Others can correspond to large pieces of the human bodies. Poselets corresponding to different parts are organized in a structured hierarchical model. The model parameters are learned in a max-margin framework. The advantage of this representation is that it infers the human pose by pulling information across various levels of details, ranging from the coarse shape of the whole body, to the fine-detailed information of small rigid parts. Our representation combines the benefits of both traditional part-based methods (e.g. pictorial structure) and exemplar-based methods for human parsing. In addition to localized rigid parts, our model also outputs other useful information about various portions of the human bodies. As future work, we would like to explore how to fully exploit this information for human parsing, person detection, action recognition, gesture recognition, human computer interaction, and other tasks related to understanding humans. We would also like to extend our representation to the temporal domain and apply it on videos.

Acknowledgement: We thank David Forsyth and anonymous reviewers for helpful comments. This work was supported in part by NSF under IIS-0803603 and IIS-1029035, and by ONR under N00014-01-1-0890 and N00014-10-1-0934 as part of the MURI program. YW is also supported in part by an NSERC postdoc fellowship. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of NSF, ONR, NSERC.

References

[1] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR*, 2009.
[2] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. In *ECCV*, 2010.
[3] L. Bourdev and J. Malik. Poselets: Body part detectors training using 3d human pose annotations. In *ICCV*, 2009.
[4] N. Dalal and B. Triggs. Histogram of oriented gradients for human detection. In *CVPR*, 2005.
[5] M. Eichner and V. Ferrari. Better appearance models for pictorial structures. In *BMVC*, 2009.
[6] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE PAMI*, 2009.
[7] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1):55–79, 2005.
[8] V. Ferrari, M. Marín-Jiménez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR*, 2008.
[9] H. Jiang and D. R. Martin. Globel pose estimation using non-tree models. In *CVPR*, 2008.
[10] T. Joachims, T. Finley, and C.-N. Yu. Cutting-plane training of structural SVMs. *Machine Learning*, 2008.
[11] S. X. Ju, M. J. Black, and Y. Yaccob. Cardboard people: A parameterized model of articulated image motion. In *FG*, 1996.
[12] X. Lan and D. P. Huttenlocher. Beyond trees: Common-factor models for 2d human pose recovery. In *ICCV*, 2005.
[13] G. Mori. Guiding model search using segmentation. In *ICCV*, 2005.
[14] G. Mori and J. Malik. Estimating human body configurations using shape context matching. In *ECCV*, 2002.
[15] G. Mori, X. Ren, A. Efros, and J. Malik. Recovering human body configuration: Combining segmentation and recognition. In *CVPR*, 2004.
[16] D. Ramanan. Learning to parse images of articulated bodies. In *NIPS*, 2007.
[17] D. Ramanan, D. A. Forsyth, and A. Zisserman. Strike a pose: Tracking people by finding stylized poses. In *CVPR*, 2005.
[18] X. Ren, A. Berg, and J. Malik. Recovering human body configurations using pairwise constraints between parts. In *ICCV*, 2005.
[19] B. Sapp, C. Jordan, and B. Taskar. Adaptive pose priors for pictorial structures. In *CVPR*, 2010.
[20] B. Sapp, A. Toshev, and B. Taskar. Cascaded models for articulated pose estimation. In *ECCV*, 2010.
[21] G. Shakhnarovich, P. Viola, and T. Darrell. Fast pose estimation with parameter sensitive hashing. In *ICCV*, 2003.
[22] L. Sigal and M. J. Black. Measure locally, reason globally: Occlusion-sensitive articulated pose estimation. In *CVPR*, 2006.
[23] P. Srinivasan and J. Shi. Bottom-up recognition and parsing of the human body. In *CVPR*, 2007.
[24] J. Sullivan and S. Carlsson. Recognizing and tracking human action. In *ECCV*, 2002.
[25] T.-P. Tian and S. Sclaroff. Fast globally optimal 2d human detection with loopy graph models. In *CVPR*, 2010.
[26] D. Tran and D. Forsyth. Improved human parsing with a full relational model. In *ECCV*, 2010.
[27] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *JMLR*, 6:1453–1484, 2005.
[28] Y. Wang and G. Mori. Multiple tree models for occlusion and spatial constraints in human pose estimation. In *ECCV*, 2008.
[29] W. Yang, Y. Wang, and G. Mori. Recognizing human actions from still images with latent poses. In *CVPR*, 2010.
[30] L. Zhu, Y. Chen, Y. Lu, C. Lin, and A. Yuille. Max margin AND/OR graph learning for parsing the human body. In *CVPR*, 2008.