Journal of Machine Learning Research 13 (2012) 3075-3102 Submitted 10/11; Revised 9/12; Published 10/12

Discriminative Hierarchical Part-based Models for Human Parsing and Action Recognition

Yang Wang YWANG@CS.UMANITOBA.CA
Department of Computer Science
University of Manitoba
Winnipeg, MB, R3T 2N2, Canada

Duan Tran DDTRAN2@UIUC.EDU
Zicheng Liao LIAO17@UIUC.EDU
David Forsyth DAF@UIUC.EDU
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA

Editor: Isabelle Guyon and Vassilis Athitsos

Abstract

We consider the problem of parsing human poses and recognizing their actions in static images 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. In this paper, we introduce hierarchical poselets—a new representation for modeling the pose configuration of human bodies. 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. The hierarchical poselets are organized in a hierarchical way via a structured model. Human parsing can be achieved by inferring the optimal labeling of this hierarchical model. The pose information captured by this hierarchical model can also be used as a intermediate representation for other high-level tasks. We demonstrate it in action recognition from static images.

Keywords: human parsing, action recognition, part-based models, hierarchical poselets, max-margin structured learning

1. Introduction

Modeling human bodies (or articulated objects in general) in images is a long-lasting problem in computer vision. Compared with rigid objects (e.g., faces and cars) which can be reasonably modeled using several prototypical templates, human bodies are much more difficult to model due to the wide variety of possible pose configurations.

A promising solution for dealing with the pose variations is to use part-based models. Part-based representations, such as cardboard people (Ju et al., 1996) or pictorial structure (Felzenszwalb and Huttenlocher, 2005), provide an elegant framework for modeling articulated objects, such as human bodies. 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, for example, 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. Part-based models have been used extensively in various computer vision applications involving humans, such as human parsing (Felzenszwalb and Huttenlocher, 2005; Ramanan, 2006), kinematic tracking (Ramanan et al., 2005), action recognition (Yang et al., 2010) and human-object interaction (Yao and Fei-Fei, 2010).

Considerable progress has been made to improve part-based models. For example, there has been a line of work on using better appearance models in part-based models. A representative example is the work by Ramanan (2006), who learns color histograms of parts from an initial edge-based model. Ferrari et al. (2008) and Eichner and Ferrari (2009) further improve the part appearance models by reducing the search space using various tricks, for example, the relative locations of part locations with respect to a person detection and the relationship between different part appearances (e.g., upper-arm and torso tend to have the same color), Andriluka et al. (2009) build better edge-based appearance models using the HOG descriptors (Dalal and Triggs, 2005). Sapp et al. (2010b) develop efficient inference algorithm to allow the use of more expensive features. There is also work (Johnson and Everingham, 2009; Mori et al., 2004; Mori, 2005; Srinivasan and Shi, 2007) on using segmentation as a pre-processing step to provide better spatial support for computing part appearances.

Another line of work is on improving configuration priors in part-based models. Most of them focus 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 (Lan and Huttenlocher, 2005), loopy graphs (Jiang and Martin, 2008; Ren et al., 2005; Tian and Sclaroff, 2010; Tran and Forsyth, 2010), mixtures of trees (Wang and Mori, 2008). There is also work on building spatial priors that adapt to testing examples (Sapp et al., 2010a).

Most of the previous work on part-based models use rigid parts that are anatomically meaningful, for example, torso, head, half limbs. Those rigid parts are usually represented as rectangles (e.g., Andriluka et al. 2009; Felzenszwalb and Huttenlocher 2005; Ramanan 2006; Ren et al. 2005; Sigal and Black 2006; Wang and Mori 2008) or parallel lines (e.g., Ren et al. 2005). However, as pointed out by some recent work (Bourdev and Malik, 2009; Bourdev et al., 2010), rigid parts are not necessarily the best representation since rectangles and parallel lines are inherently difficult to detect in natural images.

In this paper, we introduce a presentation of parts inspired by the early work of Marr (1982). The work in Marr (1982) recursively represents objects as generalized cylinders in a coarse-to-fine hierarchical fashion. In this paper, we extend Marr's idea for two problems in the general area of "looking at people". The first problem is human parsing, also known as human pose estimation. The goal is to find the location of each body part (torso, head, limbs) of a person in a static image. We use a part-based approach for human parsing. The novelty of our work is that 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 propose a new representation called "hierarchical poselets" to capture this hierarchy of parts. We infer the human pose using this hierarchical representation.

The hierarchical poselet also provides rich information about body poses that can be used in other applications. To demonstrate this, we apply it to recognize human action in static images. In this application, we use hierarchical poselets to capture various pose information of the human body, this information is further used as some intermediate representation to infer the action of the person.

A preliminary version of this work appeared in Wang et al. (2011). We organize the rest of the paper as follows. Section 2 reviews previous work in human parsing and action recognition. Section 3 introduces hierarchical poselet, a new representation for modeling human body configurations. Section 4 describes how to use hierarchical poselets for human parsing. Section 5 develops variants of hierarchical poselets for recognizing human action in static images. We present experimental results on human parsing and action recognition in Section 6 and conclude in Section 7.

2. Previous Work

Finding and understanding people from images is a very active area in computer vision. In this section, we briefly review previous work in human parsing and action recognition that is most related to our work.

Human parsing: Early work related to finding people from images is in the setting of detecting and tracking people with kinematic models in both 2D and 3D. Forsyth et al. (2006) provide an extensive survey of this line of work.

Recent work has examined the problem in static images. Some of these approaches are exemplar-based. For example, Toyama and Blake (2001) track people using 2D exemplars. Mori and Malik (2002) and Sullivan and Carlsson (2002) estimate human poses by matching pre-stored 2D templates with marked ground-truth 2D joint locations. Shakhnarovich et al. (2003) use local sensitive hashing to allow efficient matching when the number of exemplars is large.

Part-based models are becoming increasingly popular in human parsing. Early work includes the cardboard people (Ju et al., 1996) and the pictorial structure (Felzenszwalb and Huttenlocher, 2005). Tree-structured models are commonly used due to its efficiency. But there are also methods that try to alleviate the limitation of tree-structured models, include common-factor models (Lan and Huttenlocher, 2005), loopy graphs (Jiang and Martin, 2008; Ren et al., 2005; Tian and Sclaroff, 2010; Tran and Forsyth, 2010), mixtures of trees (Wang and Mori, 2008).

Many part-based models use discriminative learning to train the model parameters. Examples include the conditional random fields (Ramanan and Sminchisescu, 2006; Ramanan, 2006), max-margin learning (Kumar et al., 2009; Wang et al., 2011; Yang and Ramanan, 2011) and boosting (Andriluka et al., 2009; Sapp et al., 2010b; Singh et al., 2010). Previous approaches have also explored various features, including image segments (superpixels) (Johnson and Everingham, 2009; Mori et al., 2004; Mori, 2005; Sapp et al., 2010a,b; Srinivasan and Shi, 2007), color features (Ramanan, 2006; Ferrari et al., 2008), gradient features (Andriluka et al., 2009; Johnson and Everingham, 2010; Wang et al., 2011; Yang and Ramanan, 2011).

Human action recognition: Most of the previous work on human action recognition focuses on videos. Some work (Efros et al., 2003) uses global template for action recognition. A lot of recent work (Dollár et al., 2005; Laptev et al., 2008; Niebles et al., 2006) uses bag-of-words models. There is also work (Ke et al., 2007; Niebles and Fei-Fei, 2007) using part-based models.

Compared with videos, human action recognition from static images is a relatively less-studied area. Wang et al. (2006) provide one of the earliest examples of action recognition in static images. Recently, template models (Ikizler-Cinbis et al., 2009), bag-of-words models (Delaitre et al., 2010), part-based models (Delaitre et al., 2010; Yang et al., 2010) have all been proposed for static-image action recognition. There is also a line of work on using contexts for action recognition in static images, including human-object context (Desai et al., 2010; Gupta et al., 2009; Yao and Fei-Fei, 2010) and group context (Lan et al., 2010; Maji et al., 2011).

3. Hierarchical Poselets

Our pose representation is based on the concept of "poselet" introduced in Bourdev and Malik (2009). 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 (Bourdev and Malik, 2009; Bourdev et al., 2010).

In this paper, we propose a new representation called *hierarchical poselets*. 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 part-based human modeling 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 by Bourdev and Malik (2009), 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, for example, "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 (Ramanan et al., 2005).

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 Figure 1. To avoid terminological confusion, we will use "part" to denote one of the 20 parts in Figure 1 and use "primitive part" to denote rigid body parts (i.e., torso, head, half limbs) from now on.

In this paper, we choose the 20 parts and the hierarchical structure in Figure 1 manually. Of course, it is possible to define parts corresponding to other combinations of body segments, for example, left part of the whole body. It may also be possible to learn the connectivity of parts automatically from data, for example, using structure learning methods (Koller and Friedman, 2009). We would like to leave these issues as future work.

We use a procedure similar to Yang et al. (2010) 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

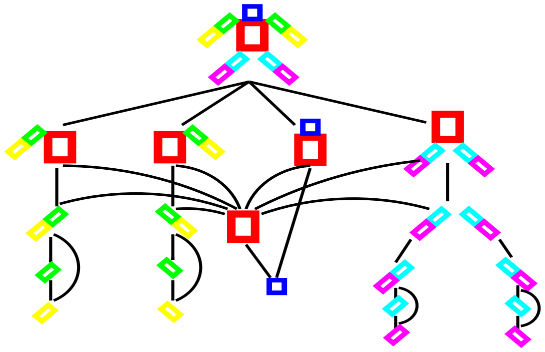


Figure 1: An illustration of the hierarchical pose representation. The black edges indicate the connectivity among different parts.

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 (Dalal and Triggs, 2005) 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 (Felzenszwalb et al., 2010), 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 Figure 2. Examples of poselets and their corresponding HOG templates for other body parts are shown in Figure 3.

A poselet of a primitive part contains two endpoints. For example, for a poselet of upper-left leg, one endpoint corresponds 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 in human parsing when we need to infer the endpoints of a primitive part for a test image.

4. Human Parsing

In this section, we describe how to use hierarchical poselets in human parsing. We first develop an undirected graphical model to represent the configuration of the human pose (Section 4.1). We then develop the inference algorithm for finding the best pose configuration in the model (Section 4.2) and the algorithm for learning model parameters (Section 4.3) from training data.

4.1 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 $l_i = (x_i, y_i, z_i)$.

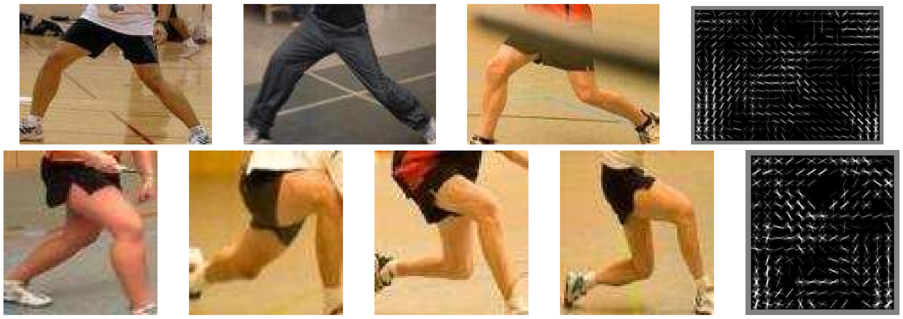


Figure 2: Examples of two poselets for the part "legs". Each row corresponds to a poselet. We show several patches from the poselet cluster. The last column shows the HOG template of the poselet.

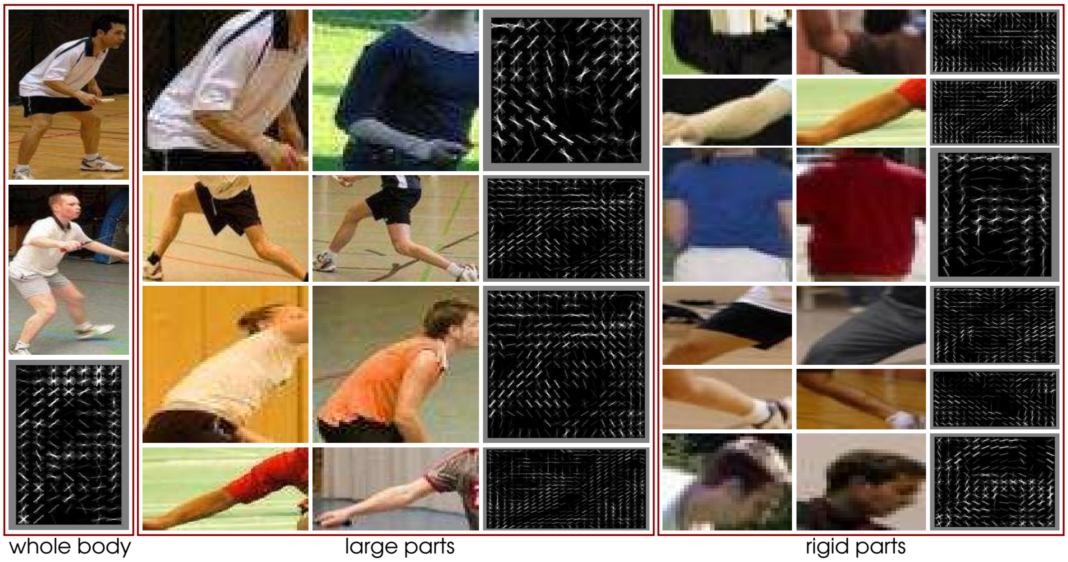


Figure 3: Visualization of some poselets learned from different body parts on the UIUC people data set, including whole body, large parts (top to bottom: torso+left arm, legs, torso+head, left arm), and rigid parts (top to bottom: upper/lower left arm, torso, upper/lower left leg, head). For each poselet,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

Here (x_i, y_i) defines the image location, and z_i is the index of the corresponding poselet for this part, that is, $z_i \in \{1, 2, \ldots, \mathcal{P}_i\}$, 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, for example, scale and foreshortening.

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 Figure 1. We define the score of labeling an image I with the pose L as:

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

The details of the potential functions in Equation 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:

$$\begin{aligned}\psi(l_i, l_j) &= \alpha_{i;j;z_i;z_j}^{\top} \text{bin}(x_i - x_j, y_i - y_j)\\ &= \sum_{a=1}^{\mathcal{P}_i} \sum_{b=1}^{\mathcal{P}_j} \mathbb{1}_a(z_i) \mathbb{1}_b(z_j) \alpha_{i;j;a;b}^{\top} \text{bin}(x_i - x_j, y_i - y_j).\end{aligned}$$

Similar to Ramanan (2006), 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:

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

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:

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

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 Yang et al. (2010), 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. For example, look at the examples in Figure 2 and Figure 3, the poselets capture various characteristic patterns for large parts, such as the “A”-shape for the legs in the first row of Figure 2.

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. (2010b). The method in Sapp et al. (2010b) 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 (Bourdev and Malik, 2009; Bourdev et al., 2010) uses a simple Hough voting scheme for person detection, that is, 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 develop 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 Zhu et al. (2008). But there are two important differences. First, the appearance models used in Zhu et al. (2008) 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 Zhu et al. (2008) are conceptually similar to poselets in our case. But the OR-nodes in Zhu et al. (2008) are defined manually, while our poselets are learned.

Our work on human parsing 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 (Mori and Malik, 2002; Shakhnarovich et al., 2003; Sullivan and Carlsson, 2002) 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.

4.2 Inference

Given an image I, the inference problem is to find the optimal pose labeling L^* that maximize the score $F(L,I)$, that is, $L^* = \arg\max_L F(L,I)$. 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:

$$\begin{aligned} m_i(l_j) &= \max_{l_i}(u(l_j) + \psi(l_i,l_j)), \\ u(l_j) &= \phi(l_j) + \sum_{k \in \text{kids}_j} m_k(l_j). \end{aligned} \tag{2}$$

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 Ramanan (2006). 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 Equation 2 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. 3).

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.

4.3 Learning

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

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

The inference scheme in Section 4.2 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 (Tsochantaridis et al., 2005) as follows:

$$\min_{w,\xi} \ \frac{1}{2}||w||^2 + C\sum_n \xi^n, \quad \text{s.t. } \forall n, \ \forall L: \tag{3}$$

$$w^\top \Phi(I^n, L^n) - w^\top \Phi(I^n, L) \geq \Delta(L, L^n) - \xi^n. \tag{4}$$

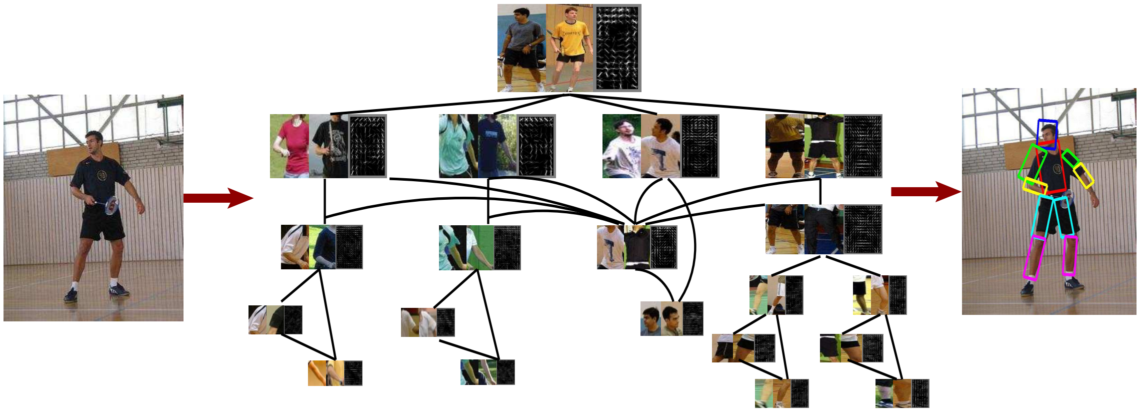


Figure 4: A graphical illustration of applying our model on a test image. For each part (please refer to Figure 1), we show the inferred poselet by visualizing two sample patches from the corresponding poselet cluster and the SVM HOG template.

Consider a training image I^n, the constraint in Equation 4 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 (Tsochantaridis et al., 2005).

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:

$$\Delta_i(L_i,L_i^n) = \lambda \cdot \mathbb{1}(z_i \neq z_i^n) + d((x_i,y_i),(x_i^n,y_i^n)), \tag{5}$$

where $\mathbb{1}(\cdot)$ is an indicator function that takes 1 if its argument is true, and 0 otherwise. The intuition of Equation 5 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 Equation 5 will be zero. Otherwise it will incur a loss λ (we choose $\lambda = 10$ in our experiments). The second term in Equation 5, $d((x_i,y_i),(x_i^n,y_i^n))$, 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 labelings of primitive parts are correct.

The optimization problem in Equations (3,4) is convex and can be solved using the cutting plane method implemented in the SVM-struct package (Joachims et al., 2008). However we opt to use a simpler stochastic subgradient descent method to allow greater flexibility in terms of implementation.

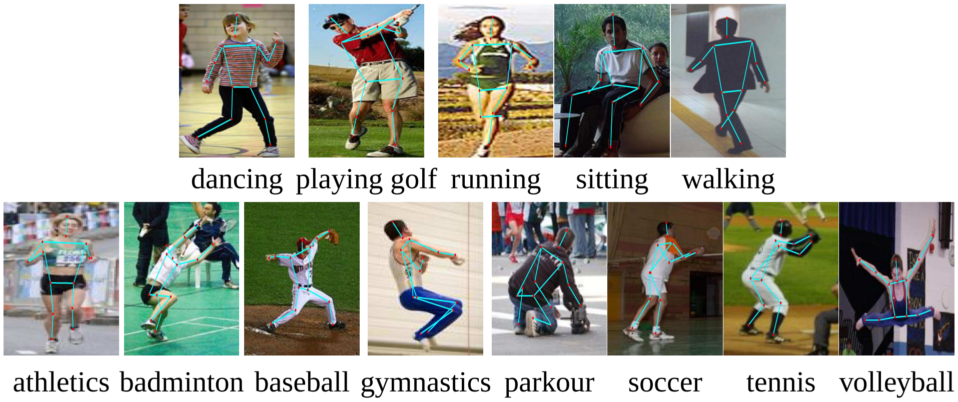


Figure 5: Human actions in static images. We show some sample images and their annotations on the two data sets used in our experiments (see Section 6). Each image is annotated with the action category and joints on the human body. It is clear from these examples that static images convey a lot of information about human actions.

First, it is easy to show that Equations (3,4) can be equivalently written as:

$$\begin{aligned}&\min_w \frac{1}{2}||w||^2 + C\sum_n \mathcal{R}^n(L),\\ &\text{where}\quad \mathcal{R}^n(L) = \max_L \Big(\Delta(L,L^n) + w^\top\Phi(I^n,L) - w^\top\Phi(I^n,L^n)\Big).\end{aligned}$$

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:

$$L^\star = \arg\max_L \Big(\Delta(L,L^n) + w^\top\Phi(I^n,L)\Big). \tag{6}$$

Equation 6 is called loss-augmented inference (Joachims et al., 2008). 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 Equation 6 can be solved in a similar way to the inference problem in Section 4.2. 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 Joachims et al. (2008) for more details.

5. Action Recognition

The hierarchical poselet is a representation general enough to be used in many applications. In this section, we demonstrate it in human action recognition from static images.

Look at the images depicted in Figure 5. We can easily perceive the actions of people in those images, even though only static images are given. So far most work in human action recognition has been focusing on recognition from videos. While videos certainly provide useful cues (e.g., motion) for action recognition, the examples in Figure 5 clearly show that the information conveyed by static images is also an important component of action recognition. In this paper, we consider the problem of inferring human actions from static images. In particular, we are interested in exploiting the human pose as a source of information for action recognition.

Several approaches have been proposed to address the problem of static image action recognition in the literature. The first is a standard pattern classification approach, that is, learning a classifier based on certain image feature representations. For example, Ikizler-Cinbis et al. (2009) learn SVM classifiers based on HOG descriptors. The limitation with this approach is that it completely ignores the pose of a person. Another limitation is that SVM classifiers implicitly assume that images from the same action category can be represented by a canonical prototype (which are captured by the weights of the SVM classifier). However, the examples in Figure 5 clearly show that humans can have very varied appearances when performing the same action, which are hard to characterize with a canonical prototype.

Another approach to static image action recognition is to explicitly recover the human pose, then use the pose as a feature representation for action recognition. For example, Ferrari et al. (2009) estimate the 2D human pose in TV shots. The estimated 2D poses can be used to extract features which in turn can be used to retrieve TV shots containing people with similar poses to a query. As point out in Yang et al. (2010), the problem with this approach is that 2D human pose estimation is still a very challenging problem. The output of the state-of-the-art pose estimation system is typically not reliable enough to be directly used for action recognition.

The work in Yang et al. (2010) is the closest to ours. It uses a representation based on human pose for action recognition. But instead of explicitly recovering the precise pose configuration, it represents the human pose as a set of latent variables in the model. Their method does not require the predicted human pose to be exactly correct. Instead, it learns which components of the pose are useful for differentiating various actions.

The pose representation in Yang et al. (2010) is limited to four parts: upper body, left/right arm, and legs. Learning and inference in their model amounts to infer the best configurations of these four parts for a particular action. A limitation of this representation is that it does not contain pose information about larger (e.g., whole body) or smaller (e.g., half-limbs) parts. We believe that pose information useful for discerning actions can vary depending on different action categories. Some actions (e.g., *running*) have distinctive pose characteristics in terms of both the upper and lower bodies, while other actions (e.g., *pointing*) are characterized by only one arm. The challenge is how to represent the pose information at various levels of details for action recognition.

In this section, we use hierarchical poselets to capture richer pose information for action recognition. While a richer pose representation may offer more pose information (less bias), it must also be harder to estimate accurately (more variance). In this paper, we demonstrate that our rich pose representation (even with higher variance) is useful for action recognition.

5.1 Action-Specific Hierarchical Poselets

Since our goal is action recognition, we choose to use an action-specific variant of the hierarchical poselets. This is similar to the action-specific poselets used in Yang et al. (2010). The difference is that the action-specific poselets in Yang et al. (2010) are only defined in terms of four parts—left/right arms, upper-body, and legs. These four parts are organized in a star-like graphical model. In contrast, our pose representation captures a much wider range of information across various pieces of the human body. So ours is a much richer representation than Yang et al. (2010).

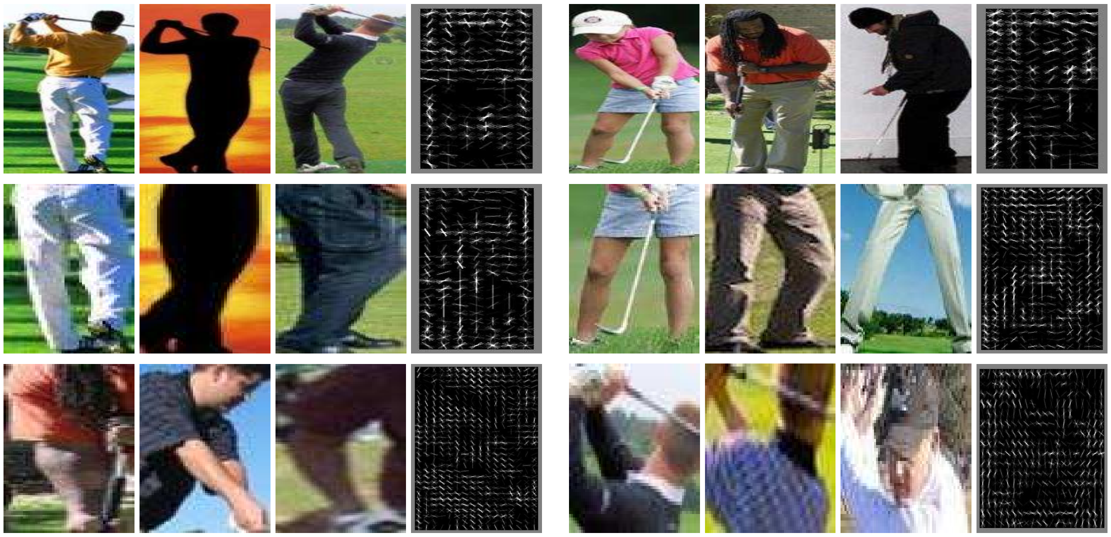


Figure 6: Examples of poselets for “playing golf”. For each poselet, we visualize several patches from the corresponding cluster and the SVM HOG template. Notice the multi-scale nature of the poselets. These poselets cover various portions of the human bodies, including the whole body (1st row), both legs (2nd row), one arm (3nd row), respectively.

The training images are labeled with ground-truth action categories and joints on the human body (Figure 5). We use the following procedure to select poselets for a specific part (e.g., *legs*) of a particular action category (e.g., *running*). We first collect training images of that action category (*running*). Then we cluster the joints on the part (*legs*) into several clusters based on their relative (x,y) coordinates with respect to some reference joint. Each cluster will correspond to a “running legs” poselet. We repeat this process for the part in other action categories. In the end, we obtain about 15 to 30 clusters for each part. Figures 6 and 7 show examples of poselets for “playing golf” and “running” actions, respectively.

Similarly, we train a classifier based on HOG features (Dalal and Triggs, 2005) to detect the presence of each poselet. Image patches in the corresponding poselet cluster are used as positive examples and random patches as negative examples for training the classifier. Similar to the model in Sec. 4, we use different cell sizes when constructing HOG features for different parts. Large cell sizes are used for poselets of large body parts (e.g., whole body and torso), while small cell sizes are used for small body parts (e.g., half limbs). Figure 6 and Figure 7 show some examples of the learned SVM weights for some poselets.

5.2 Our Model

Let I be an image containing a person, $Y \in \mathcal{Y}$ be its action label where $\mathcal{Y}$ is the action label alphabet, L be the pose configuration of the person. The complete pose configuration is denoted as $L = \{l_i\}_{i=1}^{K}$ ($K = 20$ in our case), where $l_i = (x_i, y_i, z_i)$ represents the 2D image location and the index of the corresponding poselet cluster for the i-th part. The complete pose L can be represented by a graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ shown in Figure 1. A vertex $i \in \mathcal{V}$ denotes the i-th part and an edge $(i, j) \in \mathcal{E}$ represents

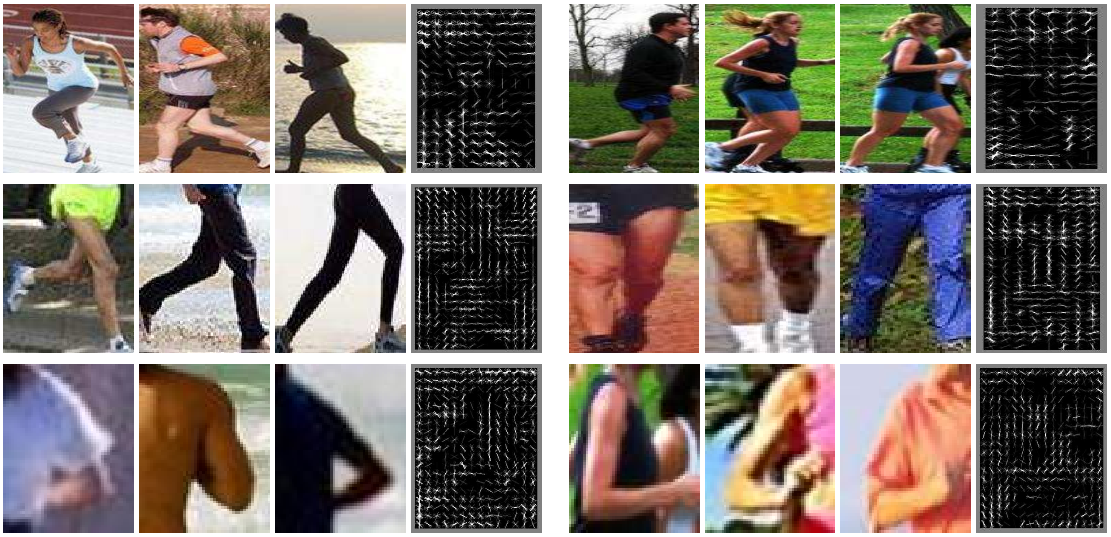


Figure 7: Examples of poselets for "running". For each poselet, we visualize several patches from the corresponding cluster and the SVM HOG template. Similar to Figure 6, these poselets cover various portions of the human bodies

the spatial constraint between the i-th and the j-th parts. We define the following scoring function to measure the compatibility of the triple (I,L,Y):

$$F(I,L,Y) = \omega_Y(I) + \sum_{i \in \mathcal{V}} \phi_Y(I,l_i) + \sum_{i,j \in \mathcal{E}} \psi_Y(l_i,l_j). \tag{7}$$

Here we use the subscript to explicitly emphasize that these functions are specific for a particular action label Y. The details of the potential functions in Equation 7 are as follows.

Root appearance $\omega_Y(I)$*:* This potential function models the compatibility of the action label Y and the global appearance of an image I. It is parametrized as:

$$\omega_Y(I) = \alpha_Y^\top \cdot f(I). \tag{8}$$

Here $f(I)$ is a feature vector extracted from the whole image I without considering the pose. In this paper, we use the HOG descriptor (Dalal and Triggs, 2005) of I as the feature vector $f(I)$. The parameters α_Y can be interpreted as a HOG template for the action category Y. Note that if we only consider this potential function, the parameters $\{\alpha_Y\}_{Y \in \mathcal{Y}}$ can be obtained from the weights of a multi-class linear SVM trained with HOG descriptors $f(I)$ alone without considering the pose information.

Part appearance $\phi_Y(I,l_i)$*:* This potential function models the compatibility of the configuration l_i of the i-th part and the local image patch defined by $l_i = (x_i, y_i, z_i)$, under the assumption that the action label is Y. Since our goal is action recognition, we also enforce that the poselet z_i should comes from the action Y. In other words, if we define $\mathcal{Z}_i^Y$ as the set of poselet indices for the i-th part corresponding to the action category Y, this potential function is parametrized as:

$$\phi_Y(I,l_i) = \begin{cases} \beta_{i,Y}^\top \cdot f(I,l_i) & \text{if } z_i \in \mathcal{Z}_i^Y; \\ -\infty & \text{otherwise.} \end{cases} \tag{9}$$

Here $f(I,l_i)$ is the score of placing the SVM HOG template z_i at location (x_i,y_i) in the image I.

Pairwise part constraint $\psi(l_i,l_j)$*:* This potential function models the compatibility of the configurations between the i-th and the j-th parts, under the assumption that the action label is Y. We parametrize this potential function using a vectorized counts of spatial histogram bins, similar to Ramanan (2006); Yang et al. (2010). Again, we enforce poselets z_i and z_j to come from action Y as follows:

$$\psi_Y(l_i,l_j) = \begin{cases} \gamma_{i,Y}^\top \cdot \text{bin}(l_i - l_j) & \text{if } z_i \in \mathcal{Z}_i^Y, z_j \in \mathcal{Z}_j^Y; \\ -\infty & \text{otherwise.} \end{cases} \tag{10}$$

Here $\text{bin}(\cdot)$ is a vector all zeros with a single one for the occupied bin.

Note that if the potential functions and model parameters in Equations(7,8,9,10) do not depend on the action label Y, the part appearance $\phi(\cdot)$ and pairwise part constraint $\psi(\cdot)$ exactly recover the human parsing model in Section 4.

5.3 Learning and Inference

We define the score of labeling an image I with the action label Y as follows:

$$H(I,Y) = \max_L F(I,L,Y). \tag{11}$$

Given the model parameters $\Theta = \{\alpha,\beta,\gamma\}$, Equation 11 is a standard MAP inference problem in undirected graphical models. We can approximately solve it using message passing scheme similar to that in Section 4.2. The predicted action label Y^* is chosen as $Y^* = \arg\max_Y H(I,Y)$.

We adopt the latent SVM (Felzenszwalb et al., 2010) framework for learning the model parameters. First, it is easy to see that Equation 7 can be written as a linear function of model parameters as $F(I,L,Y) = \Theta^\top \Phi(I,L,Y)$, where Θ is the concatenation of all the model parameters (i.e., α, β and γ) and $\Phi(I,L,Y)$ is the concatenation of the corresponding feature vectors. Given a set of training examples in the form of $\{I^n,L^n,Y^n\}_{n=1}^N$, the model parameters are learned by solving the following optimization problem:

$$\min_{\Theta,\xi} \; \frac{1}{2}||\Theta||^2 + C\sum_n \xi^n, \quad \text{s.t. } \forall n, \;\; \forall Y: \tag{12}$$

$$H(I^n,Y^n) - H(I^n,Y) \geq \Delta(Y,Y^n) - \xi^n. \tag{13}$$

It is easy to show that Equations (12,13) can be equivalently written as:

$$\min_\Theta \frac{1}{2}||\Theta||^2 + C\sum_n \mathcal{R}^n, \tag{14}$$

$$\text{where } \; \mathcal{R}^n = \max_{Y,L}\Big(\Delta(Y,Y^n) + \Theta^\top \cdot \Phi(I^n,Y)\Big) - \max_L \Theta^\top \cdot \Phi(I^n,L,Y^n).$$

The problem in Equation 14 is not convex, but we can use simple stochastic sub-gradient descent to find a local optimum. Let us define:

$$(Y^*,L^*) = \arg\max_{Y,L}(\Delta(Y,Y^n) + \Theta^\top \cdot \Phi(I^n,L,Y)),$$

$$L' = \arg\max_L(\Theta^\top \cdot \Phi(I^n,L,Y^n)).$$

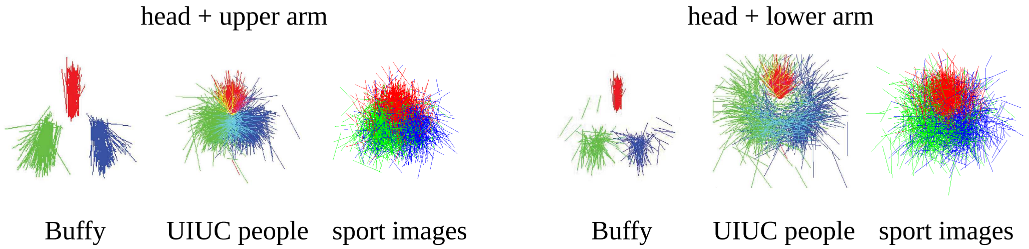


Figure 8: Scatter plots of heads (red) and upper/lower arms (blue and green) with respect to fixed upper body position on three data sets.

Then the gradient of Equation 14 can be computed as:

$$\Theta + C\sum_n \Big(\Phi(I^n, L^*, Y^*) - \Phi(I^n, L', Y^n)\Big).$$

To initialize the parameter learning, we first learn a pose estimation model using the labeled (I^n, L^n) collected from training examples with class label Y. The parameters of these pose estimation models are used to initialize β_Y and γ_Y. The parameters α_Y are initialized from a linear SVM model based on HOG descriptors without considering the poses.

6. Experiments

In this section, we present our experimental results on human parsing (Section 6.1) and action recognition (Section 6.2).

6.1 Experiments on Human Parsing

There are several data sets popular in the human parsing community, for example, Buffy data set (Ferrari et al., 2008), PASCAL stickmen data set (Eichner and Ferrari, 2009). But these data sets are not suitable for us for several reasons. First of all, they only contain upper-bodies, but we are interested in full-body parsing. Second, as pointed out in Tran and Forsyth (2010), there are very few pose variations in those data sets. In fact, previous work has exploited this property of these data sets by pruning search spaces using upper-body detection and segmentation (Ferrari et al., 2008), or by building appearance model using location priors (Eichner and Ferrari, 2009). Third, the contrast of image frames of the Buffy data set 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 Tran and Forsyth (2010) for more details. In our work, we choose to use two data sets[1] containing very aggressive pose variations. The first one is the UIUC people data set introduced in Tran and Forsyth (2010). The second one is a new sport image data set we have collected from the Internet which has been used in Wang et al. (2011). Figure 8 shows scatter plots of different body parts of our data sets compared with the Buffy data set (Ferrari et al., 2008) using a visualization style similar to Tran and Forsyth (2010) . It is clear that the two data sets used in this paper have much more variations.

[1] Both data sets can be downloaded from `http://vision.cs.uiuc.edu/humanparse`.

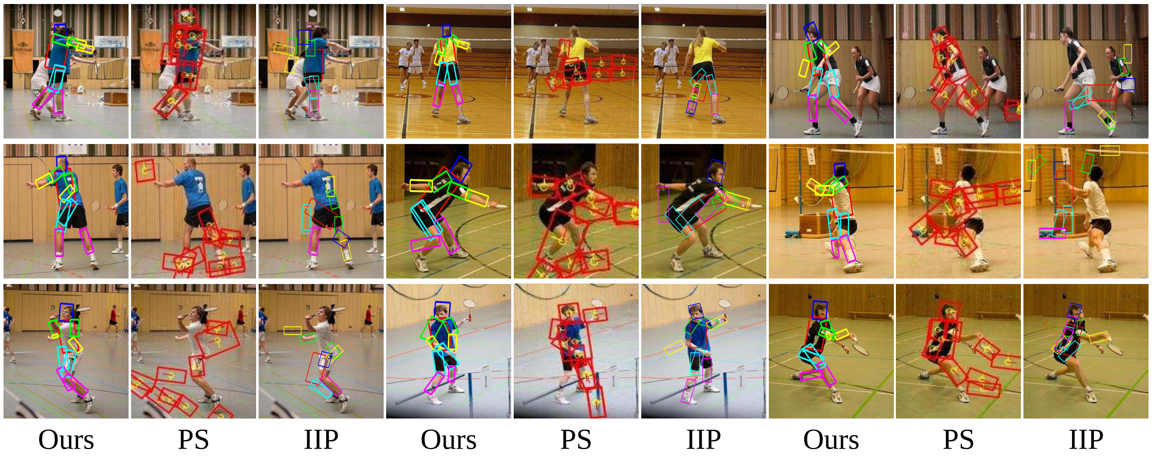


Figure 9: Examples of human body parsing on the UIUC people data set. We compare our method with the pictorial structure (PS) (Andriluka et al., 2009) and the iterative image parsing (IIP) (Ramanan, 2006). Notice the large pose variations, cluttered background, self-occlusions, and many other challenging aspects of the data set.

6.1.1 UIUC PEOPLE DATA SET

The UIUC people data set (Tran and Forsyth, 2010) 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 or standing. Sample images and their parsing results are shown in the first three rows of Figure 9. 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. (2009), and the iterative parsing method by Ramanan (2006). The results are also shown in Figure 9.

To quantitatively evaluate different methods, we measure the percentage of correctly localized body parts. Following the convention proposed in Ferrari et al. (2008), 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. We also show the result (3rd row, Table 1(a)) of using only the basic-level poselets corresponding to the rigid parts. It is clear that our full model using hierarchical poselets outperforms using rigid parts alone.

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 Figure 10, 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 data set 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 Felzenszwalb et al. (2010) and Andriluka et al.

Method	Torso	Upper leg		Lower leg		Upper arm		Forearm		Head
Ramanan (2006)	44.1	11.7	7.3	25.5	25.1	11.3	10.9	**25.9**	**25**	30.8
Andriluka et al. (2009)	70.9	37.3	35.6	23.1	22.7	22.3	30.0	9.7	10.5	59.1
Our method (basic-level)	79.4	53.8	53.4	47.8	39.7	17.8	21.1	11.7	16.6	65.2
Our method (full model)	**86.6**	**58.3**	**54.3**	**53.8**	**46.6**	**28.3**	**33.2**	23.1	17.4	**68.8**

(a) UIUC people data set

Method	Torso	Upper leg		Lower leg		Upper arm		Forearm		Head
Ramanan (2006)	28.7	7.4	7.2	17.6	20.8	8.3	6.6	**20.2**	**21**	12.9
Andriluka et al. (2009)	71.5	44.2	43.1	30.7	31	**28**	**29.6**	17.3	15.3	**63.3**
Our method (basic-level)	73.3	45.0	47.6	40.4	39.9	19.4	27.0	13.3	9.9	47.5
Our method (full model)	**75.3**	**50.1**	**48.2**	**42.5**	**36.5**	23.3	27.1	12.2	10.2	47.5

(b) Sport image data set

Table 1: Human parsing results by our method and two comparison methods (Ramanan, 2006; Andriluka et al., 2009) on two data sets. 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. As a comparison, we also show the results of using only rigid parts (basic-level).

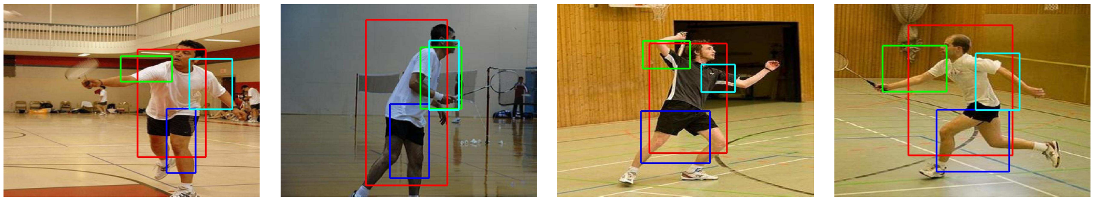


Figure 10: Examples of other information produced by our model. On each image, we show bounding boxes corresponding to the whole body, left arm, right arm and legs. The size of each bounding box is estimated from its corresponding poselet cluster.

(2009). Since most images contain one person, we only consider the detection with the best 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.

	Our method	Felzenszwalb et al. (2010)	Andriluka et al. (2009)
UIUC people	**66.8**	48.58	50.61
Sport image	**63.94**	45.61	59.94

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

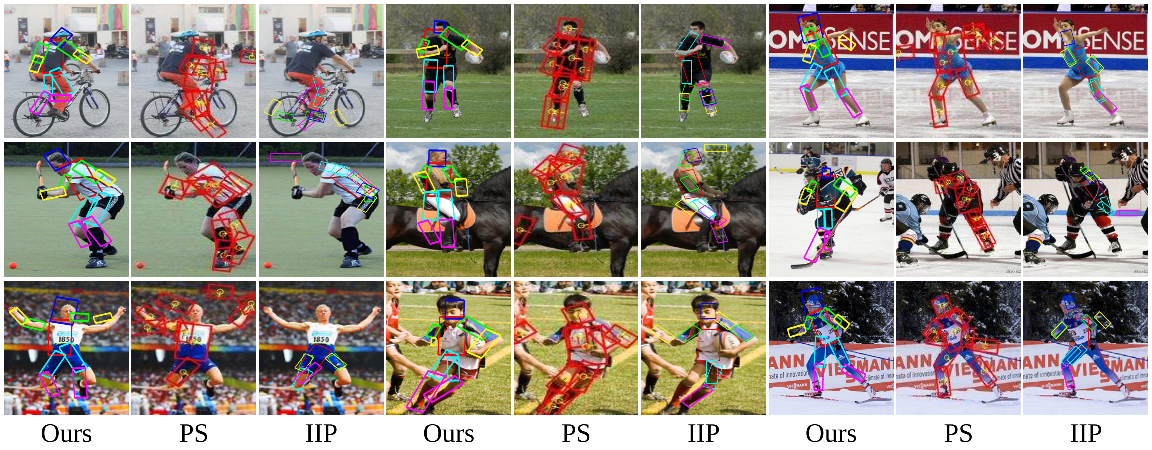


Figure 11: Examples of human body parsing on the sport image data set. We compare our method with the pictorial structure (PS) (Andriluka et al., 2009) and the iterative image parsing (IIP) (Ramanan, 2006).

6.1.2 SPORT IMAGE DATA SET

The UIUC people data set is attractive because it has very aggressive pose and spatial variations. But one limitation of that data set 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 data set from more than 20 sport categories, including acrobatics, American football, croquet, cycling, hockey, figure skating, soccer, golf and horseback riding. There are in total 1299 images. We randomly choose 649 of them for training and the rest for testing. The last three rows of Figure 9 show examples of human parsing results, together with results of Andriluka et al. (2009) and Ramanan (2006) on this data set. 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.

6.1.3 KINEMATIC TRACKING

To further illustrate our method, we apply the model learned from the UIUC people data set for kinematic tracking by independently parsing the human figure in each frame. In Figure 12, we show our results compared with applying the method in Ramanan (2006). 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 Ramanan (2006) 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 Ramanan (2006), a human body is represented by the set of primitive parts. Kinematic constraints are used 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

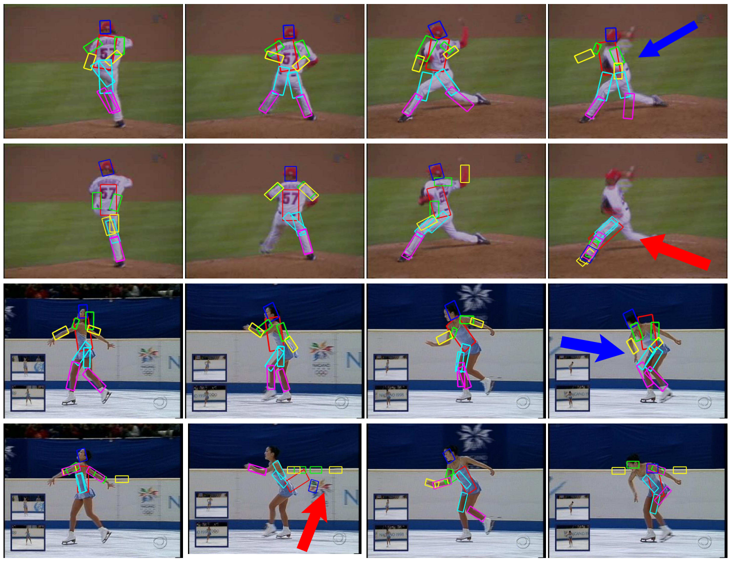


Figure 12: Examples of kinematic tracking on the baseball and figure skating data sets. The 1st and 3rd rows are our results. The 2rd and 4th rows are results of Ramanan (2006). Notice how mistakes of our method (blue arrows) still look like valid human poses, while those of Ramanan (2006) (red arrows) can be wild.

results of Ramanan (2006), 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.

6.2 Experiments on Action Recognition

We test our approach on two publicly available data sets: the still images data set (Ikizler et al., 2008) and the Leeds sport data set (Johnson and Everingham, 2010). Both data sets contain images of people with ground-truth pose annotations and action labels.

6.2.1 STILL IMAGE DATA SET

We first demonstrate our model on the still image data set collected in Ikizler et al. (2008). This data set contains more than 2000 static images from five action categories: dancing, playing golf, running, sitting, and walking. Sample images are shown in the first two rows of Figure 5. Yang et al. (2010) have annotated the pose with 14 joints on the human body on all the images in the data set. Following Yang et al. (2010), we choose 1/3 of the images from each category to form the training data, and the remaining ones as the test data.[2]

[2] A small number of images/annotations we obtained from the authors of Yang et al. (2010) are somehow corrupted due to some file-system failure. We have removed those images from the data set.

method	overall	avg per-class
Our approach	**65.15**	**70.77**
Yang et al. (2010)*	63.49	68.37
SVM mixtures	62.8	64.05
Linear SVM	60.32	61.5

Table 3: Performance on the still image data set. We report both overall and average per-class accuracies. *The results are based on our own implementation.

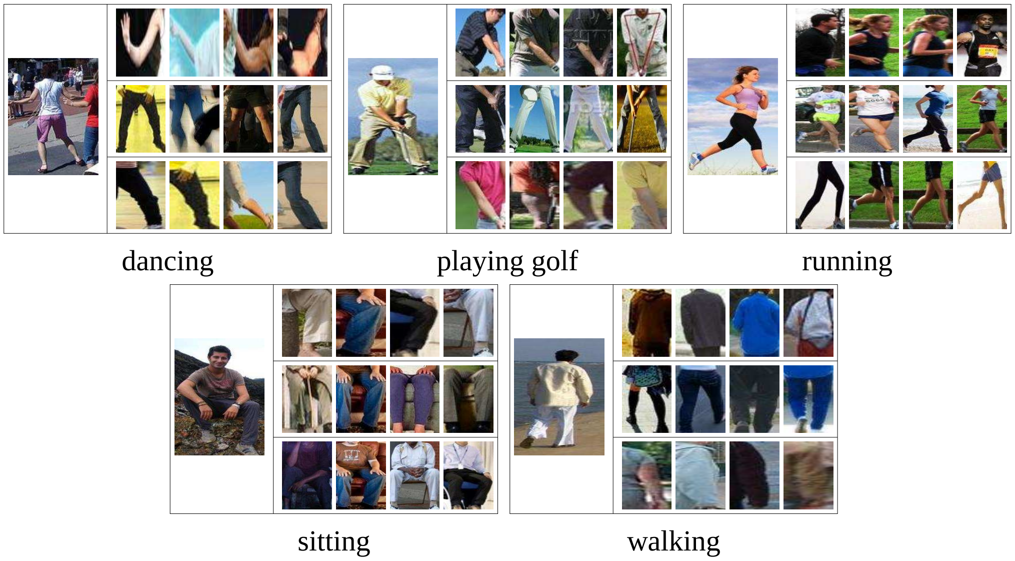
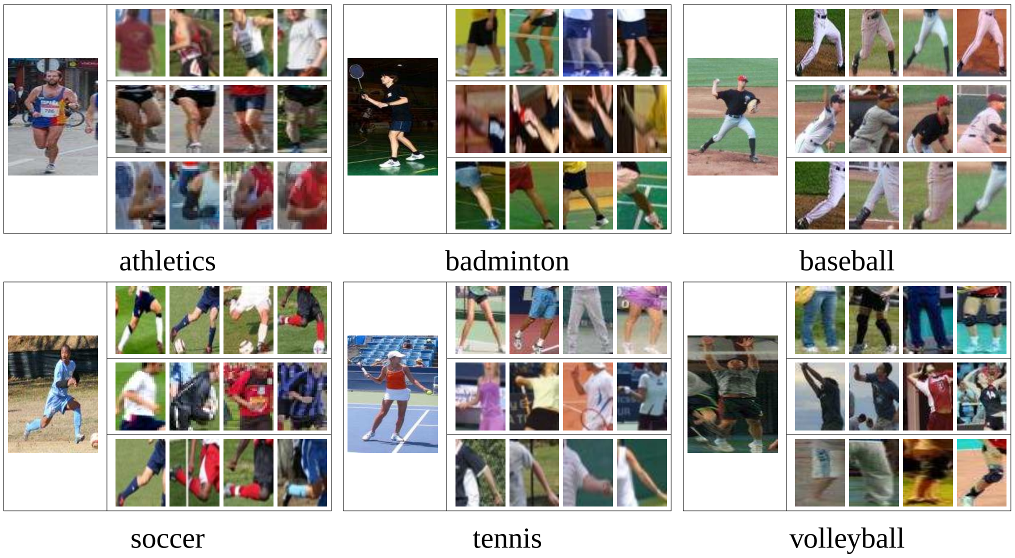


Figure 13: Visualization of some inferred poselets on the still image data set. These test images have been correctly recognized by our model. For a test image, we show three poselets that have high responses. Each poselet is visualized by showing several patches from its cluster.

We compare our approach with two baseline method. The first baseline is a multi-class SVM based on HOG features. For the second baseline, we use mixtures of SVM models similar to that in Felzenszwalb et al. (2010). We set the number of mixtures for each class to be the number of whole-body poselets. From Table 3, we can see that our approach outperforms the baseline by a large margin. Our performance is also better than the reported results in Yang et al. (2010). However, the accuracy numbers are not directly comparable since the training/testing data sets and features are not completely identical. In order to do a fair comparison, we re-implemented the method in Yang et al. (2010) by only keeping the parts used in Yang et al. (2010). Our full model performs better.

In Figure 13, we visualize several inferred poselets on some examples whose action categories are correctly classified. Each poselet is visualized by showing several patches from the corresponding poselet cluster.



athletics badminton baseball

soccer tennis volleyball

Figure 14: Visualization of some inferred poselets on the Leeds sport data set. These test images have been correctly recognized by our model. For a test image, we show three poselets that have high responses. Each poselet is visualized by showing several patches from its cluster.

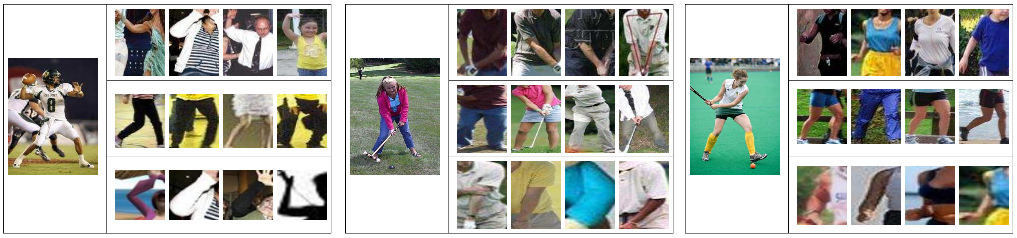
method	overall	avg per-class
Our approach	**54.6**	**54.6**
SVM mixtures	52.7	49.13
Linear SVM	52.7	52.93

Table 4: Performance on the Leeds sport data set. We report both overall and average per-class accuracies.

6.2.2 LEEDS SPORT DATA SET

The Leeds sport data set (Johnson and Everingham, 2010) contains 2000 images from eight different sports: athletics, badminton, baseball, gymnastics, parkour, soccer, tennis, volleyball. Each image in the data set is labeled with 14 joints on the human body. Sample images and the labeled joints are shown in the last four rows of Figure 5. This data set is very challenging due to very aggressive pose variations.

We choose half of the images for training, and the other half for testing. The performance is shown in Table 4. Again, we compare with the HOG-based SVM and SVM mixtures as the baselines. We can see that our method still outperforms the baseline. Similarly, we visualize the inferred poselets on some examples in Figure 14.



American football→dancing croquet→playing golf field hockey→running

Figure 15: Visualization of inferred poses on unseen actions. Here the actions of the test images (*American football*, *croquet* and *field hockey*) are not available during training. Our model recognizes these examples as *dancing*, *playing golf*, *running*, respectively. Some of the results (e.g., *croquet*→ *golfing*) make intuitive sense. Others (e.g., *football*→*dancing*) might not be intuitive at first. But if we examine the poselets carefully, we can see that various pieces of the football player are very similar to those found in the dancing action.

6.2.3 UNSEEN ACTIONS

An interesting aspect of our model is that it outputs not only the predicted action label, but also some rich intermediate representation (i.e., action-specific hierarchical poselets) about the human pose. This information can potentially be exploited in various contexts. As an example, we apply the model learned from the still image data set to *describe* images from sports categories not available during training. In Figure 15, we show examples of applying the model learned from the still image data set to images with unseen action categories. The action categories (*American football*, *croquet* and *field hockey*) for the examples in Figure 15 are disjoint from the action categories of the still image data set. In this situation, our model obviously cannot correctly predict the action labels (since they are not available during training). Instead, it classifies those images using the action labels it has learned. For example, it classifies “American football” as “dancing”, “croquet” as “playing golf”, “field hockey” as “running”. More importantly, our model outputs poselets for various parts which support its prediction. From these information, we can say a lot about “American football” even though the predicted action label is wrong. For example, we can say it is closer to “dancing” than “playing golf” because the pose of the football player in the image is similar to certain type of dancing legs, and certain type of dancing arms.

7. Conclusion and Future Work

We have presented hierarchical poselets, a new representation for modeling human poses. 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 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. We have demonstrate the applications of this representation in human parsing and human action recognition from static images. Recently, similar ideas (Sun and Savarese, 2011) have been applied in other applications, such as object detection.

As future work, we would like to explore how to automatically construct the parts and the hierarchy using data-driven methods. This will be important in order to extend hierarchical poselets to other objects (e.g., birds) that do not have obvious kinematic structures. We also like to apply the hierarchical poselet representation to other vision tasks, such as segmentation.

Acknowledgments

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. Yang Wang was also supported in part by an NSERC postdoc fellowship when the work was done. 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, or NSERC.

References

Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2009.

Lubomir Bourdev and Jitendra Malik. Poselets: Body part detectors training using 3d human pose annotations. In *IEEE International Conference on Computer Vision*, 2009.

Lubomir Bourdev, Subhransu Maji, Thomas Brox, and Jitendra Malik. Detecting people using mutually consistent poselet activations. In *European Conference on Computer Vision*, 2010.

Navneet Dalal and Bill Triggs. Histogram of oriented gradients for human detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005.

Vincent Delaitre, Ivan Laptev, and Josef Sivic. Recognizing human actions in still images: a study of bag-of-features and part-based representations. In *British Machine Vision Conference*, 2010.

Chaitanya Desai, Deva Ramanan, and Charless Fowlkes. Discriminative models for static human-object interactions. In *Workshop on Structured Models in Computer Vision*, 2010.

Piotr Dollár, Vincent Rabaud, Garrison Cottrell, and Serge Belongie. Behavior recognition via sparse spatio-temporal features. In *ICCV'05 Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 2005.

Alexei A. Efros, Alexander C. Berg, Greg Mori, and Jitendra Malik. Recognizing action at a distance. In *IEEE International Conference on Computer Vision*, pages 726–733, 2003.

Marcin Eichner and Vittorio Ferrari. Better appearance models for pictorial structures. In *British Machine Vision Conference*, 2009.

Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, January 2005.

Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1672–1645, 2010.

Vittorio Ferrari, Manuel Marín-Jiménez, and Andrew Zisserman. Progressive search space reduction for human pose estimation. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2008.

Vittorio Ferrari, Manuel Marín-Jiménez, and Andrew Zisserman. Pose search: retrieving people using their pose. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2009.

David A. Forsyth, Okan Arikan, Leslie Ikemoto, James O'Brien, and Deva Ramanan. Computational studies of human motion: Part 1, tracking and motion synthesis. *Foundations and Trends in Computer Graphics and Vision*, 1(2/3):77–254, July 2006.

Abhinav Gupta, Aniruddha Kembhavi, and Larry S. Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(10):1775–1789, 2009.

Nazli Ikizler, R. Gokberk Cinbis, Selen Pehlivan, and Pinar Duygulu. Recognizing actions from still images. In *International Conference on Pattern Recognition*, 2008.

Nazli Ikizler-Cinbis, R. Gokberk Cinbis, and Stan Sclaroff. Learning actions from the web. In *IEEE International Conference on Computer Vision*, 2009.

Hao Jiang and David R. Martin. Globel pose estimation using non-tree models. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2008.

Thorsten Joachims, Thomas Finley, and Chun-Nam Yu. Cutting-plane training of structural SVMs. *Machine Learning*, 2008.

Sam Johnson and Mark Everingham. Combining discriminative appearance and segmentation cues for articulated human pose estimation. In *International Workshop on Machine Learning for Vision-based Motion Analysis*, 2009.

Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *British Machine Vision Conference*, 2010.

Shanon X. Ju, Michael J. Black, and Yaser Yaccob. Cardboard people: A parameterized model of articulated image motion. In *International Conference on Automatic Face and Gesture Recognition*, pages 38–44, 1996.

Yan Ke, Rahul Sukthankar, and Martial Hebert. Event detection in crowded videos. In *IEEE International Conference on Computer Vision*, 2007.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

Pawan Kumar, Andrew Zisserman, and Philip H. S. Torr. Efficient discriminative learning of parts-based models. In *IEEE International Conference on Computer Vision*, 2009.

Tian Lan, Yang Wang, Weilong Yang, and Greg Mori. Beyond actions: Discriminative models for contextual group activities. In *Advances in Neural Information Processing Systems*. MIT Press, 2010.

Xiangyang Lan and Daniel P. Huttenlocher. Beyond trees: Common-factor models for 2d human pose recovery. In *IEEE International Conference on Computer Vision*, volume 1, pages 470–477, 2005.

Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2008.

Subhransu Maji, Lubomir Bourdev, and Jitendra Malik. Action recognition from a distributed representation of pose and appearance. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2011.

David Marr. *A Computational Investigation into the Human Representation and Processing of Visual Information*. W. H. Freeman, 1982.

Greg Mori. Guiding model search using segmentation. In *IEEE International Conference on Computer Vision*, volume 2, pages 1417–1423, 2005.

Greg Mori and Jitendra Malik. Estimating human body configurations using shape context matching. In *European Conference on Computer Vision*, volume 3, pages 666–680, 2002.

Greg Mori, Xiaofeng Ren, Alyosha Efros, and Jitendra Malik. Recovering human body configuration: Combining segmentation and recognition. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 326–333, 2004.

Juan Carlos Niebles and Li Fei-Fei. A hierarchical model of shape and appearance for human action classification. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2007.

Juan Carlos Niebles, Hongcheng Wang, and Li Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. In *British Machine Vision Conference*, volume 3, pages 1249–1258, 2006.

Deva Ramanan. Learning to parse images of articulated bodies. In *Advances in Neural Information Processing Systems*, volume 19, pages 1129–1136, 2006.

Deva Ramanan and Cristian Sminchisescu. Training deformable models for localization. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 206–213, 2006.

Deva Ramanan, David A. Forsyth, and Andrew Zisserman. Strike a pose: Tracking people by finding stylized poses. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 271–278, 2005.

Xiaofeng Ren, Alexander Berg, and Jitendra Malik. Recovering human body configurations using pairwise constraints between parts. In *IEEE International Conference on Computer Vision*, volume 1, pages 824–831, 2005.

Benjamin Sapp, Chris Jordan, and Ben Taskar. Adaptive pose priors for pictorial structures. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010a.

Benjamin Sapp, Alexander Toshev, and Ben Taskar. Cascaded models for articulated pose estimation. In *European Conference on Computer Vision*, 2010b.

Greg Shakhnarovich, Paul Viola, and Trevor Darrell. Fast pose estimation with parameter sensitive hashing. In *IEEE International Conference on Computer Vision*, volume 2, pages 750–757, 2003.

Leonid Sigal and Michael J. Black. Measure locally, reason globally: Occlusion-sensitive articulated pose estimation. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 2041–2048, 2006.

Vivek Kumar Singh, Ram Nevatia, and Chang Huang. Efficient inference with multiple heterogenous part detectors for human pose estimation. In *European Conference on Computer Vision*, 2010.

Praveen Srinivasan and Jianbo Shi. Bottom-up recognition and parsing of the human body. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2007.

Josephine Sullivan and Stefan Carlsson. Recognizing and tracking human action. In *European Conference on Computer Vision LNCS 2352*, volume 1, pages 629–644, 2002.

Min Sun and Silvio Savarese. Articulated part-base model for joint object detection and pose estimation. In *IEEE International Conference on Computer Vision*, 2011.

Tai-Peng Tian and Stan Sclaroff. Fast globally optimal 2d human detection with loopy graph models. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010.

Kentaro Toyama and Andrew Blake. Probabilistic exemplar-based tracking in a metric space. In *IEEE International Conference on Computer Vision*, volume 2, pages 50–57, 2001.

Duan Tran and David Forsyth. Improved human parsing with a full relational model. In *European Conference on Computer Vision*, 2010.

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

Yang Wang and Greg Mori. Multiple tree models for occlusion and spatial constraints in human pose estimation. In *European Conference on Computer Vision*, 2008.

Yang Wang, Hao Jiang, Mark S. Drew, Ze-Nian Li, and Greg Mori. Unsupervised discovery of action classes. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006.

Yang Wang, Duan Tran, and Zicheng Liao. Learning hierarchical poselets for human parsing. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2011.

Weilong Yang, Yang Wang, and Greg Mori. Recognizing human actions from still images with latent poses. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010.

Yi Yang and Deva Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2011.

Bangpeng Yao and Li Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010.

Long Zhu, Yuanhao Chen, Yifei Lu, Chenxi Lin, and Alan Yuille. Max margin AND/OR graph learning for parsing the human body. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2008.