Bottom-up Recognition and Parsing of the Human Body

Praveen Srinivasan (psrin@seas.upenn.edu) and Jianbo Shi (jshi@cis.upenn.edu)

GRASP Lab, University of Pennsylvania, 3330 Walnut Street Philadelphia, PA 19104; Ph: 650-906-7334; Fax: 215-573-2048

Abstract. Recognizing humans, estimating their pose and segmenting their body parts are key to high-level image understanding. Because humans are highly articulated, the range of deformations they undergo makes this task extremely challenging. Previous methods have focused largely on heuristics or pairwise part models in approaching this problem. We propose a bottom-up *growing*, similar to parsing, of increasingly more complete partial body masks guided by a composition tree. At each level of the growing process, we evaluate the partial body masks directly via shape matching with exemplars (and also image features), without regard to how the hypotheses are formed. The body is evaluated as a whole, not the sum of its parts, unlike previous approaches. Multiple image segmentations are included at each of the levels of the growing/parsing, to augment existing hypotheses or to introduce ones. Our method yields both a pose estimate as well as a segmentation of the human. We demonstrate competitive results on this challenging task with relatively few training examples on a dataset of baseball players with wide pose variation. Our method is comparatively simple and could be easily extended to other objects. We also give a learning framework for parse ranking that allows us to keep fewer parses for similar performance.

1 Introduction

Recognition, pose estimation and segmentation of humans and their body parts remain important unsolved problems in high-level vision. Action understanding and image search and retrieval are just a few of the areas that would benefit enormously from this task. There has been good previous work on this topic, but significant challenges remain ahead. We divide the previous literature on this topic into three main areas:

Top-down approaches: [4] developed the well-known pictorial structures (PS) method and applied it to human pose estimation. In the original formulation, PS does probablistic inference in a tree-structured graphical model usually with the torso as the root. PS recovers locations, scales and orientations of rigid rectangular part templates that represent an object. Pairwise potentials were limited to simple geometric relations (relative position and angle), while unary potentials were based on image gradients or edge detection. The tree structure is a limitation since many cues (e.g., symmetry of appearance of right and left legs) cannot be encoded. [13] extended the original model to encode the fact that symmetric limb pairs have similar color, and that parts have consistent color or colors in general, but how to incorporate more general cues seems unclear. [14] track people by repeatedly detecting them with a top-down PS method. [17] introduced a non-parametric belief propagation (NBP) method with occlusion reasoning to determine the pose. All these approaches estimate pose, and do not provide an

underlying segmentation of the image. Their ability to utilize more sophisticated cues beyond pixel-level cues and geometric constraints between parts is limited.

Search approaches: [12] utilized heuristic-guided search, starting from limbs detected as segments from Normalized Cuts (NCut) ([3]), and extending the limbs into a full-body pose and segmentation estimate. A follow up to this, [11], introduced an Markov-Chain Monte Carlo (MCMC) method for recovering pose and segmentation. [7] developed an MCMC technique for inferring 3-D body pose from 2-D images, but used skin and face detection as extra cues. [18] utilized a combination of top-down, MCMC and local search to infer 2-D pose.

Bottom-up/Top-down approaches [15] used bottom-up detection of parallel lines in the image as part hypotheses, and then combined these hypotheses into a full-body configuration via an integer quadratic program. [18] also fit into this category, as they use bottom-up cues such as skin pixel detection. Similarly, [6] integrated bottom-up skin color cues with a top-down, NBP process. [11] use superpixels to guide their search. While [2] estimate only segmentation and not pose for horses and humans in upright, running poses, they best utilize shape and segmentation information in their framework. [16] use bottom-up part detectors to detect part hypotheses, and then piece these hypotheses together using a simple dynamic programming (DP) procedure in much the same way as [4]. Lastly, our approach is similar to that described in [19], where bottom-up proposals are grouped hierarchically to produce a desired shape; unlike, [19] we use more sophisticated shape modelling than small groups of edges.

2 Overview of Our Method

Our method shares the same goal with traditional image parsing: explain an image in terms of a set of parse rules. It produces body part hypotheses in a bottom-up fashion and propogates them up a parsing tree for simultaneous detection and parsing of an entire body (excluding the arms). Our formulation is also similar to an AND-OR graph[5]. However, in our framework scoring depends only on the parse at the root node, while in the AND-OR graph it is a function of the entire set of rules for the parse. Further, we allow image-level features to participate at all levels of the parsing hierarchy, while AND-OR graph parsing typically restricts these features to leaf nodes.

Our approach has two important differences from traditional parsing (not including AND-OR graph parsing) approaches. First, we do not assume the whole is just the sum of its parts. Traditional parsing relies on a subtree independence property between the different levels of the parse tree to ensure efficient (polynomial time) computation of the best parse. Given the state of a node in the parse tree, the scores of subtrees rooted at the node's children are independent. While we also have tree structured parse rules, the goal is to streamline the parsing process, not to impose subtree independence. We discriminate our parse hypotheses based solely on their geometrical shapes. Traditional parsing assumes an additive decomposition of parse score across the parse rule applied and the scores of the parses used in the rule. In our framework, the score given to hypotheses at each node is a global shape matching score against a set of exemplar shapes. This shape score does not factorize over the parses that were grouped together and the parse rule applied.

Second, we allow the image features, in this case multiple image segmentations, to participate in all levels of the parse tree, not just at the leaves. Most parsing algorithms consider local hypotheses generated from the image only at the leaves of the parse tree. Such a requirement is too restrictive when we have a good (but not perfect) segmentation. Often, segments represent a large piece of the body containing several different parts (according to the parse tree). We introduce segmentation into the parse tree at all levels, either by merging with an existing hypothesis, or as a new hypothesis.

Define S, the set of initial shapes, which were detected in the image, and F a shape scoring function (larger is better). Then we wish to solve:

$$\max_{T \in 2^S} F(\bigcup_{i=1}^{|T|} T_i) \tag{1}$$

$\bigcup$ represents shape composition. In our case, we represent shape with binary masks and it is simply the pixel-wise OR operator. We make the assumption that the shape we are looking for is distinctive enough that is unlikely to appear by as a collection of shapes by random chance. For human bodies, this seems to be a reasonable assumption.

Note that F depends solely on the composition of multiple shapes, not on the individual masks themselves, scores associated with the shapes, or how the shapes were chosen. In traditional parsing, the score of a parse is usually defined in terms of the applied parse rule, and the scores of the children, allowing for efficient DP methods for computing an optimal parse.

However, if the function F does not contain special structure, solving (1) may be intractable; $|2^S|$ is large even when $|S|$ is small. We restrict the set of shapes to be considered by introducing a *shape composition tree*; we a give an example of such a tree in Figure 1. The root node of the tree represents the desired shape we wish to construct (a full body, with the exception of the arms), and we associate with it shape scoring function F. The other nodes represent shapes that are likely to be useful in constructing a mask m such that $F(m)$ is large, such as legs and the lower body. Finding these smaller parts can obviously help with finding the entire body.

For the root node, we would like to determine a set of good hypotheses that are likely to have high F score. Intuitively, a good set of hypotheses for the children (or child) of the root, can be used to find good hypotheses for the root. For example, a good lower body can help us locate the torso, giving us the lower body + torso, and in turn we can extend this to find the head, giving us the lower body + torso + head as shown in Figure 1.

Nodes always draw hypotheses directly from S in addition to their children. In addition, the parent-child relationships encoded in the tree restrict the sources of shapes that a node can draw from in generating shape hypotheses for itself (aside from S). This serves to control the computational complexity; if a node has at most 2 children with each at most n hypotheses, and we restrict composition to one shape from each child, we generate at most n^2 hypotheses from the children. However, we must prune these hypotheses along with those from S in order to prevent exponential growth in the the number of hypotheses at the root. This necessitates the association of a scoring function F_i (based on either shape, or shape with image features; see sections 2.2 and 3.1) with each node i, in addition to the root, for pruning. Shapes with low F_i value can be pruned

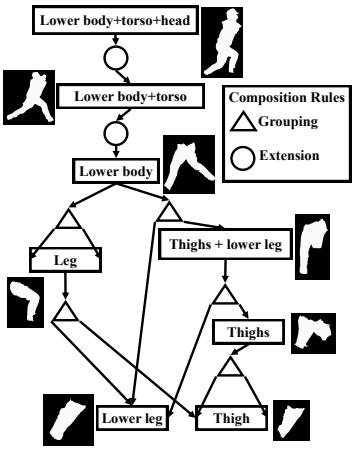

Fig. 1. Our shape composition tree, shown with an exemplar shape from our training set for each node; the exemplars are used for shape scoring. Shape composition begins at the leaf nodes of thigh and lower leg and proceeds upwards. Note that in addition to composing hypotheses from children nodes, hypotheses can always come from the initial shapes S.

away, as well as shapes that are redundant (highly similar to each other), resulting in a set of hypotheses of size n that are propogated up the tree.

We show a set of composition rules for our tree in Figure 2. The triangle and diamond symbols indicate two different types of hypothesis composition methods from children, to be discussed later. Implicitly, all nodes can draw on S for hypotheses. The leaf nodes are Thigh and Lower leg; shape composition starts at the leaves, and continues upwards to the root.

- {Lower leg, Thigh} → Leg
- {Thigh, Thigh} → Thighs
- {Thighs, Lower leg} → Thighs+Lower leg
- {Thighs+Lower leg, Lower leg} → Lower body
- {Leg, Leg} → Lower body
- {Lower body} → Lower body+torso
- {Lower body+torso} → Lower body+torso+head

Fig. 2. Our composition rules. We write them in reverse format to emphasize the bottom-up nature of the shape composition.

2.1 Multiple Segmentations

To initialize our bottom-up composition process, we need a set of intial shapes S. [12] noted that human limbs tend to be salient regions that NCut segmentation often isolate as a single segment. To make this initial shape hypothesis generation method more robust, we consider not one segmentation as in [12], but 12 different segmentations provided by NCut. We vary the number of segments from 5 to 60 in steps of 5, giving a total of 390 initial shapes per image. This allows us to segment out large parts of the body that are themselves salient, e.g. the lower body may appear as a single segment, as well as smaller parts like individual limbs or the head. Figure 3 shows for an image 2 of

the 12 segmenations with overlaid boundaries. Segments from different segmentations can overlap, or be contained within another. In our system, these segments are all treated equally. These initial shapes could be generated by other methods besides segmentation, but we found segmentation to be very effective.

Fig. 3. Two segmentations of an image, 10 and 40 segments. Red lines indicate segment boundaries for 10 segments, green lines indicate boundaries for 40 segments, and yellow indicates boundaries common to both segmentations (best viewed in color).

2.2 Shape Comparison

For each node i, we have an associated shape comparison function F_i. For the root node, this ranks the final hypotheses for us. For all other nodes, F_i ranks hypotheses so that they can be pruned. All the shape comparison functions operate the same way: we match the boundary contour of a mask against boundary contours of a set of exemplar shapes using the inner-distance shape context (IDSC) of [8].

The IDSC is an extension of the original shape context proposed in [1]. In the original shape context formulation, given a contour of n points $x_1, ..., x_n$, a shape context was computed for point x_i by the histogram

$$\#(x_j, j \neq i : x_j - x_i \in bin(k)) \tag{2}$$

Ordinarily, the inclusion function $x_j - x_i \in bin(k)$ is based on the Euclidean distance $d = \| x_j - x_i \|_2$ and the angle $acos((x_j - x_i)/d)$. However, these measures are very sensitive to articulation. The IDSC replaces these with an *inner-distance* and an *inner-angle*.

The inner-distance between x_i and x_j is the shortest path between the two points traveling through the interior of the mask. This distance is less sensitive to articulation. The inner-angle between x_i and x_j is the angle between the contour tangent at the point x_i and tangent at x_i of the shortest path leading from x_i to x_j. Figure 4 shows the interior shortest path and contour tangent.

The inner-distances are normalized by the mean inner-distance between all pairs $\{(x_i, x_j)\}$, $i \neq j$ of points. This makes the IDSC scale invariant, since angles are also scale-invariant. The inner-angles and normalized log inner-distances are binned to form a histogram, the IDSC descriptor. For two shapes with points $x_1, ..., x_n$ and $y_1, .., y_n$, IDCSs are computed at all points on both contours. For every pair of points x_i, y_j, a matching score between the two associated IDCSs is found using the Chi-Square

score ([1]). This forms an n-by-n cost matrix, which is used as input to a standard DP algorithm for string matching, allowing us to establish correspondence between the points on the two contours. The algorithm also permits occlusion of matches with a user-specified penalty. We try the alignment at several different, equally spaced starting points on the exemplar mask to handle the cyclic nature of the closed contours, and keep the best scoring alignment (and the score). Because the DP algorithm minimizes a cost (smaller is better), we negate the score since our desire is to maximize F and all F_i. The complexity of the IDSC computation and matching is dominated by the matching; with n contour points and s different starting points, the complexity is $O(sn^2)$.

If s is chosen as n, then the complexity is $O(n^3)$. However, we instead use a coarse-to-fine strategy for aligning two shapes; based on the alignment of a subsampling of the points, we can narrow the range of possible alignments for successively larger number of points, thereby greatly reducing the complexity. We use a series of 25, 50 and 100 points to do the alignment.

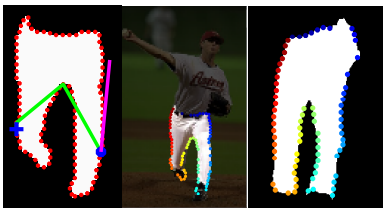

Fig. 4. IDSC Computation. **Left:** We show: shortest interior path (green) from start (blue dot) to end (blue cross); boundary contour points (red); contour tangent at start (magenta). The length of interior path is the inner-distance; the angle between contour tangent and the start of the interior path is the inner-angle. **Center:** Lower body mask hypothesis; colored points indicate correspondence established by IDSC matching with exemplar on **right**.

2.3 Composition Rule Application Procedure

Our composition process consists of five basic steps that can be used to generate the hypotheses for each node. For a particular node A, given all the hypotheses for all children nodes, we perform the following steps:

Segment inclusion: applies to all nodes We include by default all the masks in S as hypotheses for A. This allows us to cope with an input image that is itself a silhouette, which would not necessarily be broken into different limbs, for example. A leg will often appear as a single segment, not as separate segments for the thigh and lower leg; it is easier to detect this as a single segment, rather than trying to split segments into two or more pieces, and then recognize them separately. This is the only source of masks for leaf nodes in the composition tree.

Grouping: $\{B, C\} \rightarrow A$ For binary rules, we can compose hypotheses from two children such as grouping two legs into a lower body, e.g. $\{$Leg, Leg$\} \rightarrow$ Lower body. For each child, based on the alignment of the best matching exemplar to the child, we can predict which part of the segment boundary is likely to be adjacent to another part.

A pair of masks, b from B and c from C, are taken if the two masks are within 30 pixels of each other (approximately 1/10th of the image size in our images), and combined with the pixel-wise OR operator. Because we need a single connected shape for shape comparison, if the two masks are not directly adjacent we search for a mask from the segmentations that is adjacent to both, and choose the smallest such mask m. m is then combined with b and c into a single mask. If no such mask m exists, we keep the larger of a and b. Figure 5 provides an example of the composition rule, {Leg,Leg} $\rightarrow$ LowerBody.

Extension: $\{B\} \rightarrow A$ For unary rules we generate hypotheses by projecting an expected location for an additional part based on correspondence with exemplars. This is useful when bottom-up detection of a part by shape, such as the torso or head, is difficult due to wide variation of shape, or lack of distinctive shape. Once we have a large piece of the body (at least the lower body), it is more reliable to directly project a position for hypotheses. Given a hypothesis of the lower body and its correspondence to a lower body exemplar shape, we can project the exemplar's quadrilateral (quad) representing the torso on to the hypothesis (we estimate a transform with translation, rotation and scale based on the correspondence of two contour points closest to the two bottom vertices of the torso quad).

Similarly, given a mask for the lower body and torso, and its correspondence to exemplars, we can project quads for the head. With these quads, we look for all masks in S which have at least half their area contained within the quad, and combine them with the existing mask to give a new hypothesis. For each hypothesis/exemplar pair, we compose a new hypothesis. We sumamrize the general parsing process with the recursive algorithm presented in Algorithm 1.

Algorithm 1: P_A = Parse(A, S): for a particular image, given initial segments S and part name A, produce ranked and pruned parses for A.

Input: Part name A and initial shapes S
Output: P_A: set of ranked and pruned parses for A
$P_A = S$; `// Initialize parses for A to intial segments S`
foreach *rule* $\{B_i, C_i\} \rightarrow A$ *(or* $B_i \rightarrow A$*)* **do**
    P_{B_i} = Parse(B_i, S); `// Recurse`
    P_{C_i} = Parse(C_i, S); `// If binary rule, recurse`
    P_A = $P_A \cup$ Group(P_{B_i}, P_{C_i}) (or Extend(P_{B_i})); `// Add to parses of A`
end
P_A = RankByShape(P_A) or RankByImageFeatures(GetFeatures(F_A), w_A);
`// GetFeature gets parse features,` w_A `is classifier`
P_A = Prune(P_A); `// Prune redundant/low scoring parses`
return P_A; `// Return parses`

Scoring Once hypotheses have been composed, they are scored by matching to the nearest exemplar with IDSCs and DP. Correspondence is also established with the exemplar, providing an estimate of pose.

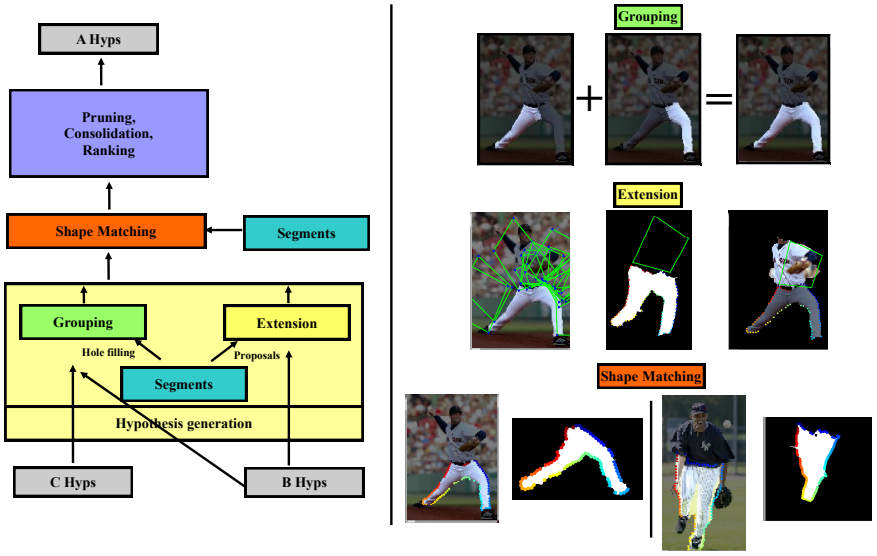

Fig. 5. Left: composition rule application. For binary rules, all pairs of child hypotheses within 30 pixels are grouped, with hole filling provided by segments if needed. For unary rules, hypotheses undergo extension using projected quads and segment proposals. Shape matching is performed on both the original segments as well as the composed/extended hypotheses. For leaf nodes, shape matching is performed only on the segments. After shape matching, the hypotheses are consolidated, pruned and ranked. **Right:** Grouping: two legs, on the left, are grouped into a lower body parse, on the right. *While recognizing legs alone from shape is not that robust, when they are combined into one shape, the distinctive shape of the lower body becomes apparent: the whole is not the sum of the parts.* Extension: the leftmost image shows a lower body parse with multiple torso quads projected from exemplars on to the image using the correspondence between the lower body hypothesis and the lower body exemplars; the center image shows the exemplar with its torso quad that yielded the best torso hypothesis, seen in the right image. Shape matching: two examples of shape matching. The lower body on the right was detected directly from the segments S, underscoring the importance of injecting the shapes from S into all levels of the composition tree.

Algorithm 2: Train(A, S): trains part classifiers and returns parses for part A and all descendants; uses input of initial segments S.

Input: Part name A and initial segments $S = \{S_1, ..., S_n\}$ for all images
Output: $P_A = \{P_A^1, ..., P_A^n\}$: set of parses for A across all images; w_A: WLR for A
$P_A = S$; `// Initialize parses for A to intial segments` S
foreach *rule* $\{B_i, C_i\} \rightarrow A$ *(or* $B_i \rightarrow A$*)* **do**
  $[P_{B_i}, w_{B_i}]$ = Train(B_i, S); `// Recurse to train child`
  $[P_{C_i}, w_{C_i}]$ = Train(C_i, S); `// If binary rule`
  foreach *image* j **do**
    $P_A^j = P_A^j \cup$ Group($P_{B_i}^j, P_{C_i}^j$) (or Extend($P_{B_i}^j$));
  end
end
F_A = GetFeatures(P_A); `// Get features for each parse`
G_A = GetGroundTruthOverlapScores(P_A); `// Get ground truth scores`
w_A = WLR(F_A, G_A); `// Do WLR on parses; get classifier` w_A
foreach *image* j **do**
  P_A^j = Prune(RankByImageFeatures(F_A^j, w_A)); `// Rank and prune`
end
return [P_A, w_A]; `// Return parses and classifer`

Pruning Many hypotheses are either low-scoring, redundant or both. We prune away these hypotheses with a simple greedy technique: we order the hypotheses by their shape score, from highest to lowest (best to worst). We add the best hypothesis to a representative set, and eliminate all other hypotheses which are similar to the just added hypothesis. We then recurse on the remaining hypotheses until the representative set reaches a fixed size. For mask similarity we use a simple mask overlap score O between masks a and b: $O(a, b) = \frac{area(a \bigcap b)}{area(a \bigcup b)}$, where $\bigcap$ performs pixel-wise AND, and the area is the number of "1" pixels. If $O(a, b)$ is greater than a threshold, a and b are considered to be similar. After this step, we have a set of hypotheses that can be passed higher in the tree, or to evaluate in the end if the node A is the root. Figure 5 illustrates the stages of the parsing process for generating the hypotheses for a single node. Also included are examples of grouping, extension, and shape matching/scoring.

3 Learning With Other Features

Besides shape, we also investigated using SIFT [9] and Probability of Boundary (PB) [10] features for ranking of parses. Traditional maximum-likelihood learning (similar to that for AND-OR graphs) would require an explicit specification of the best parse for each rule. Unfortunately, this is difficult to specify a-priori due to the exponential number of different possible parses (as combinations of the initial input shapes). Instead, for each part we train a weighted logistic regression (WLR) classifier to rerank the parses according to a variety of features. These parses are then pruned and passed up to other rules. We summarize our training procedure in Algorithm 2. It is essentially the same as

the original parsing procedure in Algorithm 1, but after generating the set of parses for a part, a WLR is trained to re-rank the parses. During testing, instead of ranking parses by shape score, they are ranked by the learnt WLR.

3.1 Features

For PB, we simply computed the average PB value along the boundary of the mask. For SIFT features, we created a SIFT codebook and computed codebook histograms for each mask as additional features. We extracted SIFT features from a dataset of 450 additional baseball images, using the code from http://vision.ucla.edu/ vedaldi/code/sift/sift.html with the default settings. These features were clustered via k-means into a 200 center codebook. For a test image and a given mask, SIFT features are extracted, associated with the nearest center in the codebook, and then a histogram over the frequency of each center in the mask is computed. This gives a 200 dimension vector of codebook center frequencies.

3.2 Learning the classifiers

The result is a 202 dimensional feature vector (texture: 200, shape: 1, PB: 1). Given these features, for each part we can learn a scoring function that best ranks the masks for that part. We also need ground truth scores for each part; from the ground truth segmentation of the parts in the image, we compute the overlap score between the mask and the ground truth labeling, giving a score in $[0, 1]$.

We optimize a WLR energy function that more heavily weights examples that have a high ground-truth score; a separate classifier is learnt for each part. f_i^j is the feature vector associated with the ith mask of the jth image, s_i^j is its ground truth score, and w parametrizes the energy function we wish to learn. For m images with n_j parses each, we have

$$E(w) = (\prod_{j=1}^{m}\prod_{i=1}^{n_j}(\frac{\exp(w^\mathsf{T} f_i^j)}{\sum_{i=1}^{n_j}\exp(w^\mathsf{T} f_i^j)})^{s_i^j})(\exp(\frac{-w^\mathsf{T} w}{\sigma^2})) \tag{3}$$

$$\log E(w) = \sum_{j=1}^{m}(\sum_{i=1}^{n_j} s_i^j (w^\mathsf{T} f_i^j)) - (\sum_{i=1}^{n_j} s_i^j)(\log(\sum_{i=1}^{n_j}\exp(w^\mathsf{T} f_i^j))) - \frac{w^\mathsf{T} w}{\sigma^2} \tag{4}$$

where we have added a regularization term $\exp(\frac{-w^\mathsf{T} w}{\sigma^2})$ that is a zero-mean isotropic Gaussian with variance σ^2. If, for each image j, exactly one of the ground truth scores s_i^j were 1 and the rest 0, this would be exactly logistic regression (LR). But since there may be several good shapes, we modify the LR. We can maximize this convex function via a BFGS quasi-newtonian solver using the function value above and the gradient below:

$$\partial_w \log E(w) = (\sum_{j=1}^{m}((\sum_{i=1}^{n_j} s_i^j f_i^j) - (\sum_{i=1}^{n_j} s_i^j)\frac{\sum_{i=1}^{n_j} f_i^j \exp(w^\mathsf{T} f_i^j)}{\sum_{i=1}^{n_j}\exp(w^\mathsf{T} f_i^j)})) - 2\frac{w}{\sigma^2} \tag{5}$$

Given a classifier w_p for each part p, we can rank a set of part hypotheses by simply evaluating ${w_p}^{\mathsf{T}} f_i^j$ and ranking the scores in descending order. We use the same parsing process, but with a different ranking function (a learned one), as opposed to using just shape. Algorithm 2 provides a pseudocode summary of this training procedure.

3.3 Results using shape only

We present results on the baseball dataset used in [12] and [11]. This dataset contains challenging variations in pose and appearance. We used 15 images to construct shape exemplars, and tested on $|I| = 39$ images. To generate the IDSC descriptors, we used the code provided by the authors of [8]. Boundary contours of masks were computed and resampled to have 100 evenly-spaced points. The IDSC histograms had 5 distance and 12 angle bins (in $[0, 2\pi]$). The occlusion penalty for DP matching of contours was 0.6 * (average match score). For pruning, we used a threshold of 0.95 for the overlap score to decide if two masks were similar (a, b are similar $\iff O(a, b) \geq 0.95$) for the lower body+torso and lower body + torso + head, and 0.75 for all other pruning. In all cases, we pruned to 50 hypotheses.

Because we limit ourselves to shape cues, the best mask (in terms of segmentation and pose estimate) found by the parsing process is not always ranked first; although shape is a very strong cue, it alone is not quite enough to always yield a good parse. Our main purpose was to investigate the use of global shape features over large portions of the body via shape composition. We evaluate our results in two different ways: segmentation score and projected joint position error. To the best of our knowledge, we are the first to present both segmentation and pose estimation results on this task.

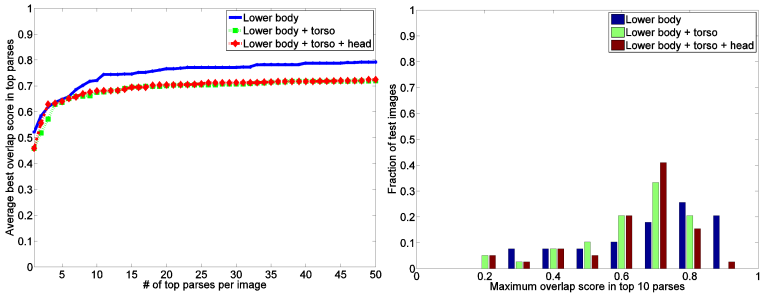

Fig. 6. Left: We plot the average (across all images) of the maximum overlap score as a function of the top k parses. **Right:** We focus on the top 10 parses, and histogram the best overlap score out of the top 10 for each image and region.

3.4 Segmentation Scoring

We present our results in terms of an overlap score for a mask with a ground truth labeling. Our composition procedure results in 50 final masks per image, ranked by

their shape score. We compute the overlap score $O(m, g)$ between each mask m and ground truth mask g. We then compute the cumulative maximum overlap score through the 50 masks. For an image i with ranked parses $p_1^i, ...p_n^i$, we compute overlap scores $o_1^i, ..., o_n^i$. From these scores, we compute the cumulative maximum $C^i(k) = \max(o_1^i, ..., o_k^i)$. The cumulative maximum gives us the best mask score we can hope to get by taking the top k parses.

To understand the behavior of the cumulative maximum over the entire dataset, we compute $M(k) = \frac{1}{|I|}\sum_{i=1}^{|I|} C^i(k)$, or the average of the cumulative maximum over all the test images for each $k = 1, ..., n$ ($n = 50$ in our case). This is the average of the best overlap score we could expect out of the top k parses for each image. We consider this a measure of both precision and recall; if our parsing procedure is good, it will have high scoring masks (recall) when k is small (precision). On left in Figure 6, we plot $M(k)$ against k for three different types of masks composed during our parsing process: lower body, lower body+torso, and lower body + head + torso. We can see that in the top 10 masks, we can expect to find a mask that is similar to the ground truth mask desired, with similarity 0.7 on average. This indicates that our parsing process does a good job of both generating hypotheses as well as ranking them.

While the above plot is informative, we can obtain greater insight into the overlap scores by examining all $C^i(k)$, $i = 1, ..., |I|$ for a fixed $k = 10$. We histogram the values of $C^i(10)$ on the right in Figure 6. We can see that most of the values are in fact well over 0.5, clustered mostly around 0.7. This confirms our belief that the composition process is effective in both recalling and ranking hypotheses, and that shape is a useful cue for segmenting human shape.

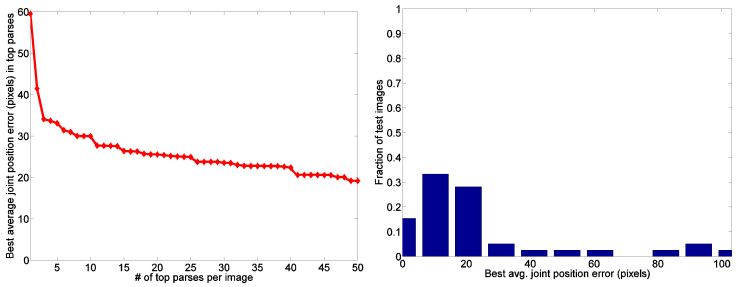

Fig. 7. Left: We plot the average (across all images) of the minimum average joint error in the top k parses as a function of k. **Right:** Taking the top 10 parses per image, we histogram the minimum average joint errors across all the images. We can see that the vast majority of average errors are roughly 20 pixels or less.

3.5 Joint Position Scoring

We also examinine the error in joint positions predicted by the correspondence of a hypothesis to the nearest exemplar. We take 5 joints: head-torso, torso-left thigh, torso-

right thigh, left thigh-left lower leg, right thigh-right lower leg. The positions of these joints are marked in the exemplars, and are mapped to a body hypothesis based on the correspondence between the two shapes. For a joint with position j in the exemplar, we locate the two closest boundary contour points p, q in the exemplar that have corresponding points p', q' in the shape mask. We compute a rotation, scaling and translation that transforms p, q to p', q', and apply these to j to obtain a joint estimate j' for the hypothesis mask. We compare j' with the ground truth joint position via Euclidean distance. For each mask, we compute the average error over the 5 joints. Given these scores, we can compute statistics in the same way as the overlap score for segmentation. On the left in Figure 7 we plot the average cumulative *minimum* $M(k)$, which gives the average best-case average joint error achieveable by keeping the top k masks. We see again that in the top 10 masks, there is a good chance of finding a mask with relatively low average joint error. On the right in Figure 7, we again histogram the data when $k = 10$.

Lastly, we show several example segmentations/registrations of images in Figure 8. Note that with the exception of the arms, our results are comparable to those of [11] (some of the images are the same), and in some cases our segmentation is better. As noted in [11], although quantitative measures may seem poor (e.g., average joint position error), qualitatively the results seem good.

4 Results with learning

Figure 9 shows plots of comparisons of shape only and shape, SIFT, and PB. For training, we used an additional 16 baseball images to train the WLR classifiers since training directly on the same images from which the shape exemplars were taken would likely have emphasized shape too much in the learning. 10 fold cross validation was performed with a range of different regularization values σ^2 to avoid overfitting. We tested on 26 baseball images.

We use the same type of plot as in the segmentation scoring previously described, and plot the results over all parts used in the parsing process. We can see that the use of additional features, particularly for the smaller parts, results in substantially better ranking of hypotheses. Even for larger regions, such as the lower body, the additional features have impact; the average cumulative maximum overlap score while keeping the top 10 parses with learning is approximately equal to the score when using only shape and keeping the top 20 parses, implying that we could keep half as many parses to obtain the same quality of result. For the largest regions, shape is clearly the most important cue, since performance is very similar. This also validates our choice of shape as a primary cue.

5 Conclusion

In summary, we present a shape composition method that constructs and verifies shapes in a bottom-up fashion. In contrast to traditional bottom-up parsing, our scoring functions at each node do not exhibit a subtree independence property; instead, we score shapes against a set of exemplars using IDSCs, which convey global shape information

Fig. 8. Body detection results. Ssegmentation has been highlighted and correspondence to the best matching exemplar indicated by colored dots. All parses were the top scoring parses for that image (images are ordered row-major), with the exception of images 4 (2nd best), 8 (3rd best), 6 (3rd best). Some images were cropped and scaled for display purposes only. Full body overlap scores for each image (images are ordered row-major): 0.83, 0.66, 0.72, 0.74, 0.76, 0.70, 0.44, 0.57 and 0.84. Average joint position errors for each image: 12.28, 28, 27.76, 10.20, 18.87, 17.59, 37.96, 18.15, and 27.79.

over both small and large regions of the body. We also infuse the process with multiple image segmentations as a pool of shape candidates at all levels, in contrast to typical parsing which only utilizes local image features at the leaf level.

We demonstrated competitive results on the challenging task of human pose estimation, on a dataset of baseball players with substantial pose variation. To the best of our knowledge, we are the first to present both quantitative segmentation and pose estimation results on this task. Note that in general, we need not start composition with the legs only; it would be entirely feasible to add other nodes (e.g. arms) as leaves.

Further, we use larger shapes (composed of multiple body limbs) than typical pose estimation methods. The notion of layers may also be useful in handling occlusion, as well as describing the shape relation of arms to the torso, since the arms often overlap the torso. Better grouping techniques (ones that introduce fewer hypotheses) are a good idea, since this would save substantial computation.

We also studied the introduction of more traditional features such as PB and SIFT codebook histograms and demonstrated that these features can make an important contribution. Our learning framework is well-suited to the parsing problem since unlike traditional maximum likelihood estimation, we do not explicitly require the best parse for each image. Instead, we simply require a function to provide a ground truth score for each parse hypothesis.

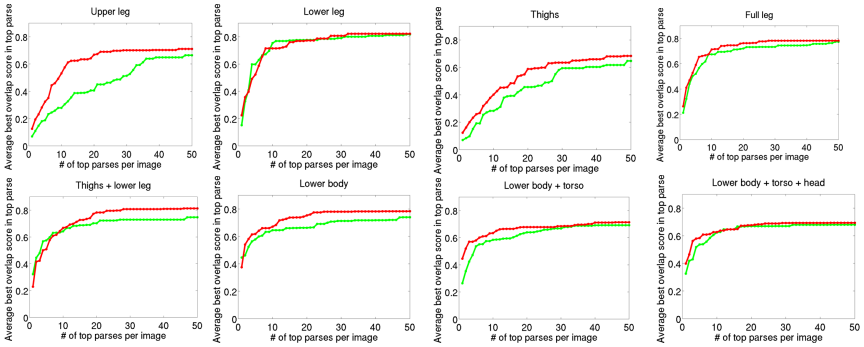

Fig. 9. Average cumulative maximum overlap scores for parsing with and without learning, across all parts. Red curves indicate performance with learning, green represents without learning. Plot methodology is same as that used for left plot in Figure 6.

References

1. S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2002.
2. E. Borenstein and J. Malik. Shape guided object segmentation. In *CVPR 2006*.
3. T. Cour, F. Benezit, and J. Shi. Spectral segmentation with multiscale graph decomposition. In *CVPR 2005*.
4. P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 2005.
5. F. Han and S.-C. Zhu. Bottom-up/top-down image parsing by attribute graph grammar. In *ICCV 2005*.
6. G. Hua, M.-H. Yang, and Y. Wu. Learning to estimate human pose with data driven belief propagation. In *CVPR 2005*.
7. M. W. Lee and I. Cohen. Proposal maps driven mcmc for estimating human body pose in static images. *CVPR 2004*.
8. H. Ling and D. W. Jacobs. Using the inner-distance for classification of articulated shapes. In *CVPR 2005*.
9. D. Lowe. Distinctive image features from scale-invariant keypoints. In *IJCV*, 2003.
10. D. R. Martin, C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2004.
11. G. Mori. Guiding model search using segmentation. In *ICCV 2005*.
12. G. Mori, X. Ren, A. A. Efros, and J. Malik. Recovering human body configurations: combining segmentation and recognition. In *CVPR 2004*.
13. D. Ramanan. Learning to parse images of articulated bodies. In *NIPS 2007*.
14. D. Ramanan, D. A. Forsyth, and A. Zisserman. Strike a pose: Tracking people by finding stylized poses. In *CVPR 2005*.
15. X. Ren, A. C. Berg, and J. Malik. Recovering human body configurations using pairwise constraints between parts. In *ICCV 2005*.
16. R. Ronfard, C. Schmid, and B. Triggs. Learning to parse pictures of people. In *ECCV 2002*.
17. L. Sigal and M. J. Black. Measure locally, reason globally: Occlusion-sensitive articulated pose estimation. In *CVPR 2006*.
18. J. Zhang, J. Luo, R. Collins, and Y. Liu. Body localization in still images using hierarchical models and hybrid search. In *CVPR 2006*.
19. L. Zhu and A. Yuille. A hierarchical compositional system for rapid object detection. In *NIPS 2005*.