Human Parsing using Stochastic And-Or Grammars and Rich Appearances

Brandon Rothrock
U.C. Los Angeles, Computer Science
rothrock@cs.ucla.edu

Song-Chun Zhu
U.C. Los Angeles, Statistics
sczhu@stat.ucla.edu

Abstract

One of the key challenges to human parsing and pose recovery is handling the variability in geometry and appearance of humans in natural scenes. This variability is due to the large number of distinct articulated configurations, clothing, and self-occlusion, as well as unknown lighting and viewpoint. In this paper, we present a stochastic grammar model that represents the body as an articulated assembly of compositional and reconfigurable parts. The reconfigurable aspect allows a compatible part to be substituted with an alternative part with different attributes, such as for clothing appearance or viewpoint foreshortening. Relations within the grammar enforce consistency between part attributes as well as geometry, allowing a richer set of appearance and geometry constraints over conventional articulated models. Part appearances are modeled by a sparse deformable image template that can still richly describe salient part structures. We describe a dynamic programming parsing algorithm for our model, and show competitive pose recovery results against the state-of-art on a challenging dataset.

1. Introduction

Human parsing and pose recovery have broad and far-reaching applications in surveillance, automotive safety, and human-computer interaction. One of the key difficulties is searching the very large configuration space of articulated parts with very cluttered and varied appearance. These variations are due to the contours of different clothing types, textures, colors, lighting, foreshortening due to viewpoint, and self-occlusion. Our approach is motivated from an image parsing and scene understanding perspective, which aims to explain the phenomenon that produces this variability.

Each part in our model is described by its geometric configuration as well as an attribute distinguishing the part specialization among other compatible parts. Although the space of all possible attributed bodies is combinatorially large, it can be represented concisely using a relatively small set of production rules. In conventional grammars such as those used for text, the composite parts formed by these production rules represent abstractions that have no direct evidence in the data except through a composition of the terminal symbols. In the case of image grammars, the presence of scale precludes the notion of having any true terminal parts, and a robust model must be able to detect parts as a whole with and without the presence of subparts. Composite parts are therefore represented by both their image appearance as well as the appearance and geometries of their constituent parts. For example, at the top-level of our grammar we may have production rules for the full body in standing and walking gaits, and each will have appearance models to detect those gaits directly in the image. Each of these gait productions will also contain relations to constrain the subpart geometries to favor those gaits. Each part in the grammar, terminal or non-terminal, has the same parameterization, which allows a recursive formulation and arbitrarily deep hierarchy.

Our stochastic grammar is comprised of a set of production rules. Each production rule defines an appearance model for the root part of the rule, a set of child parts, and a set of pairwise constraints over the geometries and attributes within the rule. The constraints between the child parts makes the grammar context-sensitive, and allows consistency to be enforced between both the pose and attributes of neighboring parts. A probability model is defined on parses of the grammar, and a dynamic programming algorithm is used to compute exact inference for certain restrictions of these constraints. We demonstrate results for human pose recovery on a challenging dataset and show competitive performance with the state-of-art.

2. Related Work and Comparison

We briefly review and contrast some relevant work in terms of body representation and their corresponding geometry and appearance models.

Body representation. The full body is commonly represented with 10 parts, consisting of torso, head, and 2 segments for each limb. Our model adds additional parts for the hands, feet, and pelvis, as well as composite parts for

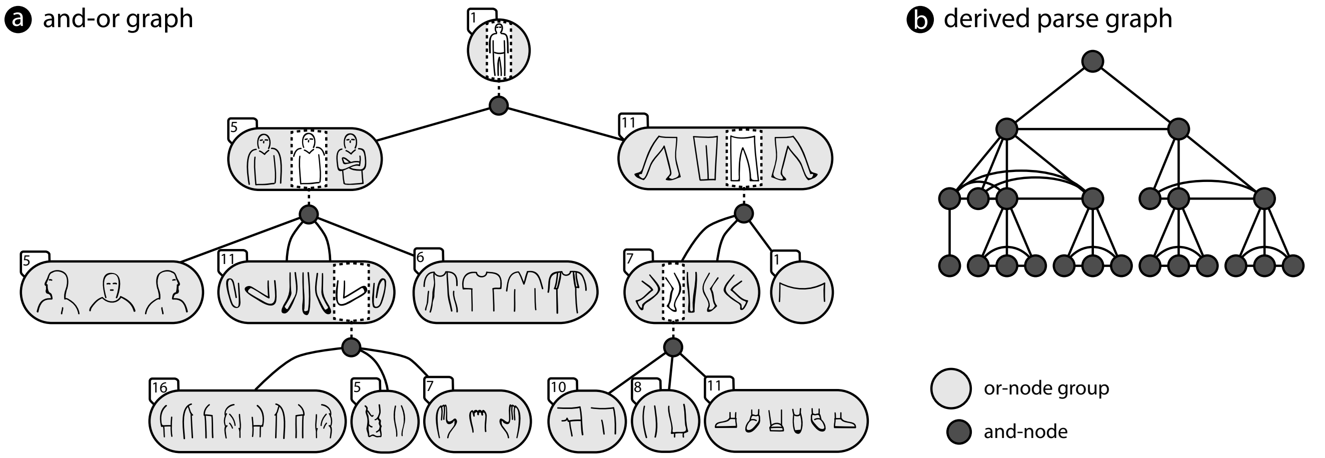

Figure 1. **And-Or graph grammar model**: (a) Pictorially represent the and-or graph, representing or-nodes as ovals containing a selection from multiple forms for a given part type. For each of these forms there is a corresponding and-node, represented as black circles, that specify the composition of that part from smaller subparts. (b) A parse graph is a derivation of the and-or graph, and contains instantiations for each and-node corresponding to every or-node selected in the derivation. The edges of the parse graph represent local geometric and type constraints between parts.

the arms, legs, upper body, lower body, and full body, for a total of 22 parts.

Hierarchical models for complex deformable objects are well motivated, and have been shown to produce promising detection performance [11, 17]. These representations are predominantly used for general object detection, however, and not articulated pose recovery. A notable exception is [19] that uses hierarchy to represent cardinal viewpoints of the body. The grammar model in [2] focuses on simple star-structured parts, and differs from our own by being context-free and handling appearance only through terminal features. Our work can be viewed as a derivation of the grammar frameworks described in [20, 5] by adding articulated geometry constraints and stronger appearance models, however, our model does not use a general MRF to keep the computing complexity of inference tractable. The work of [6] also uses an and-or graph grammar of similar construction, but designed for the task of object detection and segmentation.

Geometry models. The kinematic constraints for the body require that all parts rigidly connect at their joints. The simplest model to capture this property treats kinematic constraints as conditionally independent, which produces a tree-based graphical model. Pictorial structures [10] and much of its related work such as [16, 8, 1] use a geometry model of this form with Gaussian relations on the joints.

One criticism of the tree-based articulated model is that the independence assumptions are too strong to accurately capture realistic body poses. In particular, the position of the arms and legs are typically highly correlated. Incorporating these constraints creates a class of loopy models which are more expressive, but considerably more difficult to compute inference on. The work in [14] explores pairwise geometric constraints between arbitrary parts using Gaussian relations. Similarly, [12] augments the kinematic tree with special exclusion relations to admit efficient inference. Complete graphs under arbitrary scalar-valued potentials were studied in [3] but only for small models at a significant computational cost.

Appearance models. The quality of the appearance model is of critical importance to any image parsing system. Early work focused on finding parallel lines, which poorly detects body parts with general clothing. The shape context model used in [1] and histogram of oriented gradients (HOG) in [7, 4] have been shown to be quite robust, and are used in many modern techniques. Part symmetry is used in [14] and [8, 13] to favor color similarity between symmetric parts such as the arms and legs, but has not been shown to be particularly reliable. Similar to our own approach of learning models for each part specialization, geometry-specific HOG templates for parts are incorporated into a Hough framework in [4], and a mixture model of geometry-specific HOG templates are used in [18]. Our approach differs by incorporating compositional hierarchy, and part specializations that represent varying compositions and appearance, in addition to geometry.

3. And-Or Graph Grammar

Our body grammar is formulated as an and-or graph, which encodes the hierarchical and reconfigurable composition of parts as well as the geometric and selection constraints between parts. Most conventional grammars define a dictionary of terminal parts from which all non-terminal parts are composed. In contrast, the dictionary for our grammar consists of representations for both terminal and non-terminal parts alike. This allows our grammar model to connect to the data at all levels of the hierarchy, as opposed to only through the terminal parts. We believe this is an important property due to the presence of scale, where the evidence for a part may vanish below a certain resolution. In such a case, the object may still be detected by using evidence for parts defined at a coarser level of the hierarchy.

Each part in the dictionary is labeled with a *part form* which uniquely identifies the part, and a *part type* which indicates compatibility such that all forms of the same type are interchangeable with each other. Part compositions are specified by a production rule

$$F_i \rightarrow (\{T_1, T_2, ..., T_n\}, \mathcal{R}_{F_i}) \tag{1}$$

where F_i is a part form, the T's are part types, and $\mathcal{R}_{F_i}$ are the set of relation constraints on the constituent parts. There are two types of constraints in this set: geometric and co-occurrence. Unlike grammars used in text, there is no natural ordering of parts in the image plane, and each production rule must provide a set of geometric constrains to define how the parts can be arranged in the image relative to the root part of the production. The co-occurrence constraints encourage compatible part forms to occur together within the composition.

Part forms are represented by the and-nodes in the and-or graph, and part types are represented by the or-nodes. The part composition for each and-node is defined by production rules of the form shown above. The or-nodes indicate a selection choice between multiple part form alternatives, and are written as $T_i \rightarrow F_1|F_2|...|F_m$ to represent all the part forms of a given part type. The part type at the root of the and-or graph is designated as the start symbol $T_\mathcal{S}$.

The and-or graph can then be defined as the following tuple

$$\mathcal{G} = (V_\mathcal{G}, E_{and}, E_{or}, \mathcal{R}_\mathcal{G}, T_\mathcal{S}) \tag{2}$$

using the dictionary of parts $V_\mathcal{G}$, edges defining part form compositions E_{and}, edges defining part form selections E_{or}, geometric and co-occurrence relations $\mathcal{R}_\mathcal{G}$, and the root part $T_\mathcal{S}$.

A derivation is generated from the and-or graph by recursively selecting a part form from each or-node, starting from the root $T_\mathcal{S}$. For every and-node encountered during the derivation, a *parse node* is instantiated and placed in a *parse graph*. Parts are allowed to be shared in the grammar, and as a result the same node may be visited multiple times during a derivation. Each of these repeated visits will create a unique parse node instantiation in the parse graph. Each parse node specifies the following state

$$pn = (f, x, y, \theta, w, \ell) \tag{3}$$

containing its part form f, position in the image lattice (x, y), orientation θ, width w, and length ℓ. The parse graph is defined as

$$pg = (V_{pg}, E_{pg}, \mathcal{R}_{pg}) \tag{4}$$

where V_{pg} is the set of parse node instantiations, E_{pg} are the corresponding composition edges from E_{and}, and $\mathcal{R}_{pg}$ are the corresponding relations from $\mathcal{R}_\mathcal{G}$.

A pictorial illustration of our and-or graph is shown in figure 1, which shows several examples of part form selections in the or-nodes. At the leaves of the graph, part forms represent distinct appearance classes such as a type of clothing, viewpoint of the foot, or posture of the hand. At higher levels, the part forms largely represent distinct geometric configurations such as an upper body with arms crossed, or lower body in a walking stride. Productions for these composite forms will have constraints on their children that are consistent with these geometric configurations.

The probability model on the and-or graph is a distribution on parse graphs, and is defined in a Bayesian framework

$$p(pg|\mathbf{I}) \propto p(\mathbf{I}|pg)p(pg) \tag{5}$$

using an appearance model $p(\mathbf{I}|pg)$, and prior model $p(pg)$. These models are described in detail in the following sections.

4. Appearance Model

We assume the appearance of each parse node is conditionally independent, and factor the likelihood of the parse graph as follows

$$p(\mathbf{I}|pg) = q(\mathbf{I}_{\overline{\Lambda_{pg}}}) \prod_{v_i \in V_{pg}} p(\mathbf{I}_{\Lambda_{v_i}}|v_i) \tag{6}$$

where $q(\mathbf{I}_{\overline{\Lambda_{pg}}})$ is a background distribution over the image region not occupied by pg, and $\mathbf{I}_{\Lambda_{v_v}}$ is the image patch occupied by part v_i. Although the part appearances are treated as independent given the location of some part form, correlations between appearances are still captured in the relations on part form selection, described in the derivation model.

For each part form in the dictionary, an appearance model is trained using an adaptation of the active basis hybrid image template (HiT) model [15]. Our model constitutes a deformable template of sketch and flat elements positioned in the coordinate frame of the part. Each of these elements are allowed to perturb locally in order to fit small

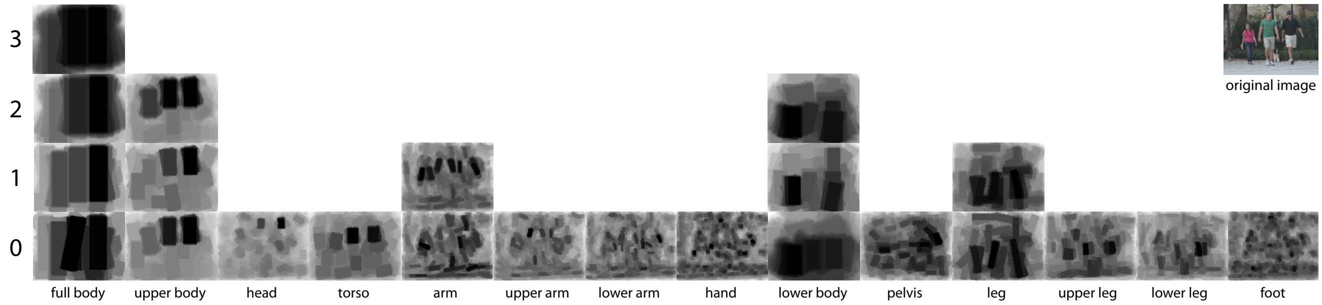

Figure 3. **Part score maps**: Part types are shown horizontally, and the hierarchy depth is shown vertically. The optimal part score maps computed during inference are 5-dimensional for position, orientation, scale, and part form, but visualized here by filling in the bounding box of the corresponding part with the maximal score. These maps are computed by dynamic programming and represent the optimal placement for a part and all its descendants in the hierarchy. As the algorithm proceeds up the grammar hierarchy, larger and larger contexts are incorporated, which cause the maps for the non-terminal parts to be incrementally updated and shown by the vertical maps.

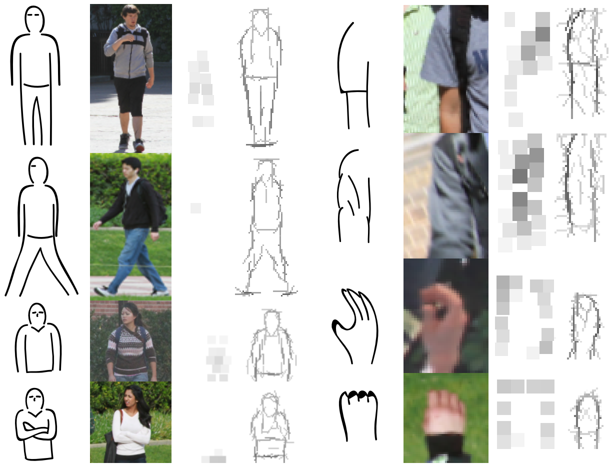

Figure 2. **Learned appearance templates**: shown are examples of 8 of the 119 learned part templates. Each template corresponds to a part form, which are annotated in the training data to be roughly consistent to the hand-drawn sketch on the left. Each template consists of sketch and flat elements, that are allowed to perturb slightly to match small deformations in the image.

variations in the data. Sketch elements are small filters designed to respond to edges at a given orientation. Flat elements, conversely, are designed to respond to local regions with little or no contrast. These features complement each other, and a response from one implies the lack of response from the other.

Each oriented sketch element, as well as the flat element, are referred to as local image prototypes. All prototypes are treated homogeneously, and each prototype need only compute a 1D response given a local image patch, written as $r(\mathbf{I}_\Lambda)$. The active basis model assumes the following log-linear form

$$p(\mathbf{I}_{\Lambda_{v_i}}|v_i) = q(\mathbf{I}_{\Lambda_{v_i}}) \prod_{j=1}^{n} \left[\exp\{\lambda_j r_j(\mathbf{I}_{\Lambda_j}) - \log z_j\}\right] \quad (7)$$

where λ_j are the model parameters, and $\log z_j$ are the normalizing constants.

The template is trained by projection pursuit, where at each iteration the best prototype element is added to the model such that the revised model moves closest to the target distribution f beginning with a background distribution q. Elements continue to be pursued and added to the model until the model is sufficiently close to f. In other words, at each iteration we wish to maximally reduce the Kullback-Leibler divergence between the current model p_k and the target distribution f. Observing that $KL(f||p_{k-1}) - KL(f||p_k) = KL(p_k||p_{k-1})$, this is equivalent to maximizing $KL(p_k||p_{k-1})$ at each iteration. This is referred to as maximizing the information gain, and can be computed as:

$$IG_k = KL(p_k||p_{k-1}) = \lambda_k E_f[r_k] - \log z_k \quad (8)$$

The expectation $E_f[r_k]$ is approximated by using the sample mean of responses from the positive training images. The normalization constant z_k and parameters λ_k are estimated using importance sampling. It is assumed that the prototype responses at different locations in the image are not correlated, and to enforce this a small neighborhood around each selected prototype is suppressed to prevent potentially correlated elements from being selected in future iterations.

Parameter estimation for this model follows maximum-likelihood, which does not use negative examples except for pooling marginal background statistics. Detection performance can be significantly improved when the parameters are reestimated under a discriminative criteria. For this we use a logistic regression model, which has the following form

$$p(y|\mathbf{x}) = \frac{1}{1 + \exp\{-y(\lambda^T \mathbf{x} + b)\}} \tag{9}$$

where y is the positive or negative class ± 1, l is the number of training examples, b is the bias, and C controls the influence of the regularization term. The feature vector $\mathbf{x}$ consists of responses from all the prototype elements in the template $(r_0, r_1, ..., r_n)$. The parameters λ are solved by minimizing the following regularized negative log-likelihood

$$\frac{1}{2}\lambda^T \lambda + C \sum_{i=1}^{l} \log(1 + \exp\{-y(\lambda^T \mathbf{x}_i + b)\}). \tag{10}$$

Once λ and b are estimated using this criteria, the λ's can be substituted directly back into the HiT model, and the bias b can replace all the individual $\log z$'s. For our experiments, we use the parameter $C = .1$ and the code from [9] to compute this optimization.

Next, we describe how the responses are computed for both the sketch and flat prototypes. Once responses are computed for each of the prototypes, the are allowed to perturb locally and independently. Computing the template response in this case simply involves taking a local max for each element within this local neighborhood.

Sketch: The sketch prototype represents a short edgelet at a specific orientation. The original active basis model uses a dictionary of Gabor basis filters at multiple orientations to compute sketch responses. Gabor filters, however, respond poorly to orientations at small scales. Instead, we use a gradient-based feature that achieves a similar result but is faster to compute and allows much smaller templates to be learned.

The image gradient is first computed and the gradient orientation for each pixel is discretized. For each orientation, an aggregate gradient responses in the local neighborhood of each pixel is computed by convolving responses only for that orientation with an elongated Gaussian filter oriented orthogonal to the gradient direction. This will produce large responses for edge segments in the same orientation of the filter.

At each location, a normalization is computed by summing responses over all orientations. The final response value for each location and orientation is then computed by dividing by the average normalization value within a local neighborhood.

The resulting feature is very similar to the popular HOG feature [7], except that the response is computed at every location, and the gradients are pooled along oriented elliptical regions instead of the square histogram cells used in HOG. It is also important for us that the features be rotatable for computing responses of an appearance template at multiple orientations, which is difficult to do with HOG. In our experiments we use 16 orientations over π and a 7×7 Gaussian filter to pool the gradient responses.

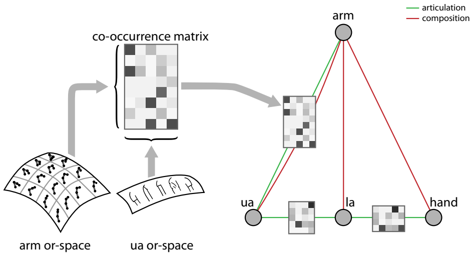

Figure 4. **Or-node co-occurrence**: the arm is composed from an upper arm (ua), lower arm (la), and hand. Local compatibility between the forms (or-node selections) of these parts are enforced by pairwise co-occurrence matrices measured between the forms of articulated pairs.

Flat: The flat prototype represents regions where there is little or no contrast. This is computed by averaging the magnitude of the image gradient over a small neighborhood around each image location using an integral image. The resulting average magnitude is then transformed using a negative sigmoid to favor regions with very small gradients.

5. Prior Model

The prior model is the product of a derivation model and geometry model

$$p(pg) = p_d(pg) \cdot p_g(pg). \tag{11}$$

The state variables of the parse nodes within pg contain the forms for all the parts, as well as their geometric states. The derivation model defines a probability on the part forms, whereas the geometry model defines a probability on their geometric states.

The set of relations $\mathcal{R}_{pg}$ contains pairwise edges that indicate constraints on the part forms and geometries of the corresponding parse nodes. We restrict these relations to not allow cycles in order to admit efficient inference. The parts of the human body naturally forms an articulated tree structure, and this is the topology we use for the relation set. One difference in our model from most articulated models is that it is also hierarchical, and therefore each parent node in pg must have a relation defined with at least one of its children. For example, an arm may decompose to an upper-arm and lower-arm, with an articulation relation between upper-arm and lower-arm, but also between upper-arm and arm.

5.1. Derivation Model

Our grammar model extends the stochastic context-free grammar (SCFG) case by allowing the selection of part forms to depend on the forms of neighboring siblings. In general, this can be represented as a product of joint probabilities of sibling forms given their parent

$$p_d(pg) = \prod_{(ij)\in E_{pg}} p(f(C(v_i))|f_i) \tag{12}$$

where $C(v_i)$ is the set of children and $f(C(v_i))$ is the corresponding child forms of parent node v_i. This allows correlations between parts forms to be modeled, for example, a short-sleeve upper arm occurs frequently with a bare skin lower arm, but never with a long-sleeve lower arm. In this case, the bare-skin lower arm template has a distinct taper not present in the sleeved template, and a strong appearance response from one will influence the upper-arm form selection. In the SCFG case, these two forms would occur independently.

We further factorize this model according to the tree model defined in $\mathcal{R}_{pg}$. Let $\mathcal{R}_{pg}(v_i)$ be the set of edges in $\mathcal{R}_{pg}$ between node v_i and any of its children $C(v_i)$. This factorization can now be expressed as

$$p(f(C(v_i))|f_i) = \frac{\prod_{(j,k)\in\mathcal{R}_{pg}(v_i)} p(f_j, f_k|f_i)}{\prod_{v_j\in C(v_i)} p(f_j|f_i)^{d(v_j)-1}} \tag{13}$$

where $d(v_j)$ represents the edge degree of node v_j according to the edges in $\mathcal{R}_{pg}(v_i)$. The marginal probabilities $p(f_j, f_k|f_i)$ and $p(f_j|f_i)$ are recorded as co-occurrence histograms from the training data, and illustrated in figure 4.

5.2. Geometry Model

The geometry model defines a probability distribution on the geometric state variables for all parse nodes in the parse graph. This model uses the same tree factorization defined by the relation set $\mathcal{R}_{pg}$, and is written as follows

$$p_g(pg) = \prod_{(ij)\in\mathcal{R}_{pg}} p(v_i, v_j). \tag{14}$$

We use a geometry model very similar to the pictorial structures model described in [10], which defines Gaussian relations for each pair of articulated parts. In the coordinate frame of the image, the Gaussian model is a fairly poor fit as articulated parts tend to lie in arc-like regions around their neighbors. By transforming the part coordinates to the reference frame of their common joint, however, this model becomes quite reasonable. Using $\mathbf{x} = (x, y, \ell, \theta)$ as shorthand for the geometric state of some parse node, these transformations are defined as

$$\begin{aligned} T_{ij}(\mathbf{x}_i) &= (x'_i, y'_i, \ell_i, \cos(\theta_i + \theta_{ij}), \sin(\theta_i + \theta_{ij})) \\ T_{ji}(\mathbf{x}_j) &= (x'_j, y'_j, \ell_j, \cos(\theta_j), \sin(\theta_j)) \\ \Sigma_{ij} &= diag(\sigma_x^2, \sigma_y^2, \sigma_\ell^2, 1/k, 1/k). \end{aligned}$$

where k is the parameter of the Von Mises distribution over angles, and x' and y' are discrete locations in the transformed space. $T_{ij}(\mathbf{x}_i)$ is the transformation of part v_i to the reference frame of the joint connecting it to part v_j. The joint probability between part pairs is then a zero-mean normal distribution in the transformed coordinates

$$p(v_i, v_j) \propto \mathcal{N}(T_{ji}(\mathbf{x}_j) - T_{ij}(\mathbf{x}_i), 0, \Sigma_{ij}). \tag{15}$$

6. Parsing as Bayesian Inference

Parsing is the process of finding the parse graph with maximal posterior probability

$$\begin{aligned} pg^* &= \arg\max_{pg} p(pg|\mathbf{I}) && (16) \\ &= \arg\max_{pg} p(\mathbf{I}|pg)p_d(pg)p_g(pg). && (17) \end{aligned}$$

Given the hierarchical structure of the model, the log-posterior can be formulated as a recursive scoring function given a parse node v

$$s(v|\mathbf{I}) = \overbrace{s_a(v|\mathbf{I})}^{appearance} + \overbrace{s_g(v)}^{geometry} + \overbrace{s_d(v)}^{derivation} + \overbrace{\sum_{v_i\in C(v)} s(v_i|\mathbf{I})}^{children}$$

$$\begin{aligned} s_a(v|\mathbf{I}) &= \log p(\mathbf{I}_{\Lambda_v}|v) = \sum_{i=1}^{n} [\lambda_i r_i(\mathbf{I}_{\Lambda_i}) - \log z_i] \\ s_g(v) &= \sum_{v_i\in\mathcal{R}_{pg}(v)} \log p(v, v_i) \\ s_d(v) &= \log p(f(C(v))|f(v)). && (18) \end{aligned}$$

For an image $\mathbf{I}$, parsing is now defined as finding the parse graph pg with root part v_0 of type T_S that maximizes the score function $pg^* = \arg\max_{pg} s(v_0)$.

6.1. Inference Algorithm

Dynamic programming is used to take advantage of this recursion by tabulating the maximal scores for each part type, which can be used as a lookup in the maximization of their parent compositions. For each part form in the grammar, two tables are allocated to store their appearance scores and optimal composition scores. The size of these tables represent the discretization of all possible part geometries, i.e. locations, orientations, and scales. Let $S_A(F_i)$ and $S_C(F_i)$ be the appearance and optimal composition score tables corresponding to part form F_i.

The base case is at the leaves of the and-or graph, which have no children, leaving only the appearance term $s^*(v|\mathbf{I}) = \max_v s_a(v|\mathbf{I})$. Computing this simply involves populating the appearance tables S_A for all forms of v. In the case where v is non-terminal part, there are now child parts and relations between them. The maximization needs to find the optimal geometries and part forms for the part v

and its subparts $C(v)$. The relations within this set of parts $\mathcal{R}_{pg}(v)$ is used to incrementally compute the optimal composition scores. These relations form a tree rooted at v. Let $C_{\mathcal{R}}(v)$ be the set of children of v according to this tree. For a given relation edge (i, j), let m and n be the number of forms that parts v_i and v_j can take respectively.

The optimal location for the child part v_j given the parent v_i can then be expressed as

$$\begin{aligned} B_{v_j}(v_i) &= \max_{f_j=1,\dots,n} \left(\max_{\mathbf{x}_j} S(v_i, v_j) \right) \\ S(v_i, v_j) &= s^*(v_j) + \log \frac{p(f_i, f_j)}{p(f_j)^{|C_{\mathcal{R}}(v_j)|}} + \\ & \log p(v_i, v_j) + \sum_{v_k \in C_{\mathcal{R}}(v_j)} B_{v_k}(v_j). \end{aligned} \tag{19}$$

The outer maximization is over all the different forms for the distal part ρ_j. The inner maximization is over all the geometries of ρ_j. The terms inside are the appearance scores, an incremental derivation score, and an incremental geometry score from equation 18. Because the incremental derivation score is constant within the inner maximization, a generalized distance transform [10] can compute the maximization in time linear in the number of grid locations. The outer maximization over part forms must be done explicitly with quadratic complexity in the number of forms, and the resulting output of B_{v_j} is a set of m tables for each form of v_i. Once B is computed for the root part of $\mathcal{R}(v)$, the resulting scores are stored in the corresponding table S_C.

The algorithm computes bottom-up starting from the terminals and working toward the root by computing tables S_C of optimal scores for each part form encountered. An additional backtrack table S'_C is also computed to store the table indices of the child parts to produce the optimal score and production for each location of the root part. This is done by simply replacing the $\max$ with an $\arg\max$. Once the algorithm reaches the start node, the location in the root table with maximal score will be the globally optimal solution. The full parse can be recovered by backtracking all the part forms and geometries using their corresponding backtrack tables. A visualization of the intermediate score maps computed during the inference process is shown in figure 3.

7. Experiments

We collected our own dataset of roughly 400 outdoor pedestrians in natural poses, which are annotated with full parse graphs. This was motivated by the lack of a high-resolution full body dataset to learn detailed part models from. Fortunately, the authors of the current state-of-art method [18] has made their code available, allowing us to train their model on the same data for a direct comparison. For this experiment, we trained a full-body grammar model using 109 part forms, consisting of 69 terminal and 40 non-terminal parts. The large number of terminals correspond to different clothing of the arms, poses of the hands and feet, orientations of the head, etc, as shown in 1. The detection performance of our method is very competitive with the state-of-art, and surpasses their performance on the lower arms and legs, which are typically the most difficult parts to localize. This is largely due to the localization of non-terminal parts such as the upper-body or arm, which help inform the location of its subparts such as lower-arm even when evidence for that part is very weak or occluded.

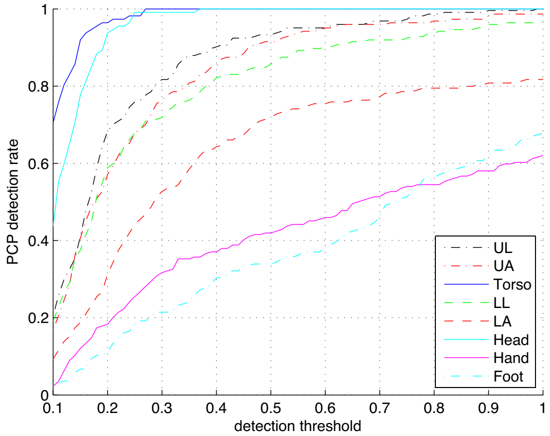

Figure 6. **Parsing evaluation**: a part is considered detected if both endpoints lie within a proportion of the ground truth part length. The percentage of correctly estimated parts (PCP) [8] is shown as a function of this threshold for each of the terminal parts.

Method of Yang et al. [18]						
head	torso	u.leg	l.leg	u.arm	l.arm	avg
1.000	1.000	.975	.839	.951	.577	.869
Our model						
head	torso	u.leg	l.leg	u.arm	l.arm	avg
1.000	1.000	.933	.857	.915	.719	.884
hand	foot					
.420	.339					

Table 1. **Pose recovery results on our outdoor pedestrian dataset**: shown are part detection rates of our method compared with the current state-of-art using a PCP threshold of 0.5. Results are computed using the same 10-part full body parameterization, although our model also produces detections for hand and foot.

8. Conclusions

We describe a stochastic image grammar for human parsing that incorporates compositional and reconfigurable parts and context-sensitive constraints to explicitly capture the vast variabilities of both articulated pose geometry and appearance of the human body. We demonstrate the viabil-

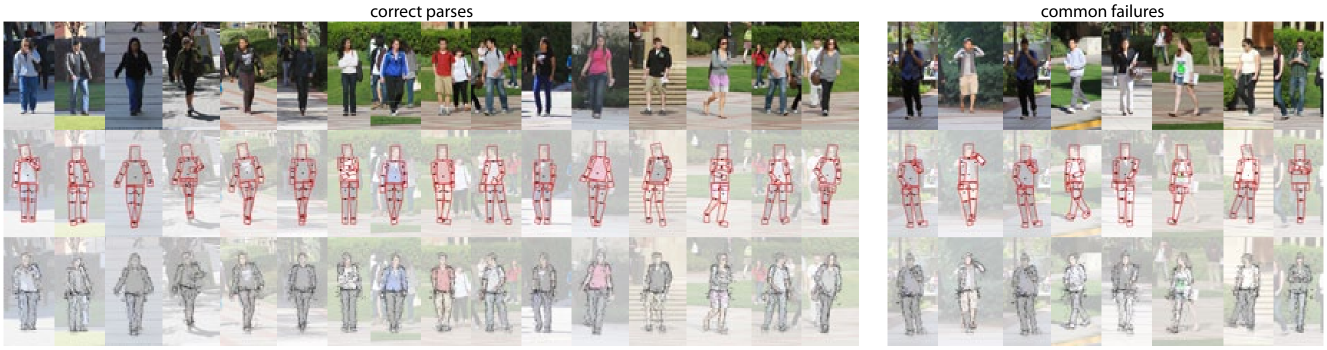

Figure 5. **Parse results**: shown here are 24 parse results from a test set of 150 images, trained on 250 examples. The detection curve for these results are shown in figure 7. The top row is the original image, the center row shows bounding boxes around the detected terminal part geometries, and the bottom row shows the matched appearance template for all parts in the parse graph. The 8 rightmost parses show some failures, which most commonly include matching a lower arm or leg to the background, or double-counting a limb.

ity of our technique by showing pose recovery performance that is competitive with the state-of-art. This representation can be easily adapted to a wide variety of deformable object classes. Furthermore, the attributes inferred from the parser can potentially be used to infer higher-level state such as gender or activity. Our dataset and the source code of our parser will be released with the publication of this paper.

Acknowledgements: Funding for this research was provided by NSF Grant IIS-1018751, NSF CNS Grant 1028381, and ONR MURI Grant N00014-10-1-0933.

References

[1] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR*, pages 1014–1021, 2009. 2

[2] M. Aycinena, L. P. Kaelbling, and T. L. Perez. Learning grammatical models for object recognition. In *Technical Report, AI Lab, Massachusetts Institute of Technology*, 2008. 2

[3] M. Bergtholdt, J. Kappes, S. Schmidt, and C. Schnörr. A study of parts-based object class detection using complete graphs. *Int. J. Comput. Vision*, 87:93–117, March 2010. 2

[4] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *International Conference on Computer Vision*, sep 2009. 2

[5] H. Chen, Z. J. Xu, Z. Q. A. Liu, and S. C. Zhu. Composite templates for cloth modeling and sketching. In *CVPR*, pages I: 943–950, 2006. 2

[6] Y. Chen, L. Zhu, C. Lin, A. L. Yuille, and H. Zhang. Rapid inference on a novel AND/OR graph for object detection, segmentation and parsing. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *NIPS*. MIT Press, 2007. 2

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages I: 886–893, 2005. 2, 5

[8] M. Eichner and V. Ferrari. Better appearance models for pictorial structures. In *BMVC*, pages xx–yy, 2009. 2, 7

[9] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008. 5

[10] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, jan 2005. 2, 6, 7

[11] S. Fidler and A. Leonardis. Towards scalable representations of object categories: Learning a hierarchy of parts. In *CVPR*, pages 1–8, 2007. 2

[12] H. Jiang and D. R. Martin. Global pose estimation using non-tree models. In *CVPR*, pages 1–8, 2008. 2

[13] D. Ramanan. Learning to parse images of articulated bodies. In *In NIPS 2007*. NIPS, 2006. 2

[14] X. F. Ren, A. C. Berg, and J. Malik. Recovering human body configurations using pairwise constraints between parts. In *ICCV*, pages I: 824–831, 2005. 2

[15] Z. Z. Si, H. F. Gong, Y. N. Wu, and S. C. Zhu. Learning mixed templates for object recognition. In *CVPR*, pages 272–279, 2009. 3

[16] L. Sigal and M. J. Black. Measure locally, reason globally: Occlusion-sensitive articulated pose estimation. In *CVPR*, pages II: 2041–2048, 2006. 2

[17] S. Todorovic and N. Ahuja. Region-based hierarchical image matching. *International Journal of Computer Vision*, 78(1):47–66, June 2008. 2

[18] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, 2011. 2, 7

[19] J. Y. Zhang, Y. X. Liu, J. B. Luo, and R. T. Collins. Body localization in still images using hierarchical models and hybrid search. In *CVPR*, pages II: 1536–1543, 2006. 2

[20] S. C. Zhu and D. Mumford. *A Stochastic Grammar of Images*. World Scientific, 2007. 2