A Deformable Mixture Parsing Model with Parselets

Jian Dong[1], Qiang Chen[1], Wei Xia[1], Zhongyang Huang[2], Shuicheng Yan[1]
[1] Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[2] Panasonic Singapore Laboratories, Singapore
{a0068947, chenqiang, weixia, eleyans}@nus.edu.sg,{zhongyang.huang}@sg.panasonic.com

Abstract

In this work, we address the problem of human parsing, namely partitioning the human body into semantic regions, by using the novel Parselet representation. Previous works often consider solving the problem of human pose estimation as the prerequisite of human parsing. We argue that these approaches cannot obtain optimal pixel level parsing due to the inconsistent targets between these tasks. In this paper, we propose to use Parselets as the building blocks of our parsing model. Parselets are a group of parsable segments which can generally be obtained by low-level over-segmentation algorithms and bear strong semantic meaning. We then build a Deformable Mixture Parsing Model (DMPM) for human parsing to simultaneously handle the deformation and multi-modalities of Parselets. The proposed model has two unique characteristics: (1) the possible numerous modalities of Parselet ensembles are exhibited as the "And-Or" structure of sub-trees; (2) to further solve the practical problem of Parselet occlusion or absence, we directly model the visibility property at some leaf nodes. The DMPM thus directly solves the problem of human parsing by searching for the best graph configuration from a pool of Parselet hypotheses without intermediate tasks. Comprehensive evaluations demonstrate the encouraging performance of the proposed approach.

1. Introduction

Human parsing [31] has drawn much attention recently for its wide applications in human-centric analysis, such as person identification [16] and clothing analysis [7, 21]. The success of human parsing relies on the seamless cooperation of human pose estimation [32], segmentation [2], and region labeling [31]. However, previous works often consider solving the problem of human pose estimation as the prerequisite of human parsing [31]. We argue that these approaches cannot obtain optimal pixel level parsing due to the inconsistent targets of these tasks.

In this paper we aim to develop a unified framework for human parsing. To this end, we reconsider the basic level representation. Although the key points [33] or rigid templates [32, 12] representation can facilitate the localization of human parts, leading to great success in human detection and pose estimation [32], it fails to provide accurate pixel-level labeling. This limitation hinders key points or templates to be the ideal building blocks for human parsing. On the other hand, there exists exciting progress of bottom-up region hypotheses based segmentation methods [5, 10], which have achieved the state-of-the-art performance [11]. More specifically, region hypotheses based segmentation is performed by first generating extensive object hypotheses based on bottom-up information and then ranking them, with the critical assumption that the object has a large probability to be tightly covered by at least one of the generated hypotheses. This assumption usually holds well for objects with homogeneous appearance. However, for objects with large appearance variance, finding a single region hypothesis to tightly cover the whole object is very difficult.

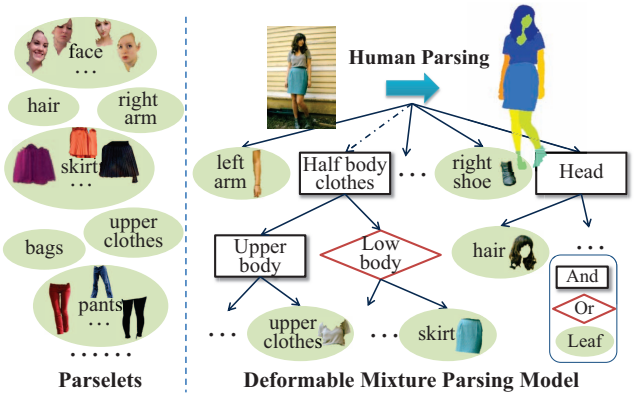

Figure 1: Parselets are image segments that can generally be obtained by low-level segmentation techniques and bear strong semantic meaning. The instantiated Parselets, which are activated by our Deformable Mixture Parsing Model, provide accurate semantic labeling for human parsing.

Based on the above observation, we propose to use **Parselets** as the building blocks for human parsing as shown in Fig. 1. The Parselets are a group of semantic image segments with the following characteristics: (1) they can generally be obtained by low-level over-segmentation

algorithms [3, 1], *i.e*. they are parsable by bottom-up techniques; (2) they have strong and consistent semantic meaning, *i.e*. they are parsable by the human knowledge. An object consisting of parts with large variance usually cannot be well segmented out by the low-level segmentation methods, *e.g*. a human body cannot be perfectly segmented by edge-based segmentation [3]. However, we argue that the localized semantic regions, *e.g*. the skirt or hair area of human in Fig. 1, often show homogeneous appearance and can be segmented out as segments. Such image segments, denoted as Parselets, explicitly encode segmentation and semantic level information.

With the Parselet representation, we propose the Deformable Mixture Parsing Model (DMPM) for human parsing. DMPM is represented as an "And-Or" graph [33] based hierarchical model to simultaneously handle the deformation and multi-modalities of Parselets. The joint learning and inference of best configuration for both appearance and structure in our DMPM guarantee the overall performance. We perform human parsing by generating extensive hypotheses for Parselets and subsequently assembling them by DMPM. The major contributions of this work can be summarized as follows:

- We propose the novel Parselet representation. By explicitly encoding segmentation and semantic information, Parselets serve as ideal building blocks for human parsing models. Human parsing is then performed with the Parselet representation, rather than with the key point [33] or rigid template [32, 12] representation. The instantiated Parselets directly provide accurate pixel-level semantic information. In practice, several over-segmentation techniques are utilized to ensure the high recall rate of Parselets.

- We build a novel Deformable Mixture Parsing Model (DMPM) for human parsing. The "co-occurrence" and "exclusive" modalities of Parselets are exhibited as the "And-Or" structure of sub-trees. To further solve the problem of Parselet occlusion or absence, we directly add the "visibility" property at the corresponding nodes. Joint learning and inference of appearance and structure parameters guarantee the overall performance. In addition, the tree structure of our DMPM allows efficient inference.

- In order to verify the effectiveness of the proposed framework, we construct a high resolution human parsing dataset consisting of 2,500 images. All the pixels in the images are thoroughly annotated with 18 types of Parselets. As far as we know, this is the largest human dataset with full parsing labels. It could serve as the benchmark for segmentation-based human analysis in the research community.

2. Related Work

Selective Search for Recognition: Selective search approaches for object recognition have achieved great success in the past few years [24, 10, 27, 5, 2, 4]. This line of works first generate a set of object hypotheses based on bottom-up information and then convert the recognition problem into a ranking problem. Compared with exhaustive sliding window scanning [9, 12], selective search usually enables more expensive and potentially more powerful recognition techniques [28, 27]. Our work differs from the above works significantly as we focus on parts instead of whole objects. We claim that region hypotheses are better hypotheses for parts than for objects toward categories with heterogeneous appearance. Gu et al. [17] also addressed the problem of segmenting and recognizing objects based on their parts. They generated part hypotheses and then formulated the problem in the generalized Hough transformation framework. Our work differs from this work significantly as their work focuses on the segmentation and is unable to exploit the hierarchical structure of the object.

Part Based Model: Hierarchical part based models can better grasp the complicated structure than rigid models and thus usually achieve better performance for articulated objects [12, 32, 34]. Pictorial Structure (PS) based methods [13, 12, 32] are the most common approaches for pose estimation and object recognition. However, unlike our DMPM, part templates are usually spread in all nodes of PS based models, which makes it inconvenient to model complicated composite relation. The stochastic image grammar model [33, 8] is also effective for modeling the hierarchical structure. However, these models rely on complex learning and inference procedures which can only be made tractable using approximate algorithms [25]. On the contrary, despite the sophisticated structure of DMPM, we show that a tractable and exact inference algorithm exists.

Human Parsing: Human parsing, namely partitioning the human body into semantic regions, plays an important role in many human-centric applications [7, 21, 22, 30, 20]. Torr and Zisserman proposed an approach for simultaneous human pose estimation and body part labeling under the CRF framework [26], which can be regarded as a continuation of combining segmentation and human pose estimation [19]. Yamaguchi et al. [31] performed human pose estimation and attribute labeling sequentially for clothing parsing. Our method differs from these methods as previous research on human parsing tends to first align human parts [32] due to the large pose variations or the complexity of the models. However, such sequential approaches may fail to capture the correlations between human appearance and structure, leading to unsatisfactory results. The proposed DMPM, which can solve human parsing in a unified framework, significantly distinguishes our work from others.

Figure 2: Human decomposition based on different basic elements. The original image, Parselet based decomposition and joint based decomposition are shown sequentially.

3. Parselets

Parselets lie at the heart of our human parsing framework. In this section, we first give the definition of human Parselets. Then we present the details of hypothesis generation and feature representation for Parselets. And finally, we briefly introduce the modalities of Parselet ensembles.

3.1. Parselet Definition

We notice that the classical part-based models [13, 32] usually divide body into parts based on joints. However, such decomposition is unsuitable for segment hypotheses because joint-based parts usually do not correspond to the segments from bottom-up cues. Considering the left image in Fig. 2, the whole dress is likely to be captured by a single segment from the bottom-up techniques. But for the right image, the upper clothes, coat and pants should intuitively correspond to three separate segments. This difference is hard to be grasped by joint based decomposition. To overcome this limitation, we propose the Parselets to serve as the building elements for our parsing model. Formally, the **Parselets** are a group of semantic image segments which have the following characteristics: (1) they can generally be obtained by low-level segmentation algorithms [3, 1, 5], *i.e*. they are parsable by the bottom-up techniques. This characteristic guarantees that Parselets can be retrieved with high possibility by the bottom-up hypothesis generation schemes. (2) They bear strong and consistent semantic meaning, *i.e*. they are parsable by the human knowledge. Since our ultimate goal is to perform human parsing, the basic elements of the parsing model should have clear semantic meaning.

We now decompose human body into homogeneous regions based on low-level cues. The homogeneous regions, which have clear semantic meaning and appear in many different images, are defined as Parselets. Through careful design, each defined Parselet will have high probability to form a single segment. Specifically, we define 18 types of Parselets as described in Table 1. These Parselets are representative and can properly cover most of human body. They engage about 98.4% of human body in our labeled datasets and can be obtained with high recall rate using the method introduced in Section 3.2. Detailed statistics are shown in the experiment section. It is worth noting that the Parselet definition is flexible to be redesigned for different applications. The only assumption here is that those semantic regions can be segmented out with high probability.

Table 1: 18 types of Parselets for human

	Parselets		
Head	hat	hair	sunglasses
Body	upper clothes skirt	coat pants	full body clothes
Foot	left/right shoe		
Skin	face	left/right arm	left/right leg
Accessory	bag	scarf	belt

3.2. Hypothesis Generation for Parselets

In order to obtain the Parselet hypotheses with high recall rate, we combine several low-level segmentation methods. As Parselets usually appear in different scales, the hierarchical segmentation algorithm should be a natural way to generate hypotheses. Here, we choose Ultrametric Contour Map (UCM) [3], which works well to preserve the boundary information. However, the merging scheme of UCM proceeds by removing the edge with smallest probability and thus only neighboring super-pixels can be merged. This may prevent non-adjacent segments from merging as a single segment and lead to unsatisfactory results for some Parselets, which are separated by noise segments. For example, the dress in the left image of Fig. 2 is split into separate segments by the stripe pattern with strong edges. Hence UCM fails to merge them in the early stage. In addition, some garments, such as a belt, may also divide a Parselet into separate segments. To handle these difficulties, we add another appearance based segmentation and merging scheme. Specifically, we first use the fast appearance based over-segmentation method [1] and sequentially merge the nearby (not necessarily adjacent) regions with the smallest similarity score in a similar manner as in [27]. We define the similarity score S between segments a and b as $S(a,b) = S_{size}(a,b) + S_{appearance}(a,b)$, both of which are normalized to [0,1]. $S_{size}(a,b)$ is defined as the fraction of the image that the region a and b jointly occupy. This factor encourages small regions to be merged early. $S_{appearance}(a,b)$ is defined as the χ^2 distance of the color and SIFT [23] histogram of segments a and b [29]. Finally, we utilize another complementary scheme, namely CPMC [5], which directly generates many segments of different scales. The segments from the above three methods are combined into the final Parselet hypothesis.

3.3. Feature Representation

Compared with exhaustive sliding window scanning [9, 12], our Parselet based representation enables complex and expensive feature design. It has been shown that the bag of words feature performs better than the rigid template for categories with large pose and view variance [28, 27, 5]. As our Parselet categorization is essentially a classification problem, we follow the state-of-the-art feature extraction-coding-pooling classification pipeline [15, 6, 4]. In this work, we adopt the Fisher Kernel (FK) + average pool-

ing [15] and enhanced feature + second order pooling [4], which have been shown with the best performance among current BoW encoding methods. In addition, as our algorithm only employs the size and appearance features which can be efficiently propagated throughout the hierarchical structure embedded in the pools of segments, the feature extraction is reasonably fast.

3.4. Parselet Ensemble

Parselets serve as the building blocks of our human parsing model. The Parselets are low-level parts from the definition. In practice, several Parselets are often grouped together in order to form the middle-level human body part, *e.g*. head, body, etc. Those middle-level parts cannot be represented by a single type of Parselets but can be modeled by the ensembles of Parselets. More specifically, the ensembles of Parselets show two kinds of modalities as follows: (1) **Co-occurrence**. The modality of co-occurrence represents the relation that several types of Parselets coexist and are merged to form a larger middle-level human part. This is the most typical modality of Parselet ensembles. For example, the "hair" usually comes with "face" to form the "head". (2) **Exclusivity**. The modality of exclusivity models the relationship of different types of Parselets that cannot coexist logically. For example, for the "lower-body" area, there are two possible Parselets, *i.e*. "skirts" and "pants". However, "skirts" and "pants" usually cannot coexist. The exclusivity for the middle-level concept "lower-body" means that only one of the two exclusive Parselets, *i.e*. "skirts" and "pants", can exist for the "lower-body".

The middle level concepts formed from Parselet ensembles can be further merged with Parselet(s) or other middle level concepts. They also exhibit co-occurrence or exclusivity modalities to form an even higher level concept. This higher level concept thus inherits all the information from its sub-components. This inheritance property guarantees that we can model complex objects (*e.g*. human) with multiple levels of concepts.

4. Human Parsing over Parselets

With the Parselets and their ensembles, we propose the Deformable Mixture Parsing Model (DMPM) for human parsing. Specifically, we propose to employ an "And-Or" graph [33] based hierarchical model to simultaneously handle the deformation and multi-modalities of Parselets. The "co-occurrence" modality is modeled as the "And" relation while "exclusivity" modality is modeled as the "Or" relation in the graph. The deformation is modeled as pairwise parent-child distance. We construct a hierarchical model, as hierarchical models have been shown to be effective for grasping the structure of objects in part based approaches [32, 33]. In addition, absence/occlusion is common for some Parselets. Hence we explicitly model this by utilizing a special structure call virtual "Leaf" node. Fig. 3

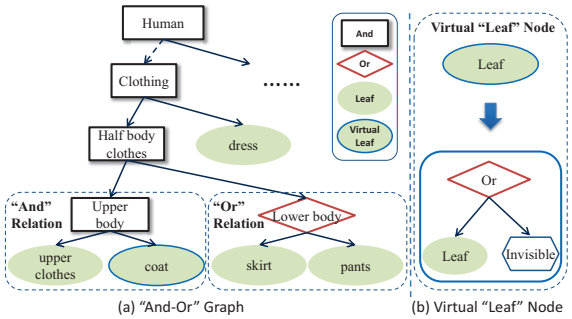

Figure 3: The subgraph from our human "And-Or" graph. The diamonds, rectangles, eclipses and eclipses with boundary represent "Or" nodes, "And" nodes, "Leaf" nodes and virtual "Leaf" nodes, respectively.

shows a subgraph from our human graph, while the full graph of our parsing model is listed in the supplemental file. In the next subsections, we will introduce our DMPM followed by the inference and learning algorithms.

4.1. Deformable Mixture Parsing Model

We first define the notations used in the following section. P represents the Parselet hypothesis segments in an image generated according to Section 3.2. For a hypothesis segment with index i, its scale (the square root of its area) and centroid are donated as s_i and $c_i = (x_i, y_i)$. Formally, a DMPM model is represented as a graph $G = (V, E)$ where V is the set of nodes and E is the set of edges. The edges are defined by the parent-child structure and kids(ν) denote the children of node ν. There are three basic types of nodes, "And","Or" and "Leaf" nodes which specify different parent-child relationships as depicted in Fig. 3 by diamonds, rectangles and eclipses respectively. Each "Leaf" node corresponds to one type of Parselets.

The state variables of the graph specify the graph configuration. Specifically, the graph topology is instantiated by a switch variable t at "Or" nodes, which indicates the set of active nodes $V(t)$. Starting from the top level, an active "Or" node $\nu \in V^O(t)$ selects a child $t_\nu \in \text{kids}(\nu)$. The active "And" or "Or" nodes have the state variables $g_\nu = (s_\nu, c_\nu)$ which specify the virtual scale and centroid of the node. The active "Leaf" nodes $\nu \in V^L(t)$ have the state variables d_ν which specify the index of the segments for Parselets. In summary, we specify the configuration of the graph by the states $z = \{(t_\nu, g_\nu) : \nu \in V^O(t)\} \bigcup \{g_\nu : \nu \in V^A(t)\} \bigcup \{d_\nu : \nu \in V^L(t)\}$ where the active nodes $V(t)$ are determined from the $\{t_\nu : \nu \in V^O(t)\}$. We then let $z_{\text{kids}(\nu)} = \{z_\mu : \mu \in \text{kids}(\nu)\}$ denote the states of all the child nodes of an "And" node $\nu \in V^A$ and let z_{t_ν} denote the state of the selected child node of an "Or" node $\nu \in V^O$.

Invisibility Modeling: Some Parselets, such as bags and scarfs, have high probability to be absent or occluded, namely invisible. In other words, these "Leaf" nodes should be with the visibility property. We explicitly model these notes by using a special structure, denoted as virtual "Leaf"

node. More specifically, we introduce an auxiliary "Invisible" type of nodes which have no appearance representation. Then the virtual "Leaf" node is represented as a structure consisting of an "Or" node, an ordinary "Leaf" node and an "Invisible" node, as shown in Fig. 3. The activated nodes in the virtual "Leaf" node structure thus explicitly suggest whether the corresponding "Leaf" node (Parselet) is visible or not. For standard "Leaf" node μ, the corresponding score is $w_\mu^L \cdot \Phi^L(P, z_\mu)$, where $\Phi^L(P, z_\mu)$ is the feature vector extracted from the segment d_μ as described in Section 3.3. For the virtual "Leaf" node with "Or" node ν, "Leaf" node μ and "Invisible" node ρ, the score is $w_\mu^L \cdot \Phi^L(P, z_\mu) + w_{\nu,\mu}^O$ or $w_{\nu,\rho}^O$ depending on the visibility of the corresponding Parselet. $w_{\nu,\mu}^O$ and $w_{\nu,\rho}^O$ are the learned weights for the visibility property, which are embedded in the "Or" node of the virtual "Leaf" node. It is worth noting that the state of the "Invisible" node fully depends on its weight in the "Or" node and its own score is always 0.

We can now write the full score associated with a state variable z:

$$S(P, z) = \sum_{\mu \in V^L(t)} w_\mu^L \cdot \Phi^L(P, z_\mu) + \sum_{\mu \in V^O(t)} w_{\mu, t_\mu}^O + \sum_{\mu \in V^A(t)} w_\mu^A \cdot \Phi^A(z_\mu, z_{\text{kids}(\mu)}). \quad (1)$$

The first term in Eqn. (1) is an appearance model that computes the local score of assigning the segment d_μ as Parselet μ. The last two terms are independent of the data and can be considered as priors of occurrence and the spatial geometry. Based on the graph structure, we can further decompose the last term of Eqn. (1) as follows:

$$S(P, z) = \sum_{\mu \in V^L(t)} w_\mu^L \cdot \Phi^L(P, z_\mu) + \sum_{\mu \in V^O(t)} w_{\mu, t_\mu}^O + \sum_{\mu \in V^A(t)} \sum_{\nu \in \text{kids}(\mu)} w_{\mu,\nu}^A \cdot \psi(d_\mu, d_\nu). \quad (2)$$

$\psi(d_\mu, d_\nu) = [dx \;\; dx^2 \;\; dy \;\; dy^2 \;\; ds]^T$ measures the geometric difference between part μ and ν, where $dx = (x_\nu - x_\mu)/\sqrt{s_\nu \cdot s_\mu}$, $dy = (y_\nu - y_\mu)/\sqrt{s_\nu \cdot s_\mu}$ and $ds = s_\nu / s_\mu$ are the relative location and scale of part ν with respect to μ.

Compared with the most prevalent hierarchical modeling approaches [32, 12], the proposed model has the following distinctive characteristics:

- We use Parselets as the basic elements for our parsing model. The parsing problem is now transferred as searching the best configuration of the hierarchical model. Once the maximization is obtained, we can directly get the accurate pixel-level segmentation and semantic labels from the corresponding Parselets.
- The "And-Or" graph structure allows both co-occurrence and exclusivity relations between different parts. Unlike previous methods [32, 12], which often use "Or" node to model the multi-view properties of the same part, the "Or" node here plays the role of selecting the best configuration among mixture of subgraphs, which is more flexible.
- We explicitly model the visibility property of the "Leaf" node, which is practical and critical for some Parselets. The introduction of a special node, *i.e.* the Invisible node, brings the flexibility for the real-life situation without adding extra model complexity.

4.2. Inference

Inference corresponds to maximizing $S(P, z)$ from Eqn. (2) over z. As graph G = (V, E) is a tree, inference can be done efficiently with dynamic programming. More specifically, we can simply iterate over all subparts starting from the leaves and moving "upstream" to the root. The message from children to their parent can be computed by the following:

$$\text{score}_\tau^I(z_\tau) = 0, \quad (3)$$

$$\text{score}_\tau^L(z_\tau) = w_\tau^L \cdot \Phi^L(P, z_\tau), \quad (4)$$

$$\text{score}_\nu^O(z_\nu) = \max_{\rho \in \text{kids}(\nu)} [m_\rho(z_\nu)], \quad (5)$$

$$m_\rho(z_\nu) = \max_{z_\rho} [\text{score}_\rho(z_\rho)] + w_{\nu,\rho}^O, \quad (6)$$

$$\text{score}_\mu^A(z_\mu) = \sum_{\rho \in \text{kids}(\mu)} n_\rho(z_\mu), \quad (7)$$

$$n_\rho(z_\mu) = \max_{z_\rho} [\text{score}_\rho(z_\rho) + w_{\mu,\rho}^A \cdot \psi(d_\mu, d_\rho)]. \quad (8)$$

At the bottom level, the scores of "Invisible" nodes and "Leaf" nodes are calculated as in Eqn. (3) and Eqn. (4). "Or" node selects the maximal response from its children for its score as in Eqn. (5) and Eqn. (6). The score of "And" node is calculated by accumulating the scores of its children plus the corresponding deformation as in Eqn. (7) and Eqn. (8). The above equations suggest that we can express the energy function recursively and hence find the optimal z using dynamic programming. In addition, the maximization over z can be partially accelerated by generalized distance transformation, which makes the whole algorithm more efficient [14, 12].

4.3. Learning

Given the labeled examples $\{P_i, z_i\}$, the max-margin framework is arguably preferable to maximum-likelihood estimation as our final goal is discrimination. Note that the scoring function of Eqn. (2) is linear in model parameters $w = (w^L, w^O, w^A)$, and can be written compactly as $S(P, z) = w \cdot \Phi(P, z)$. Thus both appearance and structure parameters can be learned in a unified framework, which is critical for achieving the state-of-the-art performance for many applications [12, 32]. Here, we formulate the structured learning problem in a max-margin framework as in [12]:

Table 2: The best IoU scores for each type of Parselets on the FS and DP datasets.

dataset	hat	hair	s-gls	u-cloth	coat	f-cloth	skirt	pants	belt	l-shoe	r-shoe	face	l-arm	r-arm	l-leg	r-leg	bag	scarf
FS	84.0	78.2	56.6	84.1	null	90.8	91.6	92.8	65.7	72.4	71.9	83.4	79.8	79.8	79.2	79.9	81.8	76.1
DP	83.5	81.0	58.8	88.6	71.9	93.9	89.3	92.5	71.0	73.2	73.8	85.6	93.4	92.9	86.7	86.5	84.8	78.2

$$\min_{w} \|w\|^2 + C\sum_{i}\xi_i$$
$$\text{s.t.} \quad w\cdot(\Phi(P_i,z_i)-\Phi(P_i,z)) \geq \Delta(z_i,z)-\xi_i, \forall z; \tag{9}$$

where $\Delta(z_i, z_j)$ is a loss function which penalizes incorrect estimate of z. This loss function gives partial credit to states which differ from the ground truth slightly. The loss function is defined as follows:

$$\Delta(z_i,z_j) = \sum_{\nu\in V^L(t_i)\bigcup V^L(t_j)} \delta(z_i^\nu, z_j^\nu), \tag{10}$$

where $\delta(z_i^\nu, z_j^\nu) = 1$, if $\nu \notin V^L(t_i)\bigcap V^L(t_j)$ or $\text{sim}(d_i^\nu, d_j^\nu) \leq \sigma$. $\text{sim}(\cdot,\cdot)$ is the intersection over union ratio of two segments d_i^ν and d_j^ν, and σ is the threshold, which is set as 0.8 in the experiments. This loss function penalizes both configurations with "wrong" topology and leaf nodes with wrong segments. The optimization problem Eqn. (9) is known as a structural SVM, which can be efficiently solved by the cutting plane solver of SVMStruct [18] and the stochastic gradient descent solver in [12].

5. Experiments

5.1. Experimental Settings

Dataset: Our experiments are conducted on two datasets. The first one is the Fashionista (FS) dataset [31], which has 685 annotated samples with 56 different clothing labels. This dataset is originally designed for fine-grained clothing parsing. To adapt this dataset for our human parsing, we merge their labels according to our Parselet definition. As there is no direct link between their annotation and our "coat" Parselet, we ignore the "coat" Parselet and merge all upper body clothing into the "upper clothes" Parselet. The second dataset, called Daily Photos (DP), contains 2500 high resolution images, which are crawled following the same strategy as the FS dataset [31]. In order to obtain quantitative evaluation results, we thoroughly annotate the semantic labels at pixel-level. Compared with FS, the DP dataset contains much more images and has consistent labels with Parselet definition for human parsing.

Evaluation Criterion: The parsing result is evaluated based on two complementary metrics. The first one is Average Pixel Accuracy (APA) [31], which is defined as the proportion of correctly labeled pixels in the whole image. This metric mainly measures the overall performance. The second metric is Intersection over Union (IoU) [11], which is widely used in evaluating segmentation and suitable for measuring the performance of each Parselet separately. We also devise two variants of IoU for Parselets to make Parselets comparable with objects. The first one is the "Merging IoU" (mIoU) which merges the hypothesis for each Parselet into an object hypothesis to obtain the object level IoU. The second one is the "Weighted IoU" (wIoU) which is calculated by accumulating each Parselet's IoU score weighted by the ratio of its pixels occupying the whole object. Note that generally mIoU is higher than wIoU.

Table 3: Comparison of Parselets versus objects in terms of the best IoU score on FS and DP datasets.

	dataset	CPMC [27]	SLIC [1]	UCM [3]	Combined
Obj IoU	FS	0.830	0.559	0.430	0.831
Par mIoU	FS	**0.895**	**0.725**	**0.604**	**0.917**
Par wIoU	FS	0.844	0.621	0.546	0.860
Obj IoU	DP	0.815	0.534	0.443	0.816
Par mIoU	DP	**0.896**	**0.722**	**0.638**	**0.928**
Par wIoU	DP	0.831	0.614	0.608	0.862

Implementation Details: We extract dense SIFT [23], HOG [9] and color moment as low-level features for Parselets. The size of Gaussian Mixture Model in FK is set to 128. The training:testing ratio is 2:1 for both datasets. The penalty parameter C is determined by 3-fold cross validation in the training set.

5.2. Hypotheses Comparison: Parselets vs. Objects

We first validate the assumption that segmentation can provide better hypotheses for Parselets than for objects with heterogeneous appearance (*e.g*. human) by comparing the best IoU scores of Parselets and objects. The best IoU score for a segmentation method is defined as the maximal IoU score between the segments produced by that method and the ground truth segments. The same hypothesis segments, which are generated through the methods introduced in Section 3.2, are used for both Parselets and objects. We calculate the best IoU of Parselets and objects for different method on two datasets. The comparison results are displayed in Table 3, from which it can be observed that the best IoU of Parselets is much higher than that of objects. This trend is consistent among different algorithms and datasets, which makes the usage of segments as Parselet hypotheses more convincing. In addition, combining all three complementary algorithms leads to the best performance and we use this setting thereafter. The detailed best IoU for each type of Parselets based on combined hypotheses are shown in Table 2 .

5.3. Evaluation for Human Parsing

Human Parsing: We now compare our proposed framework with the work of Yamaguchi *et al*. [31] for human parsing. This baseline works by first estimating the human pose and then labeling the super-pixel based on the pose estimation results. We use the public available implementation of version 0.2 and carefully tune the parameter according to [31]. The baseline method achieves 83% for FS dataset and 82% for DP dataset in terms of APA, which

Figure 4: Comparison of parsing results. Original images, our results and baseline's results [31] are shown sequentially.

Figure 5: More exemplar results from our parsing framework.

are inferior to 86% and 87% of our framework. Though APA is good at measuring the overall performance of human parsing, it fails to distinguish the performance of separate Parselets and has bias towards background. More specifically, naively assigning all segments as background results in a reasonably good APA of 78% for DP and 77% for FS. Therefore, we further employ the more discriminative IoU criterion for comparison. The detailed comparison results on all types of Parselets are reported in Table 4. It can be seen that our method performs much better than the baseline method, especially for the Parselet level results. This mainly verifies the stability of our algorithm. Unlike our method, the baseline method does not model the exclusive relation of different labels, which leads to unstable results as shown in Fig. 4. Note that their method can achieve good performance with the prior information specifying what type of Parselets appears in the image. However, such information is usually difficult to obtain for real-world applications. In addition, it can be observed that the results from our model are more robust to uncommon poses and absent/occluded parts. The baseline method estimates the human pose and labels the region separately. This non-unified nature omits the strong correlation of appearance and structure for human. On the contrast, by employing the low-level visual cues and high-level structure information in a unified framework with explicit invisibility modeling, our model is much more robust to these difficult examples. More exemplar results from our framework are shown in Fig. 5.

Parsing as Segmentation: As human parsing results in pixel-level segment labeling, our framework implicitly provides human segmentation results. We thus further compare the segmentation results between our human parsing method and the state-of-the-art image segmentation method [4], to demonstrate the effectiveness of our framework. The baseline method [4] employs the bottom-up segments as the object hypotheses and only achieves the IoU score of 73% for FS dataset and 70% for DP dataset, which

Figure 6: Comparison of human segmentation results. (a)-(d) are input images, our human parsing results, segmentation results by merging (b) and results from the segmentation method [4], respectively

is much lower than the result of Merging IoU of 83.1% and 84.6% as shown in Table 4. Some exemplar results are shown in Fig. 6, from which we can observe obvious defects for the baseline segmentation results in column (d). Such defects are avoidless for the baseline method as a single segment from the bottom-up segmentation can hardly cover the whole body tightly. On the contrary, our framework can employ the top-down knowledge and assemble several homogeneous segments into an object, which leads to much more accurate segmentation.

5.4. Human Parsing for High Level Applications

Parselets provide a middle-level representation and well bridge the gap between the low-level segments and the high-level concepts. Hence, our Parselet based parsing framework can serve as the basis for many high-level applications. Here, we build a prototype system to retrieve visually similar person as a representative. More specifically, given a query image, we first filter images in the database based on the Parselet types. For each pair of corresponding Parselets, the similarity is calculated based on the Euclidean dis-

Table 4: Comparison of human parsing IoU scores on FS and DP datasets.

	dataset	hat	hair	s-gls	u-cloth	coat	f-cloth	skirt	pants	belt	l-shoe	r-shoe	face	l-arm	r-arm	l-leg	r-leg	bag	scarf	wIoU	mIoU
Baseline [31]	FS	2.5	47.2	0.8	36.4	null	23.2	21.6	19.1	8.9	27.6	25.2	59.3	33.0	30.5	32.6	24.1	9.5	0.9	29.9	77.6
DMPM	FS	5.6	67.9	2.8	56.3	null	56.6	55.3	40.0	18.2	58.6	53.4	72.4	52.7	45.4	48.8	41.6	20.6	1.2	51.7	83.1
Baseline [31]	DP	1.3	43.5	0.6	21.3	19.5	21.8	12.2	28.7	4.8	25.6	21.7	52.6	32.4	28.3	23.5	18.4	8.5	1.2	24.6	76.6
DMPM	DP	28.9	74.8	9.6	42.5	39.4	61.0	50.3	66.3	16.6	57.0	51.8	78.1	62.7	59.3	52.6	35.5	12.7	9.3	53.0	84.6

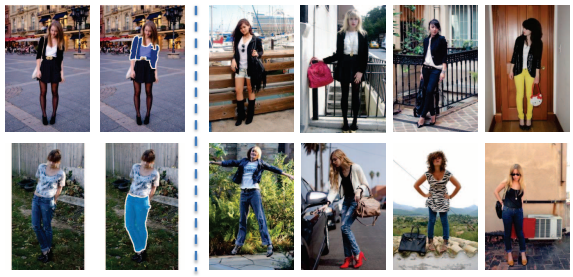

Figure 7: Top retrieval results from our visually similar person retrieval system. The retrieval results (right columns) are visually similar to the query human for the highlighted Parselets (the second column) independent of pose and uninterested regions.

tance of the extracted features. Then the similarity between images is defined as the sum of Parselet-level similarities weighted by the faction of their pixels occupying the object. Such a system can be extended for clothing retrieval, person identification and many other human centric analysis. Fig. 7 shows some top retrieval results for Parselets such as upper clothes + coat and pants, respectively. It can be observed that the visually similar persons are successfully retrieved independent of pose and uninterested regions. Here, we do not pursue this further for the space limitation.

6. Conclusions and Future Work

In this paper, we proposed an effective framework for human parsing. By reconsidering the human parsing problem, we utilized the novel Parselets as the basic elements. A unique Deformable Mixture Parsing Model (DMPM) was built to jointly learn and infer the best configuration for both appearance and structure effectively. Extensive experimental results clearly demonstrated the effectiveness of the proposed framework. In the future, we plan to further explore how to adequately utilize the top-down information and integrate the fine-grained attribute into our framework.

7. Acknowledgment

This research is supported by the Singapore National Research Foundation under its International Research Centre @Singapore Funding Initiative and administered by the IDM Programme Office.

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *TPAMI*, 2012.

[2] P. Arbeláez, B. Hariharan, C. Gu, S. Gupta, L. Bourdev, and J. Malik. Semantic segmentation using regions and parts. In *CVPR*, 2012.

[3] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *TPAMI*, 2011.

[4] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu. Semantic segmentation with second-order pooling. In *ECCV*. 2012.

[5] J. Carreira and C. Sminchisescu. Cpmc: Automatic object segmentation using constrained parametric min-cuts. *TPAMI*, 2012.

[6] K. Chatfield, V. Lempitsky, and A. Vedaldi. The devil is in the details: an evaluation of recent feature encoding methods. In *BMVC*, 2011.

[7] H. Chen, A. Gallagher, and B. Girod. Describing clothing by semantic attributes. In *ECCV*. 2012.

[8] S. chun Zhu and D. Mumford. A stochastic grammar of images. In *Foundations and Trends in Computer Graphics and Vision*, 2006.

[9] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[10] I. Endres and D. Hoiem. Category independent object proposals. *ECCV*, 2010.

[11] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results.

[12] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object Detection with Discriminatively Trained Part-Based Models. *TPAMI*, 2010.

[13] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 2005.

[14] P. F. Felzenszwalb and D. P. Huttenlocher. Distance transforms of sampled functions. *Theory of Computing*, 2012.

[15] J. S. Florent Perronnin and T. Mensink. Improving the Fisher Kernel for Large-Scale Image Classification. In *ECCV*, 2010.

[16] A. C. Gallagher and T. Chen. Clothing cosegmentation for recognizing people. In *CVPR*, 2008.

[17] C. Gu, J. J. Lim, P. Arbeláez, and J. Malik. Recognition using regions. In *CVPR*, 2009.

[18] T. Joachims, T. Finley, and C.-N. J. Yu. Cutting-plane training of structural svms. *Machine Learning*, 2009.

[19] P. Kohli, J. Rihan, M. Bray, and P. H. Torr. Simultaneous segmentation and pose estimation of humans using dynamic graph cuts. *IJCV*, 2008.

[20] S. Liu, J. Feng, Z. Song, T. Zhang, H. Lu, C. Xu, and S. Yan. Hi, magic closet, tell me what to wear! In *ACM MM*, 2012.

[21] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, and S. Yan. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *CVPR*, 2012.

[22] X. Liu, L. Lin, S.-C. Zhu, and H. Jin. Trajectory parsing by cluster sampling in spatio-temporal graph. In *CVPR*, 2009.

[23] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.

[24] B. C. Russell, W. T. Freeman, A. A. Efros, J. Sivic, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *CVPR*, 2006.

[25] M. Sun and S. Savarese. Articulated part-based model for joint object detection and pose estimation. In *ICCV*, 2011.

[26] P. H. Torr and A. Zisserman. Human pose estimation using a joint pixel-wise and part-wise formulation. 2013.

[27] K. E. A. van de Sande, J. R. R. Uijlings, T. Gevers, and A. W. M. Smeulders. Segmentation as selective search for object recognition. In *ICCV*, 2011.

[28] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman. Multiple kernels for object detection. In *ICCV*, 2009.

[29] J. Wang, J. Yang, K. Yu, F. Lv, and T. Huang. Locality-constrained linear coding for image classification. In *CVPR*, 2010.

[30] X. Wang and T. Zhang. Clothes search in consumer photos via color matching and attribute learning. In *ACM MM*, 2011.

[31] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg. Parsing clothing in fashion photographs. In *CVPR*, 2012.

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

[33] L. Zhu, Y. Chen, Y. Lu, C. Lin, and A. Yuille. Max margin and/or graph learning for parsing the human body. In *CVPR*, 2008.

[34] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, 2012.