Towards Unified Human Parsing and Pose Estimation

Jian Dong[1], Qiang Chen[1], Xiaohui Shen[2], Jianchao Yang[2], Shuicheng Yan[1]
[1] Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[2] Adobe Research, San Jose, CA, USA
{a0068947, chenqiang, eleyans}@nus.edu.sg,{xshen, jiayang}@adobe.com

Abstract

We study the problem of human body configuration analysis, more specifically, human parsing and human pose estimation. These two tasks, i.e. *identifying the semantic regions and body joints respectively over the human body image, are intrinsically highly correlated. However, previous works generally solve these two problems separately or iteratively. In this work, we propose a unified framework for simultaneous human parsing and pose estimation based on semantic parts. By utilizing Parselets and Mixture of Joint-Group Templates as the representations for these semantic parts, we seamlessly formulate the human parsing and pose estimation problem jointly within a unified framework via a tailored And-Or graph. A novel Grid Layout Feature is then designed to effectively capture the spatial co-occurrence/occlusion information between/within the Parselets and MJGTs. Thus the mutually complementary nature of these two tasks can be harnessed to boost the performance of each other. The resultant unified model can be solved using the structure learning framework in a principled way. Comprehensive evaluations on two benchmark datasets for both tasks demonstrate the effectiveness of the proposed framework when compared with the state-of-the-art methods.*

1. Introduction

Human parsing (partitioning the human body into semantic regions) and pose estimation (predicting the joint positions) are two main topics of human body configuration analysis. They have drawn much attention in the recent years and serve as the basis for many high-level applications [1, 24, 5]. Despite their different focuses, these two tasks are highly correlated and complementary. On one hand, most works on pose estimation usually divide the body into parts based on joint structure [24]. However, such joint-based decomposition ignores the influence of clothes, which may significantly change the appearance/shape of a person. For example, it is hard for joint-based models to accurately locate the knee positions of a person wearing long dress as shown in Figure 1. In this case, the human parsing results can provide valuable context information for locating the missing joints. On the other hand, human parsing can be formulated as inference in a conditional random field (CRF) [17, 5]. However, without top-down information such as human pose, it is often intractable for CRF to distinguish ambiguous regions (*e.g.*, the left shoe v.s. the right shoe) using local cues as illustrated in Figure 1. Despite the strong connection of these two tasks, the intrinsic consistency between them has not been fully explored, which hinders the two tasks from benefiting each other. Only very recently, some works [23, 18] began to link these two tasks with the strategy of performing parsing and pose estimation sequentially or iteratively. While effective, this paradigm is suboptimal, as errors in one task will propagate to the other.

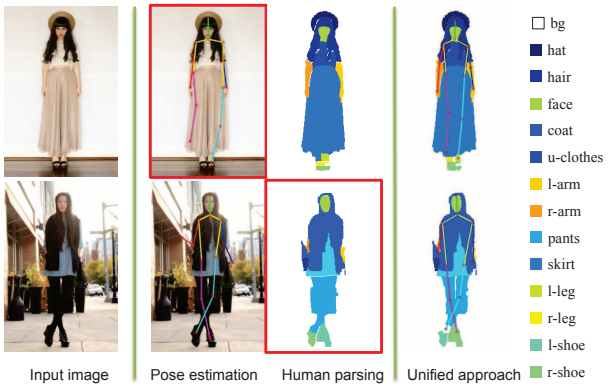

Figure 1. Motivations for unified human parsing and pose estimation. The images in top row show the scenario where pose estimation [24] fails due to joints occluded by clothing (*e.g.*, knee covered by dress) while the human parsing works fine. The images in bottom row show the scenario where human parsing [5] is not accurate when body regions are crossed together (*e.g.*, the intersection of the legs). Thus, the human parsing and pose estimation may benefit each other, and more satisfactory results (the right column) can be achieved for both tasks using our unified framework.

In this work, we aim to seamlessly integrate human parsing and pose estimation under a unified framework. To this end, we first unify the basic elements for both tasks by proposing the concept of "semantic part". A semantic part is either a region with contour (*e.g.*, hair, face and skirt) re-

lated to the parsing task, or a joint group (*e.g.*, right arm with wrist, elbow and shoulder joints) serving for pose estimation. For the representation of semantic regions, we adopt the recently proposed Parselets [5]. Parselets are defined as a group of segments which can be generally obtained by low-level over-segmentation algorithms and bear strong semantic meaning. Unlike the raw pixels used by traditional parsing methods [17], which are not directly compatible with the template based representation for pose estimation, Parselets allow us to easily convert the human parsing task into the structure learning problem as in pose estimation. For pose estimation, we employ joint groups instead of single joints as basic elements since joints themselves are too fine-grained for effective interaction with Parselets. We then represent each joint group as one Mixture of Joint-Group Templates (MJGT), which can be regarded as a mixture of pictorial structure models defined on the joints and their interpolated keypoints. This design ensures that the semantic region and joint group representation of the semantic parts are at the similar level and thus can be seamlessly connected together.

By utilizing Parselets and MJGTs as the semantic parts representation, we propose a Hybrid Parsing Model (HPM) for simultaneous human parsing and pose estimation. The HPM is a tailored "And-Or" graph [25] built upon these semantic parts, which encodes the hierarchical and reconfigurable composition of parts as well as the geometric and compatibility constraints between parts. Furthermore, we design a novel grid-based pairwise feature, called Grid Layout Feature (GLF), to capture the spatial co-occurrence/occlusion information between/within the Parselets and MJGTs. The mutually complementary nature of these two tasks can thus be harnessed to boost the performance of each other. Joint learning and inference of best configuration for both human parsing and pose related parameters guarantee the overall performance. The major contributions of this work include:

- We build a novel Hybrid Parsing Model for unified human parsing and pose estimation. Unlike previous works, we seamlessly integrate two tasks under a unified framework, which allows joint learning of human parsing and pose estimation related parameters to guarantee the overall performance.
- We propose a novel Grid Layout Feature (GLF) to effectively model the geometry relation between semantic parts in a unified way. The GLF not only models the deformation as in the traditional framework but also captures the spatial co-occurrence/occlusion information of those semantic parts.
- HPM achieves the sate-of-the-art for both human parsing and pose estimation on two public datasets, which verifies the effectiveness of joint human parsing and pose estimation, and thus well demonstrates the mutually complementary nature of both tasks.

2. Related Work

2.1. Human Pose Estimation

Human pose estimation has drawn much research attention during the past few years [1]. Due to the large variance in viewpoint and body pose, most recent works utilize mixture of models at a certain level [24, 14]. Similar to the influential deformable part models [6], some methods [14] treat the entire body as a mixture of templates. However, since the number of plausible human poses is exponentially large, the number of parameters that need to be estimated is prohibitive without a large dataset or a part sharing mechanism. Another approach [24] focuses on directly modeling modes only at the part level. Although this approach has combinatorial model richness, it usually lacks the ability to reason about large pose structures at a time. To strike a balance between model richness and complexity, many works begin to investigate the mixtures at the middle level in hierarchical models, which have achieved promising performance [4, 15, 16, 13]. As we aim to perform simultaneous human parsing and pose estimation, we tailor the above techniques for the proposed HPM by utilizing the mixture of joint-group templates as basic representation for body joints.

2.2. Human Parsing

There exist several inconsistent definitions for human parsing in literature. Some works [19, 21, 22] treat human parsing as a synonym of human pose estimation. In this paper, we follow the convention of scene parsing [12, 17] and define human parsing as partitioning the human body into semantic regions. Though human parsing plays an important role in many human-centric applications [3], it has not been fully studied. Yamaguchi *et al*. [23] performed human pose estimation and attribute labeling sequentially for clothing parsing. However, such sequential approaches may fail to capture the correlations between human appearance and structure, leading to unsatisfactory results. Dong *et al*. proposed the concept of Parselets for direct human parsing under the structure learning framework [5]. Recently, Torr and Zisserman proposed an approach for joint human pose estimation and body part labeling under the CRF framework [18], which can be regarded as a continuation of the theme of combining segmentation and human pose estimation [11, 8, 20]. Due to the complexity of this model, the optimization cannot be carried out directly and thus is conducted by first generating a pool of pose candidates and then determining the best pixel labeling within this restricted set of candidates. Our method differs from previous approaches as we aim to solve human parsing and pose estimation simultaneously in a unified framework, which allows joint learning of all parameters to guarantee the overall performance.

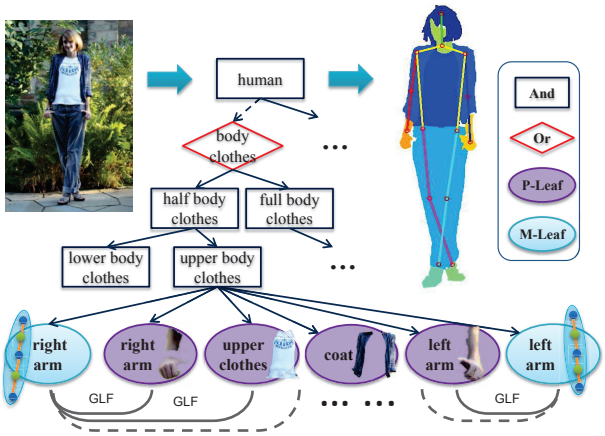

Figure 2. Illustration of the proposed Hybrid Parsing Model. The hierarchical and reconfigurable composition of semantic parts are encoded under the And-Or graph framework. The "P-Leaf" nodes encode the region information for parsing while the "M-Leaf" nodes capture the joint information for pose estimation. The pairwise connection between/within "P-Leaf"s and "M-Leaf" is modelled through Grid Layout Feature (GLF). HPM can simultaneously perform parsing and pose estimation effectively.

3. Unified Human Parsing and Pose Estimation

In this section, we introduce the framework of the proposed Hybrid Parsing Model and detail the key components.

3.1. Unified Framework

We first give some probabilistic motivations for our approach. Human parsing can be formally formulated as a pixel labeling problem. Given an image I, the parsing system should assign the label mask $L \equiv \{l_i\}$ to each pixel i, such as face or dress, from a pre-defined label set. Human pose estimation aims to predict the joint positions $X \equiv \{x_j\}$, which is a set of image coordinates x_j for body joints j. As human parsing and pose estimation are intuitively strongly correlated, ideally one would like to perform MAP estimation over joint distribution $p(X, L|I)$. However, previous works either estimate $p(X|I)$ and $p(L|I)$ separately [24] or estimate $p(X|I)$ and $p(L|X, I)$ sequentially [23]. The first case obviously ignores the strong correlation between joint positions X and parsing label mask L. The second approach may also be suboptimal, as errors in estimating X will propagate to L.

To overcome the limitations of previous approaches, we propose the Hybrid Parsing Model (HPM) for unified human parsing and pose estimation by directly estimating MAP over $P(X, L|I)$. The proposed HPM uses Parselets and Mixture of Joint-Group Templates (MJGT) as the semantic part representation (which will be detailed in Section 3.2) under the "And-Or" graph framework. This instantiated "And-Or" graph encodes the hierarchical and reconfigurable composition of semantic parts as well as the geometric and compatibility constraints between them. Formally, an HPM 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 according to the parent-child relation and "kids(ν)" denotes the children of node ν. Unlike the traditional And-Or graph, we define four basic types of nodes, namely, "And","Or", "P-Leaf" and "M-Leaf" nodes as depicted in Figure 2. Each "P-Leaf" node corresponds to one type of Parselets encoding pixel-wise labeling information, while each "M-Leaf" node represents one type of MJGTs for joint localization. The graph topology is specified by the switch variable t at "Or" nodes, which indicates the set of active nodes $V(t)$. $V^O(t), V^A(t), V^{L_P}(t)$ and $V^{L_M}(t)$ represent the active "Or", "And", "P-Leaf" and "M-Leaf" nodes, respectively. Starting from the top level, an active "Or" node $\nu \in V^O(t)$ selects a child $t_\nu \in \text{kids}(\nu)$. P represents the set of Parselet hypotheses in an image and z denotes the state variables for the whole graph. We then define $z_{\text{kids}(\nu)} = \{z_\mu : \mu \in \text{kids}(\nu)\}$ as 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 $\nu \in V^O$.

Based on the above representation, the conditional distribution on the state variable z and the data can then be formulated as the following energy function (Gibbs distribution):

$$E(I, z) = \sum_{\mu \in V^O(t)} E^O(z_\mu) + \sum_{\mu \in V^A(t)} E^A(z_\mu, z_{\text{kids}(\mu)}) + \sum_{\mu \in V^{L_P}(t)} E^{L_P}(I, z_\mu) + \lambda \sum_{\mu \in V^{L_M}(t)} E^{L_M}(I, z_\mu). \quad (1)$$

The "P-Leaf" component $E^{L_P}(.)$ links the model with the pixel-wise semantic labeling, while the 'M-Leaf" component $E^{L_M}(.)$ models the contribution of keypoints. The "And" component $E^A(.)$ captures the geometry interaction among nodes. The final "Or" component $E^O(.)$ encodes the prior distribution/compatibility of different parts. It is worth noting that there exists pairwise connection at the bottom level in our "And-Or" graph as shown in Figure 2. This ensures that more sophisticated pairwise modeling can be utilized to model the connection between/within "P-Leaf" and "M-Leaf" nodes. We approach this by designing the Grid Layout Feature (GLF). The detailed introduction of each component and GLF are given below.

3.2. Representation for Semantic Parts

In this subsection, we give details of the representation for the semantic parts. More specifically, we utilize Parselets and Mixture of Joint-Group Templates (MJGT) as the representation for regions and joint groups.

3.2.1 Region Representation with Parselets

Traditional CRF-based approaches for human parsing [8, 13] are inconsistent with structure learning approaches widely used for pose estimation. To overcome this difficulty, we employ the recently proposed Parselets [5] as

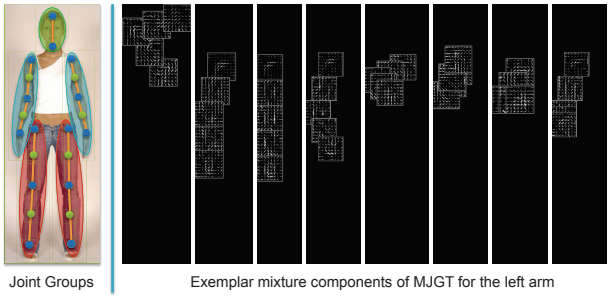

Figure 3. The left image shows our joint-group definition (marked as ellipses). Each group consist of several joints (marked as blue dots) and their interpolated points (marked as green dots). We represent each group as one Mixture of Joint-Group Templates (MJGT). Some exemplar mixture components of the MJGT for the right arm are shown on the right side.

building blocks for human parsing. In a nutshell, Parselets are a group of semantic image segments with the following characteristics: (1) can generally be obtained by low-level over-segmentation algorithms; and (2) bear strong and consistent semantic meanings. With a pool of Parselets, we can convert the human parsing task into the structure learning problem, which can thus be unified with pose estimation under the "And-Or" graph framework.

As Parselet categorization can be viewed as a region classification problem, we follow [5] by utilizing the state-of-the-art classification pipelines [9, 2] for feature extraction. The parsing node score can then be calculated by

$$E^{L_P}(I, z_\mu) = w_\mu^{L_P} \cdot \Phi^{L_P}(I, z_\mu),$$

where $\Phi^{L_P}(.)$ is the concatenation of appearance features for the corresponding Parselet of node μ.

3.2.2 Mixture of Joint-Group Templates

The HoG template based structure learning approaches have shown to be effective for human pose estimation [24, 13, 14]. Most of these approaches treat keypoints (joints) as basic elements. However, joints are too fine-grained for effective interaction with Parselets. Since joints and Parselets have no apparent one-to-one correspondence (*e.g.*, knee joints may be visible or be covered by pants, dress or skirt), direct interaction between all joints (plus additional interpolated keypoints) and the Parselets is almost intractable. Hence, we divide the common 14 joints for pose estimation [24, 13] into 5 groups (*i.e.* left/right arm, left/right leg and head), as shown in Figure 3. Each joint group is modeled by one Mixture of Joint-Group Templates (MJGT). MJGT can be regarded as a mixture of pictorial structure models [7, 24] defined on the joints and interpolated keypoints (blue points and green points in Figure 3). We choose MJGT defined on joint groups as the building block for modeling human pose mainly for three reasons: (1) there are much fewer joint groups than keypoints, which allows more complicated interaction with Parselets; (2) with the reduced complexity in each component brought by the mixture models, we can employ the linear HoG template + spring deformation representation for pictorial structure modeling [24, 14] to ensure the effectiveness of pose estimation; and (3) each component of an MJGT can easily embed mid-level status information (*e.g.*, the average mask).

In practice, we set the number of mixtures as 32/16/16 for MJGT to handle the arms/legs/head group variance respectively. The training data are split into different components based on the clusters of the joint configurations. In addition, an average mask is attached to each component of MJGTs to unify the interaction between Parselet and MJGT, which will be discussed in Section 3.3. The state of the instantiated mask for a component of an MJGT is fully specified by the scale and the position of the root node.

For an MJGT model μ, we can now write the score function associated with a configuration of component m and positions c as in [24, 14]:

$$S_\mu(I, c, m) = b_m + \sum_{i \in \mathcal{V}_\mu} w_i^{\mu,m} \cdot \mathbf{f}_i(I, c_i) + \sum_{(i,j) \in \mathcal{E}_\mu} w_{(i,j)}^{\mu,m} \cdot \mathbf{f}_{i,j}(c_i, c_j),$$

where $\mathcal{V}_\mu$ and $\mathcal{E}_\mu$ are the node and edge set, respectively. $\mathbf{f}_i(I, c_i)$ is the HoG feature extracted from pixel location c_i in image I and $\mathbf{f}_{i,j}(c_i, c_j)$ is the relative location $([dx, dy, dx^2, dy^2])$ of joint i with respect to j. Each M-Leaf node can be seen as the wrapper of an MJGT model. Hence the score of M-Leaf is equal to that of the corresponding MJGT model. As the state variable z_μ contains the component and position information for M-Leaf node μ, the final score can be written more compactly as follows:

$$E^{L_M}(I, z_\mu) = w_\mu^{L_M} \cdot \Phi^{L_M}(I, z_\mu),$$

where $\Phi^{L_M}(.)$ is the concatenation of the HoG features and the relative geometric features for all the components within the joint group.

3.3. Pairwise Geometry Modeling

According to our "And-Or" graph construction, there exist three types of pairwise geometry relations in the HPM: (1) Parselet-Parselet, (2) Parselet-MJGT, and (3) parent-child in "And" nodes. Articulated geometry relation, such as relative displacement and scale, is widely used in the pictorial structure models to capture the pairwise connection. We follow this tradition to model the parent-child interaction (3) as in [24]. However, the pairwise relation of (1) and (2) is much more complex. For example, as shown in Figure 4, the "coat" Parselet has been split into two parts and its relation with the "upper clothes" Parselet can hardly be accurately modeled by using only their relative center positions and scales. Furthermore, as Parselets and MJGTs essentially model the same person by different representations, a more precise constraint than the articulated geometry should be employed to ensure their consistency.

To overcome the above difficulties, we propose a Grid Layout Feature (GLF) to model the pairwise geometry re-

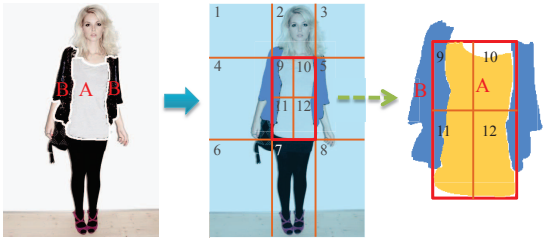

Figure 4. Grid Layout Feature (GLF): GLF measures the pixel spatial distribution relation of two masks. To calculate GLF of mask B with respect to mask A, the image is first divided into 12 spatial bins based on the tight bounding box of A as shown in (b), which includes 8 surrounding and 4 central bins. GLF consists of two parts: (1) the ratio of pixels of mask B falling in the 12 bins , and (2) the ratio of pixels of the interaction of mask A and B falling in the 4 central bins as shown in (c).

lation between two nodes. More specially, as a region mask can be derived from each Parselet or MJGT (the average mask is utilized for MJGTs), the relation between two nodes can be measured by the pixel spatial distribution relation of their corresponding masks. As illustrated in Figure 4, to measure the GLF of mask A with respect to mask B, we first calculate the tight bounding box of A and then divide the whole image into 12 spatial bins, denoted by $R_i, i = 1, \cdots, 12$. The 12 spatial bins consist of 8 cells outside of the bounding box and 4 central bins inside it. We then count the pixels of mask B falling in each bin ($|B \cap R_i|$). Besides the spatial relation, we also model the level of overlap between mask A and B, which has two main functions, *i.e.* (1) to avoid the overlap between Parselets and (2) to encourage the overlap between corresponding Parselets and MJGTs. This is achieved by further counting pixels of the insertion region between A and B in the four central bins ($|A \cap B \cap R_i|$) as shown in Figure 4 (c). The resultant 16 dimension feature is normalized by the total pixel number of mask B ($|B|$). By swapping mask A and mask B, we can get another complementary feature centered at the mask B, which is then concatenated with the original one to form the final 32 dimension sparse vector. Formally, we define the Grid Layout Feature as follows:

$$PG(A,B) = \left[\begin{array}{l} \frac{|B \cap R_i|}{|B|}, i = 1, \cdots, 12; \\ \frac{|A \cap B \cap R_i|}{|B|}, i = 9, \cdots, 12 \end{array} \right],$$
$$\psi_G(A,B) = [PG(A,B); PG(B,A)],$$

where $\psi_G(A,B)$ is the GLF between mask A and B. With GLF, the interaction between Parseles, such as "coat" and "upper clothes", can be effectively captured. Furthermore, as each mixture component of an MJGT is attached with an average mask, interaction (1) and (2) can be easily unified with the help of GLF.

We can then write out the score of the "And" node, whose child nodes consist of multiple Parselets/MJGTs, as follows:

$$E^A(z_\mu, z_{\text{kids}(\mu)}) = \sum_{\nu \in \text{kids}(\mu)} w^A_{\mu,\nu} \cdot \psi(\mu,\nu) + \sum_{\omega,\upsilon \in \text{kids}(\mu),(\omega,\upsilon)\in E} w^A_{\omega,\upsilon} \cdot \psi_G(\omega,\upsilon),$$

where $\psi_G(\omega,\upsilon)$ is the GLF feature between Parselet/MJGT ω and υ. $\psi(\mu,\nu) = [dx \;\; dx^2 \;\; dy \;\; dy^2 \;\; ds]^T$ is the articulated geometry feature to measure 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. As the horizontal relations (Parselet-Parselet, Parselet-MJGT) only exist between the "Leaf" nodes under a common "And" node, the GLF term will be removed for those "And" nodes not connected to "Leaf" nodes. By concatenating all geometry interaction features, the score can be written compactly as:

$$E^A(z_\mu, z_{\text{kids}(\mu)}) = w^A_\mu \cdot \Phi^A(z_\mu, z_{\text{kids}(\mu)}).$$

3.4. Summary

Finally, we summarize the proposed HPM model. For a Parselet hypothesis with index i, its scale (the square root of its area) and centroid can be directly calculated. The switch variable t at "Or" nodes indicates the set of active nodes $V(t)$. The active "And", "Or" and "M-Leaf" nodes have the state variables $g_\nu = (s_\nu, c_\nu)$ which specify the (virtual) scale and centroid of the nodes. The active "P-Leaf" nodes $\nu \in V^{L_P}(t)$ have the state variables d_ν which specify the index of the segments for Parselets, while the active "M-Leaf" nodes $\nu \in V^{L_M}(t)$ have the state variables d_ν which specify the active component index of the MJGTs. In summary, we specify the configuration of the graph by the states $z = \{(t_\nu, g_\nu) : \nu \in V^O(t)\} \bigcup \{g_\nu : \nu \in V^A(t)\} \bigcup \{d_\nu : \nu \in V^{L_P}(t)\} \bigcup \{(d_\nu, g_\nu) : \nu \in V^{L_M}(t)\}$. The full score associated with a state variable z can now be written as:

$$\begin{aligned} S(I,z) = &\sum_{\mu \in V^O(t)} w^O_{\mu,t_\mu} + \sum_{\mu \in V^A(t)} w^A_\mu \cdot \Phi^A(z_\mu, z_{\text{kids}(\mu)}) + \\ &\sum_{\mu \in V^{L_P}(t)} w^{L_P}_\mu \cdot \Phi^{L_P}(I, z_\mu) + \lambda \sum_{\mu \in V^{L_M}(t)} w^{L_M}_\mu \cdot \Phi^{L_M}(I, z_\mu), \end{aligned} \tag{2}$$

where w^O_{μ,t_μ} measures priors of occurrence for different parts and λ controls the relative weight of the pose and parsing related terms.

4. Inference

The inference corresponds to maximizing $S(I,z)$ from Eqn. (2) over z. As our model follows the summarization principle [26], it naturally leads to a dynamic programming type algorithm that computes optimal part configurations from bottom to up. As the horizontal relation only exists between the "Leaf" nodes under a common "And" node, if we have already calculated the states of all nodes in the second layer, the following inference can be performed effectively on a tree due to the Markov property of our model. In other words, if we regard all cliques containing an "And" in the second layer and all its child "Leaf" nodes as super nodes, the original model can be converted to a tree model.

Hence, the maximization over positions and scales for upper level nodes can be computed very efficiently using distance transforms with linear complexity as in [6].

Since the cycles only exist in the first and second layers, the main computation cost for the proposed model lies in passing the message from "Leaf" nodes to their parent "And" node. However, there are only a limited number of "Leaf" nodes under each "And" node. Furthermore, with the filtering through appearance and spatial constraints, there are usually less than 30 hypotheses for each type of Parselets. Hence, though there are cycles at the bottom level, the algorithm is still reasonably fast.

5. Learning

We solve the unified human parsing and pose estimation under the structural learning framework. We follow the setting of [5] to perform the Parselet selection and training. As pose annotation contains no information about mixture component labeling of joint-groups, we derive these labels using k-means algorithm based on joint locations as in [24, 14]. Though such assignment is derived heuristically, it is usually found that treating these labels as latent variables will not improve the performance as these labels tend not to change over iterations [24, 14]. We thus directly use the cluster membership as the supervised definition of mixture component labels for training examples.

As the scoring function of Eqn. (2) is linear in model parameters $w = (w^{L_P}, w^{L_M}, w^O, w^A)$, it can be written compactly as $S(I, z) = w \cdot \Phi(I, z)$. Then both pose and parsing related parameters can be learned in a unified framework. Thus we learn all the parameters simultaneously rather than learning local subsets of the parameters independently or iteratively to guarantee the overall performance. Given the labeled examples $\{(I_i, z_i)\}$, the structured learning problem can be formulated in a max-margin framework as in [6]:

$$\begin{aligned} &\min_{w} \|w\|^2 + C\sum_i \xi_i \\ &\text{s.t.} \quad w \cdot (\Phi(I_i, z_i) - \Phi(I_i, z)) \geq \Delta(z_i, z) - \xi_i, \forall z, \end{aligned} \tag{3}$$

where $\Delta(z_i, z_j)$ is a loss function which penalizes the incorrect estimate of z. This loss function should give partial credit to states which differ from the ground truth slightly, and thus is defined based on [13, 5] as follows:

$$\Delta(z_i, z_j) = \sum_{\nu \in V^{L_P}(t_i) \bigcup V^{L_P}(t_j)} \delta(z_i^\nu, z_j^\nu) + \lambda \sum_{\nu \in V^{L_M}(t_i)} \min(2 * \mathrm{PCP}(z_i^\nu, z_j^\nu), 1),$$

where $\delta(z_i^\nu, z_j^\nu) = 1$, if $\nu \notin V^L(t_i) \bigcap V^L(t_j)$ or $\mathrm{sim}(d_i^\nu, d_j^\nu) \leq \sigma$. $\mathrm{sim}(\cdot, \cdot)$ 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 term penalizes both configurations with "wrong" topology and leaf nodes with wrong segments. The second term penalizes the derivation from the correct poses, where $\mathrm{PCP}(z_i^\nu, z_j^\nu)$ is the average PCP score [8] of all points in the corresponding MJGT. The optimization problem Eqn. (3) is known as a structural SVM, which can be efficiently solved by the cutting plane solver of SVMStruct [10] and the stochastic gradient descent solver in [6].

6. Experiments

6.1. Experimental Settings

Dataset: Simultaneous human parsing and pose estimation requires annotation for both body joint positions and pixel-wise semantic labeling. Traditional pose estimation datasets, such as the Parse [24] and Buffy [8], are of insufficient resolution and lack the pixel-wise semantic labeling. Hence we conduct the experiments on two recently proposed human parsing datasets. The first one is the Fashionista (FS) dataset [23], which has 685 annotated samples with clothing labels and joint annotation. This dataset is originally designed for fine-grained clothing parsing. To adapt this dataset for human parsing, we merge their labels according to the Parselet definition as in [5]. The second Daily Photos (DP) dataset [5] contains 2500 high resolution images. Due to its lack of pose information, we label the common 14 joint positions in the same manner as in [23].

Evaluation Criteria: There exist several competing evaluation protocols for human pose estimation throughout the literature. We adopt the probability of a correct pose (PCP) method described in [24], which appears to be the most common variant. Unlike pose estimation, human parsing is rarely studied and with no common evaluation protocols. Here, we utilize two complementary metrics as in [23, 5] to allow direct comparison with previous works. The first one is Average Pixel Accuracy (APA) [23], which is defined as the proportion of correctly labeled pixels in the whole image. This metric mainly measures the overall performance over the entire image. Since most pixels are background, APA is greatly affected by mislabeling a large region of background pixels as body parts. The second metric, Intersection over Union (IoU), is widely used in evaluating segmentation and more suitable for measuring the performance for each type of semantic regions. In addition, the accuracy of labels for some parts, such as "upper clothes" and "skirt" should be more important than the accuracy for "scarf", which seldom appears in images. Hence, besides the "Average IoU" (aIoU), we also calculate "Weighted IoU" (wIoU) which is calculated by accumulating each Parselet's IoU score weighted by the ratio of its pixels occupying the whole body.

Implementation Details: We use the same definition of Parselets and settings for feature extraction as in [5]. The dense SIFT, HoG and color moment are extracted as low-level features for Parselets. The size of Gaussian Mixture Model in FK is set to 128. For pose estimation, we follow [24] by using the 5×5 HoG cells for each template. The training : testing ratio is 2:1 for both datasets as in [5]. The penalty parameter C and relative weight λ are deter-

Table 1. Comparison of human pose estimation PCP scores on FS and DP datasets.

method	dataset	torso	ul leg	ur leg	ll leg	lr leg	ul arm	ur arm	ll arm	lr arm	head	avg
[24]	FS	100.0	94.2	93.0	90.9	90.1	86.5	85.2	62.3	61.9	99.2	86.3
[23]	FS	99.6	94.1	95.1	89.6	91.9	85.8	86.9	62.1	63.6	99.3	86.8
raw MJGT	FS	100.0	91.9	91.6	83.9	82.5	80.4	81.1	54.7	58.2	97.5	82.2
HPM	FS	99.5	95.3	95.6	92.2	92.7	89.9	90.9	69.6	69.7	99.1	**89.5**
[24]	DP	99.8	91.2	93.9	90.3	90.0	89.1	88.8	66.9	61.7	99.5	87.1
[23]	DP	99.8	92.0	94.2	90.9	90.0	89.5	88.7	68.2	62.6	99.5	87.5
raw MJGT	DP	99.8	90.0	92.3	89.0	88.7	85.6	85.7	60.4	48.0	99.6	83.9
HPM	DP	99.8	95.5	96.4	93.3	92.7	92.4	91.7	72.8	69.3	99.7	**90.4**

mined by 3-fold cross validation over the training set.

6.2. Experimental Results

To the best of our knowledge, there are few works handling human parsing and pose estimation simultaneously. Hence, besides the recent representative approach [23], which performs parsing and pose estimation iteratively, we also compare the proposed method with the state-of-the-art methods designed for each task separately.

Human Pose Estimation: For human pose estimation, as the experiments are conducted on these two new datasets, we only compare with several state-of-the-art methods with publicly available codes for retraining [24, 23]. The comparison results are shown in Table 1. Method [23] utilizes the results of [24] as initial estimation of pose for human parsing. The parsing results are then fed back as additional features to re-estimate the pose. However, the improvement of [23] over [24] is marginal probably because of its sequential optimization nature. As the error from initial pose estimation results will propagate to parsing, it is difficult for the re-estimation step to rectify the initial pose results from error-propagated parsing results. On the contrary, we perform human parsing and pose estimation simultaneously, which significantly improves the state-of-the-art performance [24, 23]. We also evaluate the raw MJGT baseline which only utilizes the MJGT representation and removes the Parselet from the "And-Or" graph. The worse results compared with the full HPM model verify the advantages of joint parsing and pose estimation.

Figure 5 shows some qualitative comparison results. It can be seen that all other methods fail in cases where joints are occluded by clothing, *e.g.*, wearing long dress or skirt. By contrast, with the help of Parselets and the pairwise constraints brought by the GLF, the proposed method can still obtain reasonable joint positions.

Human Parsing: For human parsing, we compare the proposed framework with the works [23] and [5]. In terms of APA, our method achieves 87% for FS dataset and 88% for DP dataset, which are superior to 86% and 87% of the current leading approach [5]. The improvement is not significant as APA metric is dominated by the background. Even naively assigning all segments as background results in a reasonably good APA of 78% for DP and 77% for FS. Therefore, the more discriminative IoU criterion is more suitable to measure the real performance of each algorithm. The detailed comparison results in terms of IoU are shown in Table 2. It can be seen that our framework is consistently better than other methods across different datasets and metrics. This significant improvement mainly comes from the complementary nature of two tasks and the strong pairwise modeling, which verifies the effectiveness of our unified parsing and pose estimation framework.

Some example human parsing results are shown in Figure 5. It can be observed that the sequential approach [23] performs much worse than ours. This may be owing to the errors propagated from the inaccurate pose estimation results as well as the lack of the ability to model the exclusive relation of different labels, which usually leads to cluttered results. Though this method can achieve much better performance with the additional information about the type of clothes in the target image as illustrated in [23], such information is usually difficult to obtain for real applications. Our method also outperforms the baseline [5], which has obvious artifacts for persons with joint crossed (*e.g.*, legs and foot). The lack of top-down information makes it difficult for the method [5] to distinguish the left shoe from the right shoe. On the contrary, by jointly modeling human parsing and pose estimation, our model can achieve reasonably good results for these cases. In addition, as the method [5] does not explicitly model the overlap between Parselets, the resultant Parselets may occlude each other seriously. For example, the "dress" Parselet is badly occluded by the "coat" Parselet in the right-bottom image. With the help of GLF, our unified model can effectively avoid the severe overlap of Parselets and thus leads to more promising results.

Finally, we want to emphasize that our goal is to explore the intrinsic correlation between human parsing and pose estimation. To achieve this, we propose the HPM which is a unified model built upon the unified representation and the novel pairwise geometry modeling. Separating our framework into different components leads to inferior results as demonstrated in Table 1 and 2. Though we use more annotations than methods for individual tasks, the promising results of our framework verify that human parsing and pose estimation are essentially complementary and thus performing two tasks simultaneously will boost the performance of each other.

7. Conclusions and Future Work

In this paper, we present a unified framework for simultaneous human parsing and pose estimation, as well as an effective feature to measure the pairwise geometric re-

Table 2. Comparison of human parsing IoU scores on FS and DP datasets.

method	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	aIoU	wIoU
[23]	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	23.8	29.9
[5]	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	41.0	51.7
HPM	FS	7.9	70.8	2.6	59.5	null	58.0	56.3	48.3	16.6	58.9	51.8	76.1	56.7	50.3	52.6	41.5	17.7	2.3	**42.8**	**54.3**
[23]	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	20.3	24.6
[5]	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	44.9	53.0
HPM	DP	26.4	74.2	8.3	47.9	43.6	64.7	53.6	70.7	17.2	59.7	53.0	78.9	67.9	64.7	55.1	39.9	16.2	6.6	**47.1**	**56.4**

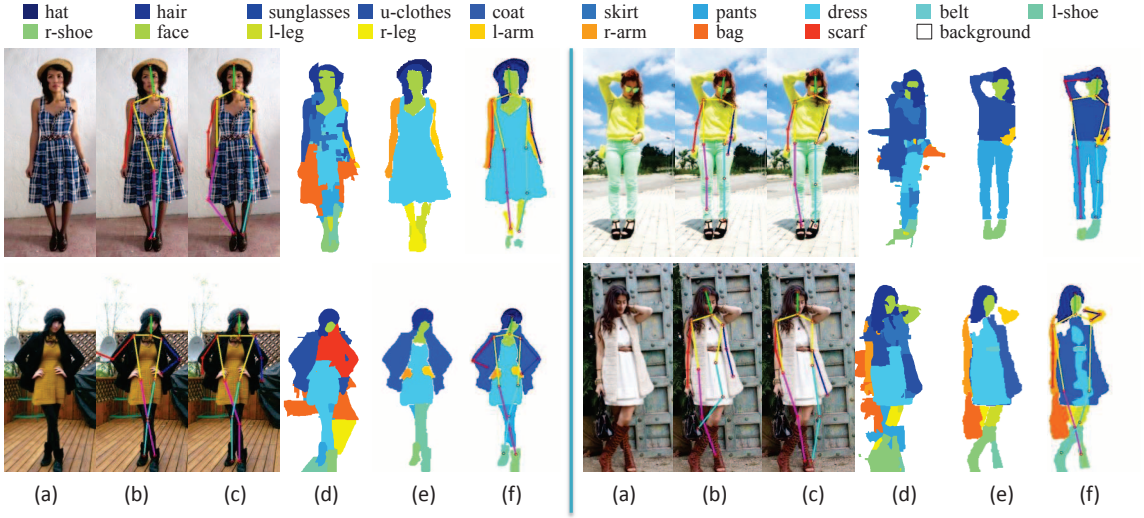

Figure 5. Comparison of human parsing and pose estimation results. (a) input image, (b) pose results from [24], (c) pose results from [23], (d) parsing results from [23], (e) parsing results from [5], and (f) our HPM results are shown sequentially.

lation between two semantic parts. By utilizing Parselets and Mixture of Deformable Templates as basic elements, the proposed Hybrid Parsing Model allows joint learning and inference of the best configuration for all parameters. The proposed framework is evaluated on two benchmark datasets with superior performance to the current state-of-the-arts in both cases, which verifies the advantage of joint human parsing and pose estimation. In the future, we plan to further explore how to integrate the fine-grained attribute analysis and extend the current framework to other object categories with large pose variance.

Acknowledgment

This work is supported by Singapore Ministry of Education under research Grant MOE2010-T2-1-087.

References

[1] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, 2009.

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

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

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

[5] J. Dong, Q. Chen, W. Xia, Z. Huang, and S. Yan. A deformable mixture parsing model with parselets. In *ICCV*, 2013.

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

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

[8] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR*, 2008.

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

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

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

[12] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing: Label transfer via dense scene alignment. In *CVPR*, 2009.

[13] B. Rothrock, S. Park, and S.-C. Zhu. Integrating grammar and segmentation for human pose estimation. 2013.

[14] B. Sapp and B. Taskar. Modec: Multimodal decomposable models for human pose estimation. In *CVPR*, 2013.

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

[16] Y. Tian, C. L. Zitnick, and S. G. Narasimhan. Exploring the spatial hierarchy of mixture models for human pose estimation. In *ECCV*. 2012.

[17] J. Tighe and S. Lazebnik. Superparsing - scalable nonparametric image parsing with superpixels. *IJCV*.

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

[19] D. Tran and D. Forsyth. Improved human parsing with a full relational model. In *ECCV*. 2010.

[20] H. Wang and D. Koller. Multi-level inference by relaxed dual decomposition for human pose segmentation. In *CVPR*, 2011.

[21] Y. Wang, D. Tran, and Z. Liao. Learning hierarchical poselets for human parsing. In *CVPR*, 2011.

[22] Y. Wang, D. Tran, Z. Liao, and D. Forsyth. Discriminative hierarchical part-based models for human parsing and action recognition. *JMLR*, 2012.

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

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

[25] 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.

[26] L. L. Zhu, Y. Chen, C. Lin, and A. Yuille. Max margin learning of hierarchical configural deformable templates (hcdts) for efficient object parsing and pose estimation. *IJCV*, 2011.