Parsing Clothing in Fashion Photographs

Kota Yamaguchi M. Hadi Kiapour Luis E. Ortiz Tamara L. Berg
Stony Brook University
Stony Brook, NY 11794, USA
{kyamagu, mkiapour, leortiz, tlberg}@cs.stonybrook.edu

Abstract

In this paper we demonstrate an effective method for parsing clothing in fashion photographs, an extremely challenging problem due to the large number of possible garment items, variations in configuration, garment appearance, layering, and occlusion. In addition, we provide a large novel dataset and tools for labeling garment items, to enable future research on clothing estimation. Finally, we present intriguing initial results on using clothing estimates to improve pose identification, and demonstrate a prototype application for pose-independent visual garment retrieval.

1. Introduction

Consider the upper east sider in her tea-dress and pearls, the banker in his tailored suit and wingtips, or the hipster in his flannel shirt, tight jeans, and black framed glasses. Our choice of clothing is tightly coupled with our socio-identity, indicating clues about our wealth, status, fashion sense, or even social tribe.

Vision algorithms to recognize clothing have a wide variety of potential impacts, ranging from better social understanding, to improved person identification [11], surveillance [26], computer graphics [14], or content-based image retrieval [25]. The e-commerce opportunities alone are huge! With hundreds of billions of dollars being spent on clothing purchases every year, an effective application to automatically identify and retrieve garments by visual similarity would have exceptional value (see our prototype garment retrieval results in Fig 1). In addition, there is a strong contextual link between clothing items and body parts – for example, we wear hats on our heads, not on our feet. For visual recognition problems such as person detection or pose identification, knowing what clothing items a person is wearing and localizing those items could lead to improved algorithms for estimating body configuration.

Despite the potential research and commercial gains of clothing estimation, relatively few researchers have explored the clothing recognition problem, mostly focused on examining the problem in limited domains [2], or recognizing only a small number of garment types [4, 13]. Our approach tackles clothing estimation at a much more general scale for real-world pictures. We consider a large number (53) of different garment types (e.g. shoes, socks, belts, rompers, vests, blazers, hats, ...), and explore techniques to accurately parse pictures of people wearing clothing into their constituent garment pieces. We also exploit the relationship between clothing and the underlying body pose in two directions – to estimate clothing given estimates of pose, and to estimate pose given estimates of clothing. We show exciting initial results on our proposed novel clothing parsing problem and also some promising results on how clothing might be used to improve pose identification. Finally, we demonstrate some results on a prototype visual garment retrieval application (Fig 1).

Figure 1: Prototype garment search application results. Query photo (left column) retrieves similar clothing items (right columns) *independent of pose and with high visual similarity*.

Our main contributions include:

- A novel dataset for studying clothing parsing, consist-

(a) Superpixels (b) Pose estimation (c) Predicted Clothing Parse (d) Pose re-estimation

Figure 2: Clothing parsing pipeline: **(a)** Parsing the image into Superpixels [1], **(b)** Original pose estimation using state of the art flexible mixtures of parts model [27]. **(c)** Precise clothing parse output by our proposed clothing estimation model (note the accurate labeling of items as small as the wearer's necklace, or as intricate as her open toed shoes). **(d)** Optional re-estimate of pose using clothing estimates (note the improvement in her left arm prediction, compared to the original incorrect estimate down along the side of her body).

ing of 158,235 fashion photos with associated text annotations, and web-based tools for labeling.

- An effective model to recognize and precisely parse pictures of people into their constituent garments.
- Initial experiments on how clothing prediction might improve state of the art models for pose estimation.
- A prototype visual garment retrieval application that can retrieve matches independent of pose.

Of course, clothing estimation is a very challenging problem. The number of garment types you might observe in a day on the catwalk of a New York city street is enormous. Add variations in pose, garment appearance, layering, and occlusion into the picture, and accurate clothing parsing becomes formidable. Therefore, we consider a somewhat restricted domain, fashion photos from Chictopia.com. These highly motivated users – fashionistas – upload individual snapshots (often full body) of their outfits to the website and usually provide some information related to the garments, style, or occasion for the outfit. This allows us to consider the clothing labeling problem in two scenarios: 1) a constrained labeling problem where we take the users' noisy and perhaps incomplete tags as the list of possible garment labels for parsing, and 2) where we consider all garment types in our collection as candidate labels.

1.1. Related Work

Clothing recognition: Though clothing items determine most of the surface appearance of the everyday human, there have been relatively few attempts at computational recognition of clothing. Early clothing parsing attempts focused on identifying layers of upper body clothes in very limited situations [2]. Later work focused on grammatical representations of clothing using artists' sketches [6]. Freifeld and Black [13] represented clothing as a deformation from an underlying body contour, learned from training examples using principal component analysis to produce eigen-clothing. Most recently attempts have been made to consider clothing items such as *t-shirt* or *jeans* as semantic attributes of a person, but only for a limited number of garments [4]. Different from these past approaches, we consider the problem of estimating a complete and precise region based labeling of a person's outfit, for general images with a large number of potential garment types.

Clothing items have also been used as implicit cues of identity in surveillance scenarios [26], to find people in an image collection of an event [11, 22, 25], to estimate occupation [23], or for robot manipulation [16]. Our proposed approach could be useful in all of these scenarios.

Pose Estimation: Pose estimation is a popular and well studied enterprise. Some previous approaches have considered pose estimation as a labeling problem, assigning most likely body parts to superpixels [18], or triangulated regions [20]. Current approaches often model the body as a collection of small parts and model relationships among them, using conditional random fields [19, 9, 15, 10], or discriminative models [8]. Recent work has extended patches to more general poselet representations [5, 3], or incorporated mixtures of parts [27] to obtain state of the art results. Our pose estimation subgoal builds on this last method [27], extending the approach to incorporate clothing estimations in models for pose identification.

Image Parsing: Image parsing has been studied as a step toward general image understanding [21, 12, 24]. We consider a similar problem (parsing) and take a related approach (CRF based labeling), but focus on estimating labelings for a particularly interesting type of object – people – and build models to estimate an intricate parse of a per-

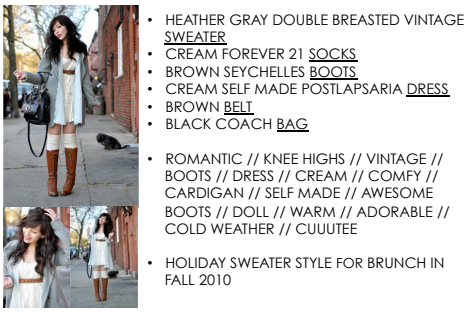


Figure 3: Example Chictopia post, including a few photos and associated meta-data about garment items and styling. If desired, we can make use of clothing tags (underlined) as potential clothing labels.

son's outfit into constituent garments. We also incorporate discriminatively trained models into the parsing process.

1.2. Overview of the Approach

We consider two related problems: 1) Predicting a clothing parse given estimates for pose, and 2) Predicting pose given estimates for clothing. Clothing parsing is formulated as a labeling problem, where images are segmented into superpixels and then clothing labels for every segment are predicted in a CRF model. Unary potentials account for clothing appearance and clothing item location with respect to body parts. Pairwise potentials incorporate label smoothing, and clothing item co-occurrence. Pose estimation is formulated as an extension to state of the art work on flexible part models [27], to incorporate estimates of clothing as an additional feature.

The remainder of the paper discusses our novel data set and labeling tools (Sec 2), our approaches to clothing parsing and pose estimation (Sec 3), results plus a peak at our prototype application for visual garment retrieval (Sec 4), and conclusions and future work (Sec 5).

2. Fashionista Dataset & Labeling Tools

We introduce a novel dataset, useful for training and testing clothing estimation techniques. This dataset consists of 158,235 photographs collected from Chictopia.com, a social networking website for fashion bloggers. On this website, fashionistas upload "outfit of the day" type pictures, designed to draw attention to their fashion choices or as a form of social interaction with peers. Because these are people who particularly care about their clothes they tend to display a wide range of styles, accessories, and garments. However, pictures are also often depicted in relatively simple poses (mostly standing), against relatively clean backgrounds, and without many other people in the picture. This makes for an ideal scenario for studying clothing!

In addition, users also provide additional outfit information in the form of tags, comments, and links, etc (*e.g.* Fig 3). We make use of the tag portion of this meta-data to extract useful information about what clothing items might be present in each photo (but can also ignore this information if we want to study clothing parsing with no prior knowledge of items). Sometimes the tags are noisy or incomplete, but often they cover the items in an outfit well.

As a training and evaluation set, we select 685 photos with good visibility of the full body and covering a variety of clothing items. For this carefully selected subset, we design and make use of 2 Amazon Mechanical Turk jobs to gather annotations. The first Turk job gathers ground truth pose annotations for the usual 14 body parts [27]. The second Turk job gathers ground truth clothing labels on superpixel regions. All annotations are verified and corrected if necessary to obtain high quality annotations.

In this ground truth data set, we observe 53 different clothing items, of which 43 items have at least 50 image regions. Adding additional labels for *hair*, *skin*, and *null* (background), gives a total of 56 different possible clothing labels – a *much larger* number than considered in any previous approach [4, 2, 6, 26, 11, 22, 25]. On average, photos include 291.5 regions and 8.1 different clothing labels. Many common garment items have a large number of occurrences in the data set (number of regions with each label denoted in parenthesis), including dress (6565), bag (4431), blouse (2946), jacket (2455), skirt (2472), cardigan (1866), t-shirt (1395), boots (1348), jeans (1136), sweater (1027), etc. However, even items probably unheard of by the fashion non-initiate, also have many occurrences – *leggings* (545), *vest* (955), *cape* (137), *jumper* (758), *wedges* (518), and *romper* (164), for example.

3. Clothing parsing

In this section, we describe our general technical approach to clothing parsing, including formal definitions of the problem and our proposed model.

3.1. Problem formulation

We formulate the clothing parsing problem as a labeling of image regions. Let I denote an image showing a person. The goal is to assign a label of a clothing or null (background) item to each pixel, analogous to the general image parsing problem. However, in this paper we simplify the clothing parsing problem by assuming that uniform appearance regions belong to the same item, as reported in [11], and reduce the problem to the prediction of a labeling over a set of superpixels. We denote the set of clothing labels by $L \equiv \{l_i\}$, where $i \in U$ denotes a region index within a set of superpixels U in I, and l_i denotes a clothing label for region indexed by i (e.g., $l_i =$ *t-shirt* or *pants*). Also let $\boldsymbol{s}_i$ denote the set of pixels in the i-th region.

In this paper, we take a probabilistic approach to the clothing parsing problem. Within our framework, we reduce the general problem to one of *maximum a posteriori*

(MAP) assignments; we would like to assign clothing labels based on the most likely joint clothing label assignments under a probability distribution $P(L|I)$ given by the model. However, it is extremely difficult to directly define such a distribution due to the varied visual appearance of clothing items. Therefore, we introduce another variable, human pose configuration, and consider the distribution in terms of interactions between clothing items, human pose, and image appearance. We denote a human pose configuration by $X \equiv \{\boldsymbol{x}_p\}$, which is a set of image coordinates $\boldsymbol{x}_p$ for body joints p, *e.g.*, *head* or *right elbow*.

Ideally, one would then like to find the joint MAP assignment over both clothing and pose labels with respect to the joint probability distribution $P(X, L|I)$ simultaneously. However, such MAP assignment problems are often computationally intractable because of the large search space and the complex structure of the probabilistic model. Instead, we split the problem into parts, solving the MAP assignment of $P(L|X, I)$ and $P(X|I)$ separately.

Our clothing parsing pipeline proceeds as follows:

1. Obtain superpixels $\{\boldsymbol{s}_i\}$ from an image I
2. Estimate pose configuration X using $P(X|I)$
3. Predict clothes L using $P(L|X, I)$
4. Optionally, re-estimate pose configuration X using model $P(X|L, I)$

Figure 2 shows an example of this pipeline. We now briefly describe each step and formally define our probabilistic model.

3.2. Superpixels

We use a recent image segmentation algorithm [1] to obtain superpixels. The algorithm provides a hierarchical segmentation, but we set the threshold value to 0.05 to obtain a single over-segmentation for each image. This process typically yields between a few hundred to a thousand regions per image, depending on the complexity of the person and background appearance (Fig 2(a) shows an example).

3.3. Pose estimation

We begin our pipeline by estimating pose $\hat{X}$ using $P(X|I)$:

$$\hat{X} \in \arg\max_X P(X|I) \ . \tag{1}$$

For our initial pose estimate, we make use of the current best implementation available to the computer vision community [27]. In addition to the above terms, this model includes an additional hidden variable representing a type label for pose mixture components, $T \equiv \{t_p\}$ for each body joint p, containing information about the types of arrangements possible for a joint. Therefore, the estimation problem is written as $(\hat{X}, \hat{T}) \in \arg\max_{X,T} P(X, T|I)$. The scoring function used to evaluate pose [27] is:

$$\begin{aligned}\ln P(X, T|I) \equiv &\textstyle\sum_p \boldsymbol{w}_p(t_p)^{\mathrm{T}} \boldsymbol{\phi}(x_p|I) + \\ &\textstyle\sum_{p,q} \boldsymbol{w}_{p,q}(t_p, t_q)^{\mathrm{T}} \boldsymbol{\psi}(x_p - x_q) - \ln Z,\end{aligned} \tag{2}$$

where, $\boldsymbol{w}$ are the model parameters, $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ are feature functions, and Z is a partition function.

3.4. Clothing labeling

Once we obtain the initial pose estimate $\hat{X}$, we can proceed to estimating the clothing labeling:

$$\hat{L} \in \arg\max_L P(L|\hat{X}, I) \ . \tag{3}$$

We model the probability distribution $P(L|X, I)$ with a second order conditional random field (CRF):

$$\begin{aligned}\ln P(L|X, I) \equiv &\sum_{i \in U} \Phi(l_i|X, I) + \sum_{(i,j) \in V} \lambda_1 \Psi_1(l_i, l_j) \\ &+ \sum_{(i,j) \in V} \lambda_2 \Psi_2(l_i, l_j|X, I) - \ln Z,\end{aligned} \tag{4}$$

where V is a set of neighboring pairs of image regions, λ_1 and λ_2 are model parameters, and Z is a partition function.

We model the unary potential function Φ using the probability of a label assignment, given the feature representation of the image region $\boldsymbol{s}_i$:

$$\Phi(l_i|X, I) \equiv \ln P(l_i|\phi(\boldsymbol{s}_i, X)). \tag{5}$$

In this paper, we define the feature vector ϕ as the concatenation of (1) normalized histograms of RGB color, and (2) normalized histogram of CIE L*a*b* color, (3) histogram of Gabor filter responses, (4) normalized 2D coordinates within the image frame, and (5) normalized 2D coordinates with respect to each body joint location $\boldsymbol{x}_p$. In our experiments, we use 10 bins for each feature type. Using a 14-joint pose estimator, this results in a 360 dimensional sparse representation for each image region. For the specific marginal probability model $P(l_i|\phi(\boldsymbol{s}, X))$, we experimentally evaluated a few distributions and found that logistic regression works well for our setting.

The binary potential function Ψ_1 is a log empirical distribution over pairs of clothing region labels in a single image:

$$\Psi_1(l_i, l_j) \equiv \ln \tilde{P}(l_i, l_j). \tag{6}$$

This term serves as a prior distribution over the pairwise co-occurrence of clothing labels (*e.g.* shirts are near blazers, but not shoes) in neighboring regions within an image. We compute the function by normalizing average frequency of neighboring label pairs in training samples.

The last binary potential in (4) estimates the probability of neighboring pairs having the same label (*i.e.* label smoothing), given their features, ψ:

$$\Psi_2(l_i, l_j|X, I) \equiv \ln P(l_i = l_j|\psi(\boldsymbol{s}_i, \boldsymbol{s}_j, X)). \tag{7}$$

In this paper, we define the feature transformation to be $\psi(\boldsymbol{s}_i, \boldsymbol{s}_j) \equiv [(\phi(\boldsymbol{s}_i) + \phi(\boldsymbol{s}_j))/2, |\phi(\boldsymbol{s}_i) - \phi(\boldsymbol{s}_j)|]$. As with the unary potential, we use logistic regression for this probability distribution.

Because of the loopy structure of our graphical model, it is computationally intractable to solve (3) exactly. Therefore, we use belief propagation to obtain an approximate MAP assignment, using the libDAI [17] implementation.

In practice, regions outside of the bounding box around pose estimation are always background. Therefore, in our experiment, we fix these outside regions to *null* and run inference only within the foreground regions.

3.5. Pose re-estimation

The original pose estimations may be inaccurate. We believe that these estimates may be improved by considering clothing predictions during pose identification (because clothes and pose are tightly coupled). Given the predicted clothing labels $\hat{L}$, we try to improve our prior MAP pose assignment $\hat{X}$ by computing the posterior MAP conditioned on $\hat{L}$ in (1):

$$\hat{X} \in \arg\max_X P(X|\hat{L}, I) \ . \tag{8}$$

To incorporate clothing item predictions in the pose estimation process we modify (1). To do this, we update the appearance feature $\boldsymbol{\phi}(x_p|I)$ in (1) to $\boldsymbol{\phi}(x_p|L, I)$, where our new appearance feature includes HoG as well as normalized histograms of clothing labels computed at the location $\boldsymbol{x}_p$.

3.6. Training

Training of our clothing parser includes parameter learning of the pose estimator $P(X|I)$ and $P(X|L, I)$, learning of potential functions in $P(L|X, I)$, and learning of CRF parameters in (4).

Pose estimator: The training procedure of [27] uses separate negative examples, sampled from scene images to use the pose estimator as a detector. Since our problem assumes a person is shown, we do not use a scene based negative set, but rather mine hard negative examples using false detections in our images. We treat a detection as negative if less than 30% of the body parts overlap with their true locations with ratio more than 60%.

Potential functions: We learn the probability distributions $P(l_i|\phi)$ and $P(l_i = l_j|\psi)$ in (5) and (7) using logistic regression with L2 regularization (liblinear implementation [7]). For each possible clothing item, e.g. *shirt* or *boots* we learn the distribution its regional features, $P(l_i|\phi)$. We learn this model using a one-versus-all approach for each item. This usually introduces an imbalance in the number of positive vs negative examples, so the cost parameter is weighted by the ratio of positive to negative samples.

CRF parameters: Our model (4) has two parameters λ_1 and λ_2. We find the best parameters by maximizing cross validation accuracy over pixels in our training data using line search and a variant of the simplex method (`fminsearch` in Matlab). In our experiment, typically both λ_1 and λ_2 preferred small values (e.g., 0.01-0.1).

Method	Pixel acc	mAGR
Full-a	**89.0** $\pm$ 0.8	63.4 $\pm$ 1.5
with truth	89.3 $\pm$ 0.8	64.3 $\pm$ 1.3
without pose	86.0 $\pm$ 1.0	58.8 $\pm$ 2.1
Full-m	88.3 $\pm$ 0.8	**69.6** $\pm$ 1.7
with truth	88.9 $\pm$ 0.7	71.2 $\pm$ 1.5
without pose	84.7 $\pm$ 1.0	64.6 $\pm$ 1.6
Unary	88.2 $\pm$ 0.8	**69.8** $\pm$ 1.8
Baseline	77.6 $\pm$ 0.6	12.8 $\pm$ 0.1

Table 1: Clothing Parsing performance. Results are shown for our model optimized for accuracy (**top**), our full model optimized for mAGR (**2nd**), our model using unary term only (**3rd**), and a baseline labeling (**bottom**).

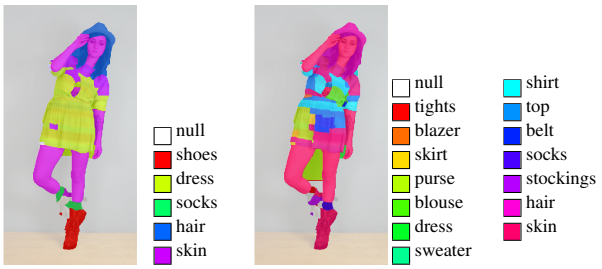
Garment	Full-m	with truth	without pose
background	95.3± 0.4	95.6± 0.4	92.5± 0.7
skin	74.6± 2.7	76.3± 2.9	78.4± 2.9
hair	76.5± 4.0	76.7± 3.9	69.8± 5.3
dress	65.8± 7.7	67.7± 9.4	50.4±10.2
bag	44.9± 8.0	47.6± 8.3	33.9± 4.7
blouse	63.6± 9.5	66.2± 9.1	52.1± 8.9
shoes	82.6± 7.2	85.0± 8.8	77.9± 6.6
top	62.0±14.7	64.6±13.1	52.0±13.8
skirt	59.4±10.4	60.6±13.2	42.8±14.5
jacket	51.8±15.2	53.3±13.5	45.8±18.6
coat	30.8±10.4	31.1± 5.1	22.5± 8.8
shirt	60.3±18.7	60.3±17.3	49.7±19.4
cardigan	39.4± 9.5	39.0±12.8	27.9± 8.7
blazer	51.8±11.2	51.7±10.8	38.4±14.2
t-shirt	63.7±14.0	64.1±12.0	55.3±12.5
socks	67.4±16.1	67.8±19.0	74.2±15.0
necklace	51.3±22.5	46.5±20.1	16.2±10.7
bracelet	49.5±19.8	56.1±17.6	45.2±17.0

Table 2: Recall for selected garments

4. Experimental Results

We evaluate the performance of our approach using 685 annotated samples from the Fashionista Dataset (described in Sec 2). All measurements use 10-fold cross validation (9 folds used for training, and the remaining for testing). Since the pose estimator contains some random components, we repeat this cross validation protocol 10 times.

In the remainder of this section we discuss quantitative (Sec 4.1) and qualitative (Sec 4.2) evaluations of our proposed clothing parsing model, demonstrate intriguing initial results on incorporating clothing estimates to improve pose identification (Sec 4.3), and finally show a prototype garment retrieval application (Sec 4.4).



(a) Items known (b) Items unknown

Figure 4: Clothing parsing with garment meta-data (left) and without meta-data (right). Confusion between garments is increased in the unconstrained case, but still improves over the baseline (Table 1).

4.1. Clothing Parsing Accuracy

We measure performance of clothing labeling in two ways, using average pixel accuracy, and using mean Average Garment Recall (mAGR). mAGR is measured by computing the average labeling performance (recall) of the garment items present in an image, and then the mean is computed across all images. Table 1 shows a comparison for 8 versions of our approach. Full-a and Full-m are our models with CRF parameters learned to optimize pixel accuracy and mAGR respectively (note that the choice of which measure to optimize for is application dependent). The most frequent label present in our images is *background*. Naively predicting all regions to be *background* results in a reasonably good **77%** accuracy. Therefore, we use this as our baseline method for comparison. Our model (Full-a) achieves a much improved **89%** pixel accuracy, close to the result we would obtain if we were to use ground truth estimates of pose (**89.3%**). If no pose information is used, clothing parsing performance drops significantly (**86%**). For mAGR, the Unary model achieves slightly better performance (**69.8%**) over the full model because smoothing in the full model tends to suppress infrequent (small) labels.

Finally, we also report results on the general clothing parsing problem (with no prior knowledge about items from meta-data). As seen in Fig 4, the full parsing problem with all 53 garment possibilities is quite challenging, but our method still obtains 80.8% pixel accuracy, a cross-validated gain of 3% over the baseline method.

4.2. Qualitative evaluation

We also test our clothing parser on all 158k un-annotated samples in our Fashionista dataset. Since we don't have ground truth labels for these photos, we just report qualitative observations. From these results, we confirm that our parser predicts good clothing labels on this large and varied dataset. Figure 5 shows some good parsing results, even handling relatively challenging clothing (e.g. small hats, and partially occluded shoes). Generally the parsing problem becomes easier in highly distinguishable appearance situations, such as on clean backgrounds, or displaying distinctive clothing regions. Failure cases (Fig 6) are observed due to ambiguous boundaries between foreground and background, when initial pose estimates are quite incorrect, or in the presence of very coarse patterns. Other challenges include pictures with out of frame body joints, close ups of individual garment items, or no relevant entity at all.

Discussion of Superpixels: Our approach assumes that each superpixel has the same clothing label and encourages over-segmentation to make this assumption nearly true. However, in some cases the superpixel segmentation does not correctly separate regions. This is likely to occur in an image with nearly invisible boundaries, such as a black-haired person wearing a black jacket with black pants. This issue is an age old segmentation problem and very difficult to solve. We could for example, consider pixel-wise labeling rather than superpixel, with the drawback of significant increase in the problem size for inference (but still might not observe significant improvements).

Method	PCP
No clothing (initial)	86.5 ± 1.5
With clothing	**86.9** ± 1.4
True clothing	89.5 ± 1.5

Table 3: Pose estimation performance. Initial state of the art performance (**top** - trained and evaluated on our data), our re-estimate of pose using a model incorporating predicted clothing estimates (**middle**), and pose re-estimation performance given ground truth clothing parse (**bottom**).

4.3. Pose Re-Estimation Accuracy

Finally, we also report initial experiments on pose re-estimation using clothing predictions. Pose estimation is a well-studied problem with very effective methods [8, 5, 3, 27]. For evaluation we measure performance as the probability of a correct pose (PCP) [27], which computes the percentage of body parts correctly overlapping with the ground truth parts. Table 3 and 4 summarizes performance. Current methods [27] obtain a cross-validated PCP of **86.5%** on our data set. Using our estimated clothing labels, we achieve **86.9%**. As motivation for future research on clothing estimation, we also observe that given true clothing labels our pose re-estimation system reaches a PCP of **89.5%**, demonstrating the potential usefulness of incorporating clothing into pose identification.

4.4. Retrieving Visually Similar Garments

We build a prototype system to retrieve garment items via visual similarity in the Fashionista dataset. For each

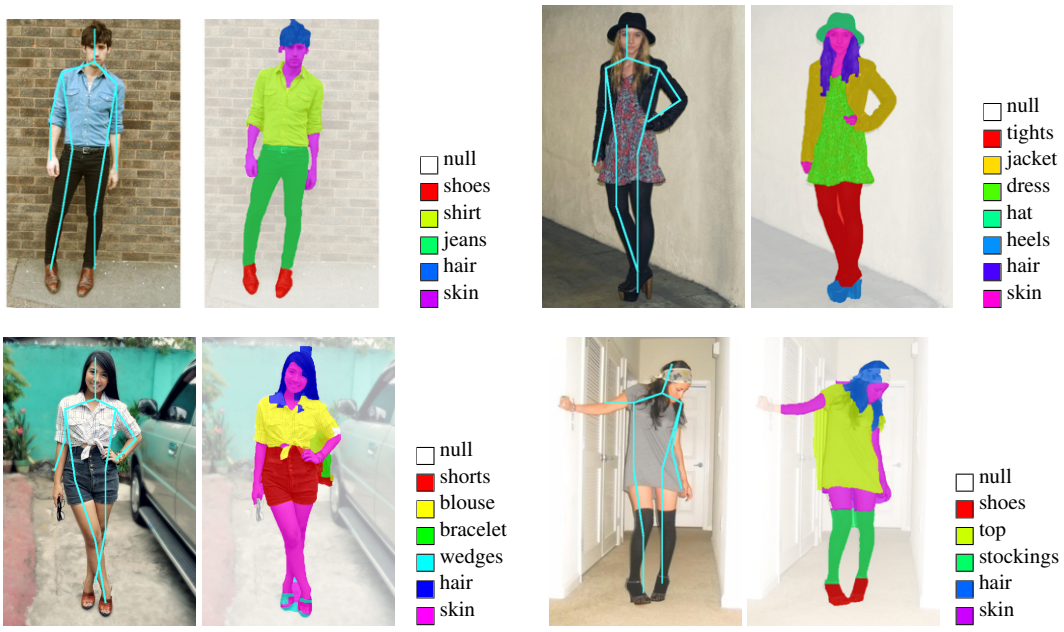
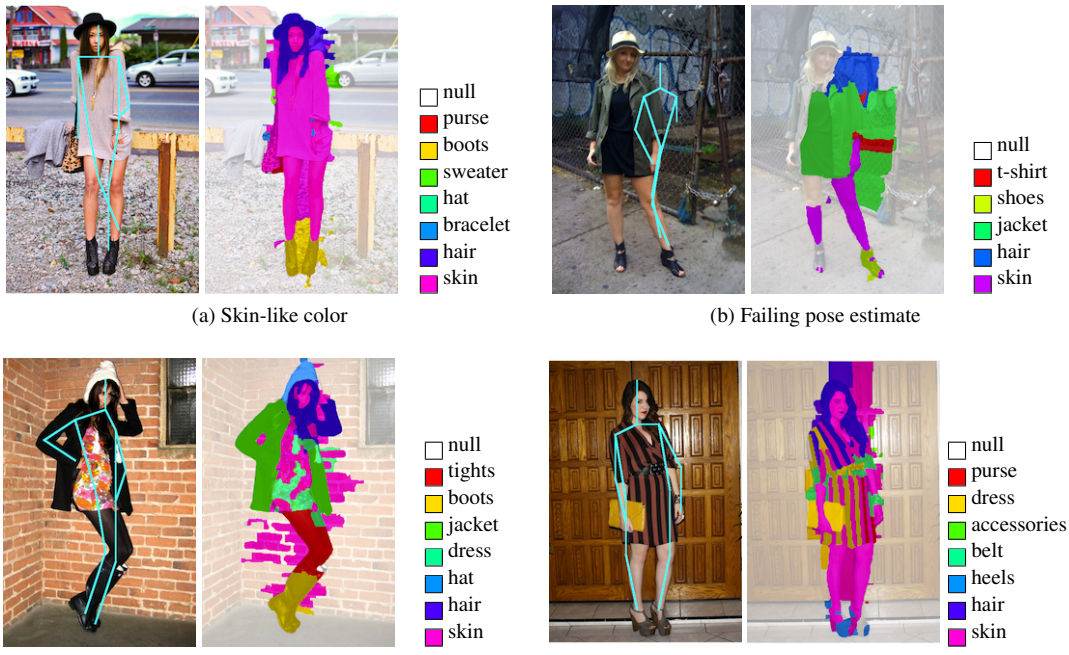


Figure 5: Example successful results on the Fahionista dataset.



(a) Skin-like color

(b) Failing pose estimate

(c) Spill in the background

(d) Coarse pattern

Figure 6: Example failure cases

parsed garment item, we compute normalized histograms of RGB and L*a*b* color within the predicted labeled region, and measure similarity between items by Euclidean distance. For retrieval, we prepare a query image and obtain a list of images ordered by visual similarity. Figure 1 shows a few of top retrieved results for images displaying *shorts*, *blazer*, and *t-shirt* (query in leftmost col, retrieval results in right 4 cols). These results are fairly representative for the more frequent garment items in our dataset. While we don't pursue this further here, this fun result demonstrates

Method	torso	ul leg	ur leg	ll leg	lr leg	ul arm	ur arm	ll arm	lr arm	head
No clothing	100.0±0.2	94.3±2.1	93.8±2.4	90.8±3.0	90.3±3.7	86.6±3.9	85.3±3.4	62.8±6.3	62.2±6.1	99.5±0.7
With clothing	99.9±0.3	94.3±2.3	95.3±2.1	89.4±3.9	93.3±3.1	84.7±3.8	86.6±3.6	61.8±5.5	64.9±6.6	99.2±1.1
True clothing	100.0±0.1	94.3±2.9	96.2±2.0	90.7±3.3	94.7±2.7	87.7±3.6	89.9±3.1	70.4±5.0	71.7±5.9	99.5±0.9

Table 4: Limb detection rate

the potential for visual garment retrieval applications of the future!

5. Conclusions and Future Work

This paper proposes an effective method to produce an intricate and accurate parse of a person's outfit. Two scenarios are explored: parsing with meta-data provided garment tags, and parsing with unconstrained label sets. A large novel data set and labeling tools are also introduced. Finally, we demonstrate intriguing initial experiments on using clothing estimates to improve human pose prediction, and a prototype application for visual garment search.

In future work, we would like to consider solutions to some of the observed challenges of clothing parsing, including: considering partial body pose estimates, using multiple segmentations to deal with inaccuracies in a single segmentation, and incorporating higher level potentials for longer range models of garment items.

References

[1] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *PAMI*, 33(5):898 –916, 2011. 2, 4

[2] A. Borràs, F. Tous, J. Lladós, and M. Vanrell. High-level clothes description based on colour-texture and structural features. In *Pattern Recognition and Image Analysis*, pages 108–116. Springer Berlin / Heidelberg, 2003. 1, 2, 3

[3] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. In *ECCV*, 2010. 2, 6

[4] L. Bourdev, S. Maji, and J. Malik. Describing people: Poselet-based attribute classification. In *ICCV*, 2011. 1, 2, 3

[5] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3D human pose annotations. In *ICCV*, 2009. 2, 6

[6] H. Chen, Z. J. Xu, Z. Q. Liu, and S. C. Zhu. Composite templates for cloth modeling and sketching. In *CVPR*, 2006. 2, 3

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

[8] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 32(9):1627 –1645, 2010. 2, 6

[9] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR*, 2008. 2

[10] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Pose search: Retrieving people using their pose. In *CVPR*, 2009. 2

[11] A. Gallagher and T. Chen. Clothing cosegmentation for recognizing people. In *CVPR*, 2008. 1, 2, 3

[12] S. Gould, T. Gao, and D. Koller. Region-based segmentation and object detection. In *NIPS*, 2009. 2

[13] P. Guan, O. Freifeld, and M. J. Black. A 2D human body model dressed in eigen clothing. *ECCV*, pages 285–298, 2010. 1, 2

[14] N. Hasler, C. Stoll, B. Rosenhahn, T. Thormählen, and H.-P. Seidel. Estimating body shape of dressed humans. *Computers and Graphics*, 33(3):211 – 216, 2009. 1

[15] E. Marcin and F. Vittorio. Better appearance models for pictorial structures. In *BMVC*, September 2009. 2

[16] S. Miller, M. Fritz, T. Darrell, and P. Abbeel. Parametrized shape models for clothing. In *ICRA*, 2011. 2

[17] J. M. Mooij. libDAI: A free and open source C++ library for discrete approximate inference in graphical models. *JMLR*, 11:2169–2173, Aug. 2010. 5

[18] G. Mori, X. Ren, A. Efros, and J. Malik. Recovering human body configurations: combining segmentation and recognition. In *CVPR*, 2004. 2

[19] D. Ramanan. Learning to parse images of articulated bodies. In *NIPS*, pages 1129–1136, 2006. 2

[20] X. Ren, A. Berg, and J. Malik. Recovering human body configurations using pairwise constraints between parts. In *ICCV*, 2005. 2

[21] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *ECCV*, 2006. 2

[22] J. Sivic, C. L. Zitnick, and R. Szeliski. Finding people in repeated shots of the same scene. In *BMVC*, 2006. 2, 3

[23] Z. Song, M. Wang, X. Hua, and S. Yan. Predicting occupation via human clothing and contexts. In *ICCV*, 2011. 2

[24] J. Tighe and S. Lazebnik. Superparsing: scalable nonparametric image parsing with superpixels. In *ECCV*, 2010. 2

[25] M. Weber, M. Bäuml, and R. Stiefelhagen. Part-based clothing segmentation for person retrieval. In *AVSS*, 2011. 1, 2, 3

[26] M. Yang and K. Yu. Real-time clothing recognition in surveillance videos. In *ICIP*, 2011. 1, 2, 3

[27] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, pages 1385 –1392, 2011. 2, 3, 4, 5, 6