DRESSING ATTRIBUTES PARSING

Daitao Xing, Jun Chen, Jin Huang

{dx383, jc7459, jh5442}@nyu.edu

ABSTRACT

This paper aims at parsing fashion clothes using an efficient method. Clothes parsing can be an extremely challenging project due to the large quantity of different clothes and types. In addition, there could be countless collocation with different styles and occlusion. We use a large dataset with thousands of fashion clothes with labels, which includes 58 main types of clothes and accessories to enable both current and future research. We finally present remarkable initial results on the topic by using MRF methods combined with location information and demonstrate a feasible prototype for clothes parsing task.

***Index Terms*—** Clothing Parsing, Markov Random Field, Segmentation and Labeling, Bag-of-Words

1. INTRODUCTION

As we may know, clothing is tightly relevant to our daily life because we choose different clothes for different occasion; the choice may vary a lot from job to entertainment and has something to do with our social status and personal taste. Dressing Attributes parsing has a great potential impact on e-commerce market. It will improve the precision of recommender systems by better understanding and predicting the consumers preference. However, this could be a challenging problem due to the large number of possible garment items, variations in configuration, garment appearance, layering, and occlusion.

In computer vison field, some interesting research [1] [2] has been done and given promising results. The basic idea for the task is to divide the image into superpixels and carry out the experiment on them. However, it can be expensive and time-consuming to analyze superpixels and give accurate labels to them. So, it raises the question on how to do clothes parsing on the integral image.

In this paper, we try to explore a method to parse different dressing attributes from photographs. We recognize different items like shoes, clothing or hat from pictures through segmentation and the label. Then we step further and recognize the different cloth types like T-shirt, suits, beachwear, shorts etc. Finally, we use Markov Random Field(*MRF*) to improve its accuracy.

Our main achievements in the project includes:

- Using a diverse dataset with more that 50 types of clothes and accessories and more than 1,000 images for the task.

- Completing the extraction of foreground.

- Training the unary and binary function for the *MRF* and finish the segmentation using graphic cut and achieve the satisfied result.

- Predicting the type of the garment and improve the result.

Of course, clothing parsing is a sophisticated problem, you can observe countless types of garments and collocation styles in one day in the New York City. Therefore, we only choose a part of the images on one website called *Chictopia*. This is a group of users who cares much about fashion and clothing and upload their daily clothes collocation online. So, it is possible for us to consider various types of garments and verify our prediction.

1.1. Related work

Some results have been carried out during previous research, but most of them focus on clothing style and appearance [3]. Later works studied on blocking models to segment clothes for highly occluded group images [4]. Recent approaches incorporated shape-based human model [5], or pose estimation and supervised region labeling [2], and achieved impressive results.

Accurate clothing parsing is, in fact, a topic risen recently. Early clothing parsing attempts to focus on identifying layers of upper body clothes in very limited situations [6]. Later work focused on grammatical representations of clothing using artists sketches [3]. Freifeld and Black [7] represented clothing as a deformation from an underlying body contour, learned from training examples using principal component analysis to produce Eigen-clothing. Latest work recognizes and precisely parse pictures of people into their constituent garments. Wei Yang, Ping Luo *et el* [1] use probabilistic function and E-SVM model to do co-parsing for joint segmentation. Kota Yamaguchi, M. Kiapour *et el* [2] use an innovate way, they do both predicting a clothing parse given estimates of pose, and the predicting pose given the estimates for clothing by pose estimation and finally achieve a prototype application for visual garment retrieval.

1.2. Overview of the Approach

Our project can be divided into the following steps:

1. Feature extraction

   Features play an important role in image labeling and are used in both unary and binary potential functions in the *MRF* model. This will be discussed in detail in Sec. 3.3.1.

2. Foreground extraction

   We use pedestrian and face detection based on superpixels and binary MRF to extract the foreground, which will be discussed in detail in Sec. 3.3.2.

3. Segmentation and Labeling

   We use the *MRF* model to do the tasks of segmentation and labeling, which will be discussed further in Sec. 3.2.

2. DATASET DESCRIPTION

We use the dataset from *Chitopia.com*, which contains more than 150,000 photographs, it is an active social website with fresh and interesting new photographs. To simplify the problem, we randomly choose about 1,000 photographs as training set and another 1,000 photographs as the testing set.

Moreover, we use a set of mat files which contain the labels [2]. The mat files give label information on both pixel and image level, which provide us a good precondition for training and testing.

3. CLOTHEING PARSING

In this section, we will describe the general approach and our model.

3.1. Problem Formulation

We regard clothing parsing problem as a classification. Let $\mathcal{I}$ be an image and $\mathcal{A}$ its attributes. The goal is to assign a label, which indicates what item it is, to each pixel, given its attributes. However, as an image has millions of pixels, predicting all the pixel labels is time-consuming. We simplify the problem by introducing the concept of *superpixel*. First, we segment the image into several superpixels and assume pixels in the same superpixel belong to the same label. Suppose $\mathcal{S} = \{s_i\}$ is the set of superpixel and $\mathcal{L} = \{l_i\}$ the set of all the labels, the task is to predict the posterior probability $\Pr(l_i|s_i)$.

3.2. Mathematical Formulation

We will assign the most probable label to a superpixel:

$$\hat{L} = \arg\max_{L} \Pr(L|I) \tag{1}$$

Similar to [2], we model the posterior probability with second-order Markov random field(*MRF*)

$$\log \Pr[L|I] = \sum_{i \in U} \phi(l_i|I) + \lambda_1 \sum_{(i,j) \in V} \psi_1(l_i, l_j) + \lambda_2 \sum_{(i,j) \in V} \psi_2(l_i, l_j|I) - \log Z \tag{2}$$

where U is the single superpixel region, V is the pairwise superpixels that are adjacent, ϕ is unary potential, ψ_1 and ψ_2 are binary potentials, λ_1 and λ_2 are smoothing parameters, and Z is partition function.

The unary potential ϕ is the logarithmic probability conditioned on the extracted features of the superpixel

$$\phi(l_i|I) = \log \Pr[l_i|f(s_i)] \tag{3}$$

where $f(s_i)$ is the feature function of ith superpixel.

The first binary potential is the logarithm of the prior probability of adjacency of two labels because we assume this result will indicate how likely two items co-occur. For example, hats are unlikely with shoes while scarves are likely with coats. The second binary potential is the logarithmic probability of the two superpixels with the same label conditional on their features

$$\psi_2(l_i, l_j|I) = \log \Pr[l_i = l_j|f(s_i), f(s_j)] \tag{4}$$

In our model, we define the binary potential function as

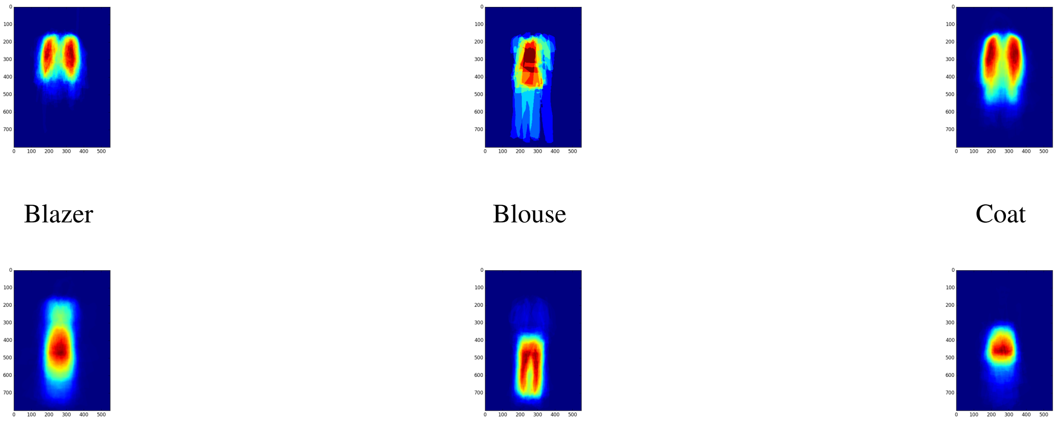
$$\psi_2 = \begin{cases} 2, & \text{if } s_i = s_j \\ 2exp(\frac{-\sum_k (x_{ik} - x_{jk})^2}{10}) & \text{otherwise.} \end{cases} \tag{5}$$

where s_i and s_j mean the superpixel in which the i- and j-th pixels are, x_{ik} and x_{jk} are their color features.

3.3. Preprocessing

3.3.1. *Feature Extraction*

In this paper, we select the following features: (1) normalized histograms of RGB color, and (2) normalized histogram of CIE L*a*b* color, (3) normalized 2D coordinates in the image frame, (4) human pose estimators [8], and (5) the combination of bag of visual words and SIFT features [9]. We experimentally evaluated on several classification models and finally found that random forest works well for our setting.



Blazer Blouse Coat

Dress Jeans Skirt

Fig. 1. Visualization of clothelets of images.

3.3.2. *Foreground extraction*

Because of the colorfulness of the background, we first separate foreground from background to simplify our work. We use pedestrian detection and face detection to extract the person in an image. To improve the accuracy and discard unnecessary background, we use *MRF* to do the binary classification and only extract superpixels that are labeled foreground.

3.3.3. *Clothelets*

We also exploit the statistical dependency between the location on the human body and garment type [10]. Its goal is to make use of the fact of the position relationship of human body. We compute a likelihood of each garment appearing in a particular relative location of the human pose. In particular, for each training example we take a region around the location of each joint (and limb), the size of which corresponds to the size of the joint part template encoded in [8]. The Fig. 1 shows the result of clothelets mask for the average output of the information.

3.3.4. *human pose estimators*

We also add pose information to our model thus to improve the performance of the function [8]. We mainly use the feature pyramid and greedy search to establish detection boxes to abstract the features of human body and give head position, leaf arm, right arm, body, left leg and right leg. By filtering the best detection, the result matrix can be used in the training step.

3.3.5. *Bag-of-words Model*

Motivated by [9], we construct a bag-of-word(*BOW*) model that can extract the SIFT feature histogram of the superpixel. We use K-means clustering algorithm to build a dictionary that can represent clothing items. Then we use generated dictionary to encode previous SIFT local features. And we apply one-vs-all support vector machine on the features. The resulting classifier can help predict to which label a superpixel belongs, regardless of its surroundings.

4. EXPERIMENT RESULTS

In this section, we will illustrate and evaluate the result of our model.

We test our model on 311 images of Fashionista dataset [2], which have 30 different labels. In the training set, we use ground truth label to extract features and use random forest to train the classifier. In the testing set, we use Simple Linear Iterative Clustering (*SLIC*) to segment image into several superpixels and assume that all the pixels in the same superpixel belong to the same label. This can greatly improve the running time of the program.

The full confusion matrix of the labels are shown in Table 1,whose trace is the accuracy and are hightailed. We achieve the satisfactory result: the best of 93.81% and the worst of 70.45% of accuracy in our unary function.

For the *MRF* mode, the average accuracy of the labeling is 90.1%, including the background.

After applying MRF, we can get the segmentation and labeling of each superpixel in an image. Some of the good results are shown in Fig. 2, in which the labels are correctly matched and displayed. We achieve average 80% of accuracy in labeling and segmentation. it can be seen that not only

Table 1. Confusion matrix for our approach

	accessories	bag	belt	blazer	blouse	boots	bracelet	coat	dress	hair	hat	jacket	jeans	pants	purse	sandals	scarf	shirt	shoes	shorts	skin	skirt	socks	stockings	suit	sunglasses	sweater	t-shirt	top	vest
accessories	0.8486	0.0141	0	0	0	0	0	0.0317	0.0106	0.007	0	0	0	0.0317	0.0035	0	0	0	0	0	0.0423	0	0	0	0.007	0	0	0.0035	0	0
bag	0	0.9106	0	0.0004	0.0013	0	0	0.0296	0.003	0.0004	0	0.0013	0.0055	0.0301	0.0004	0	0	0.0013	0.0008	0	0.0123	0.0013	0	0.0004	0.0013	0	0	0	0	0
belt	0	0	0.8502	0.004	0.0202	0	0	0.0648	0.0121	0	0	0	0.004	0.0202	0	0	0	0.0162	0	0	0	0.004	0	0	0	0	0.004	0	0	0
blazer	0	0.0029	0.0005	0.863	0.0126	0	0	0.0789	0.0053	0.0058	0	0.0005	0.0015	0.0063	0	0	0	0.0024	0	0.0005	0.0106	0.0005	0	0	0.0082	0	0.0005	0	0	0
blouse	0	0.0011	0	0.0007	0.8988	0	0	0.064	0.0014	0.0035	0	0.0004	0.0011	0.0067	0	0	0	0.0042	0	0	0.011	0	0	0	0.0042	0	0.0028	0	0	0
boots	0	0	0	0	0	0.7985	0	0	0	0	0	0	0.0073	0.044	0	0	0	0	0.1502	0	0	0	0	0	0	0	0	0	0	0
bracelet	0	0	0	0	0	0	0.7692	0	0	0	0	0	0	0	0	0	0	0.0769	0	0	0.1538	0	0	0	0	0	0	0	0	0
coat	0	0.0022	0.0001	0.0005	0.0058	0	0	0.9497	0.0021	0.0038	0	0.0002	0.002	0.0205	0.0007	0	0	0.0027	0	0	0.0046	0.0011	0	0	0.0033	0	0.0005	0.0001	0	0
dress	0	0.0014	0.0004	0.0008	0.0054	0	0	0.0229	0.8953	0.0017	0	0.0002	0.0041	0.0525	0	0	0	0.0017	0.0031	0	0.0058	0.0012	0	0	0.0033	0	0.0002	0	0	0
hair	0	0	0	0	0.0026	0	0	0.0159	0.0006	0.9574	0	0	0	0	0	0	0	0.0006	0	0	0.0207	0	0	0	0.0011	0.0009	0	0	0.0003	0
hat	0	0	0	0	0	0	0	0	0	0.1824	0.8024	0	0	0	0	0	0	0	0	0	0.0152	0	0	0	0	0	0	0	0	0
jacket	0	0.0006	0	0.0044	0.0088	0	0	0.0823	0.0031	0.0063	0	0.8637	0	0.0063	0.0006	0	0.0006	0.0025	0	0	0.0044	0	0	0	0.0138	0	0.0019	0.0006	0	0
jeans	0	0.0021	0.0002	0.0007	0.0014	0	0	0.0158	0.003	0.0002	0	0	0.8701	0.092	0	0	0	0.0011	0.0073	0	0.0034	0.0011	0	0	0.0011	0	0.0002	0	0	0
pants	0	0.0007	0	0	0.0001	0	0	0.0054	0.0026	0	0	0	0.0069	0.9742	0	0	0	0.0004	0.0053	0	0.0043	0.0001	0	0	0	0	0	0.0001	0	0
purse	0	0.0312	0	0.0016	0.0016	0	0	0.0336	0.004	0	0	0	0.0016	0.0264	0.8737	0	0	0.0016	0	0	0.0168	0.0056	0	0	0.0016	0	0.0008	0	0	0
sandals	0	0	0	0	0	0	0	0	0	0	0	0	0.0065	0.0065	0	0.7806	0	0	0.1806	0	0.0258	0	0	0	0	0	0	0	0	0
scarf	0	0.0013	0	0.0027	0.0147	0	0	0.091	0.0013	0.0201	0	0	0	0.0067	0	0	0.8327	0.004	0	0	0.0174	0	0	0	0.0067	0	0.0013	0	0	0
shirt	0	0.0012	0	0.0024	0.0178	0	0	0.0629	0.0032	0.0016	0	0.0004	0.0004	0.0115	0	0	0	0.8749	0	0	0.0143	0.0012	0	0	0.0067	0	0.0012	0.0004	0	0
shoes	0	0	0	0	0	0	0	0	0	0	0	0	0.0018	0.0131	0	0	0	0	0.9831	0	0.0018	0	0	0.0004	0	0	0	0	0	0
shorts	0	0.0024	0	0	0.0012	0	0	0.0224	0.0059	0	0	0	0.0294	0.1048	0	0	0	0.0024	0	0.8257	0.0035	0.0024	0	0	0	0	0	0	0	0
skin	0	0.0004	0.0001	0.0008	0.0036	0	0	0.0137	0.0028	0.0083	0	0.0004	0.0013	0.0294	0.0006	0	0	0.0013	0.0078	0	0.9266	0.0009	0	0	0.0015	0.0001	0.0001	0.0003	0	0.0001
skirt	0	0.005	0	0	0.0008	0	0	0.0149	0.0075	0	0	0.0004	0.0091	0.0807	0.0004	0	0	0.0004	0.0008	0	0.0099	0.8688	0	0	0.0012	0	0	0	0	0
socks	0	0	0	0	0	0	0	0	0	0	0	0	0.0227	0.1705	0	0	0	0	0.0909	0	0.0114	0	0.7045	0	0	0	0	0	0	0
stockings	0	0.0014	0	0	0	0	0	0.0014	0.0042	0	0	0	0.0169	0.1455	0	0	0	0	0.0085	0	0.0141	0	0	0.8079	0	0	0	0	0	0
suit	0	0.0012	0	0.0031	0.0087	0	0	0.0885	0.0031	0.0022	0	0.0006	0.0016	0.0112	0.0006	0	0	0.0075	0	0	0.0097	0	0	0	0.8595	0	0.0019	0.0006	0	0
sunglasses	0	0	0	0	0	0	0	0	0	0.0947	0	0	0	0	0	0	0	0	0	0	0.0368	0	0	0	0	0.8684	0	0	0	0
sweater	0	0	0	0.0041	0.0083	0	0	0.0731	0.0016	0.0031	0	0.0005	0.0005	0.0047	0	0	0.0005	0.0057	0	0	0.0093	0.0016	0	0	0.0078	0	0.8793	0	0	0
t-shirt	0	0	0	0.003	0.0211	0	0	0.0622	0.008	0	0	0.003	0.002	0.012	0	0	0.001	0.008	0	0	0.006	0.001	0	0	0.002	0	0.001	0.8696	0	0
top	0	0	0	0.0044	0.0133	0	0	0.0466	0.0022	0.0044	0	0	0	0.0044	0	0	0	0.0022	0	0	0.0044	0	0	0	0	0	0.0022	0	0.9157	0
vest	0	0	0	0.0049	0.0324	0	0	0.0939	0.0049	0.0032	0	0	0.0032	0.0081	0	0	0	0.0129	0	0	0.0016	0	0	0	0.0065	0	0.0016	0.0016	0	0.8252

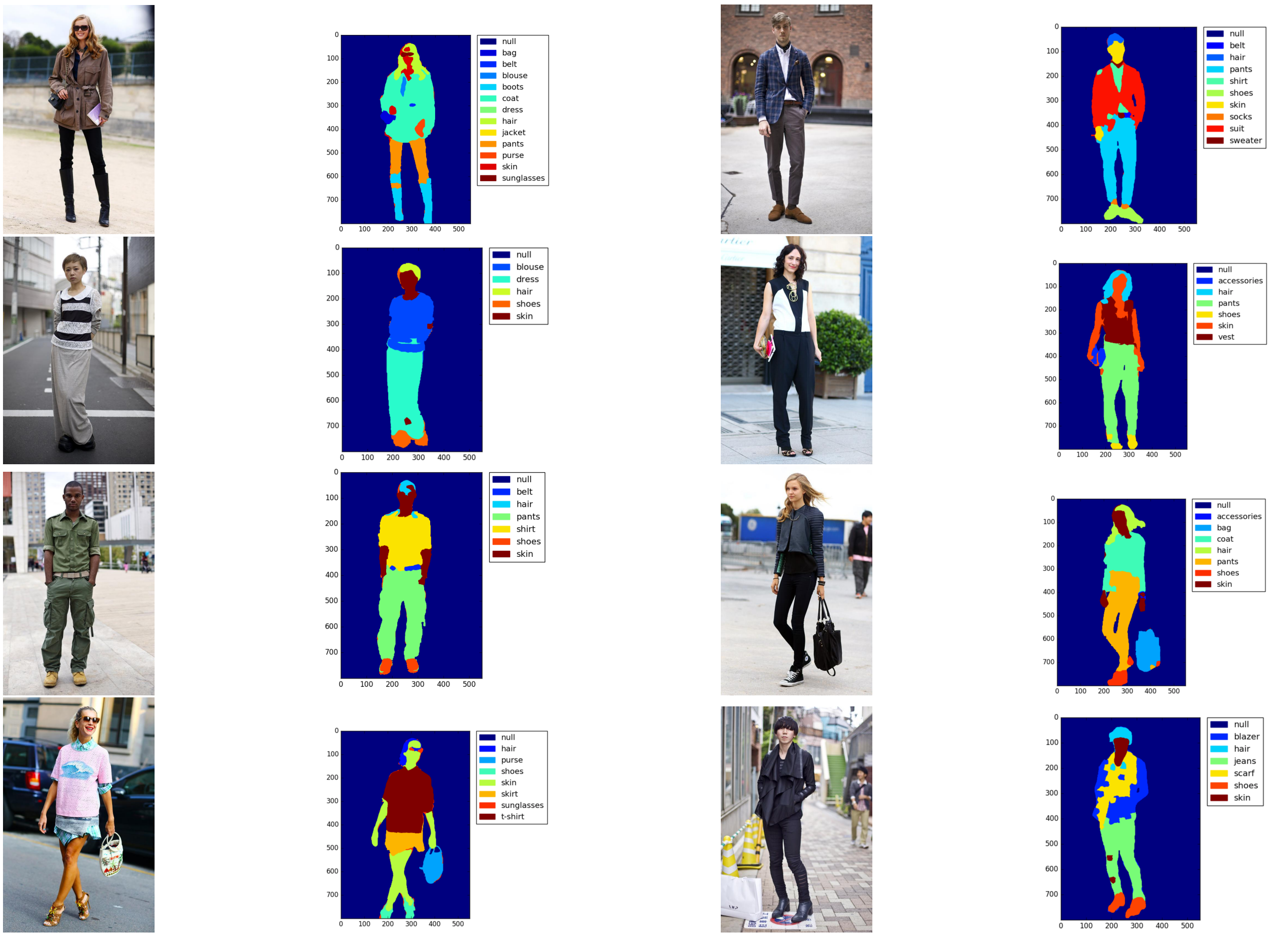


Fig. 2. Some good results.

the apparent clothes, like T-shirts, pants or bags, are correctly segmented, but also the accessory is labeled. There also exist some unsuccessful results, which are shown in Fig. 3. Some part of skin are labeled as clothes, or the clothes are in a mess. We consider that these images fail after *MRF* due to they do not have accuracy in the detection of pose, or the complexity of the texture of clothes. It is also possible that the occlusion and dressing style, e.g. a person in a black shirt and black pants, has caused the failure.

5. DISCUSSION

During the whole project, we have tried varies types of methods, in which the Gabor bank filters and Tamura Texture descriptor dont work well; the human pose estimator, BOW and SIFT work well.

The former ones are not finally used in the project for the same reason, which is that they dont work efficiently. Considering we have a large dataset to deal with, they take a rather longer time to give the result. Moreover, they do not give specific information for the clothes, because the images are in the sophisticated environment.

The later ones perform well during the experiment. The human pose estimator can improve the positional information for the distribution of the image, thus improve the accuracy. We innovatively use BOW in the feature extraction because BOW coding represents independent features, that is to say, the representation of shirts is different than that of pants. And the longer codes, the more discriminative. The SIFT, which is used in the BOW, can extract the local feature of shape. They all give satisfactory result during the repeated experiment.

6. CONCLUSION

We have tried to tackle the challenging clothing parsing problem and got some preliminary results. With the selected features and model, we have shown that the pixel-level accuracy is promising, and we have successfully separated foreground and background. And in some images, the results of segmentation and labeling are satisfactory. However, our model cannot tackle the situations in which images are too complex or ambiguous in colors and the people and surroundings in

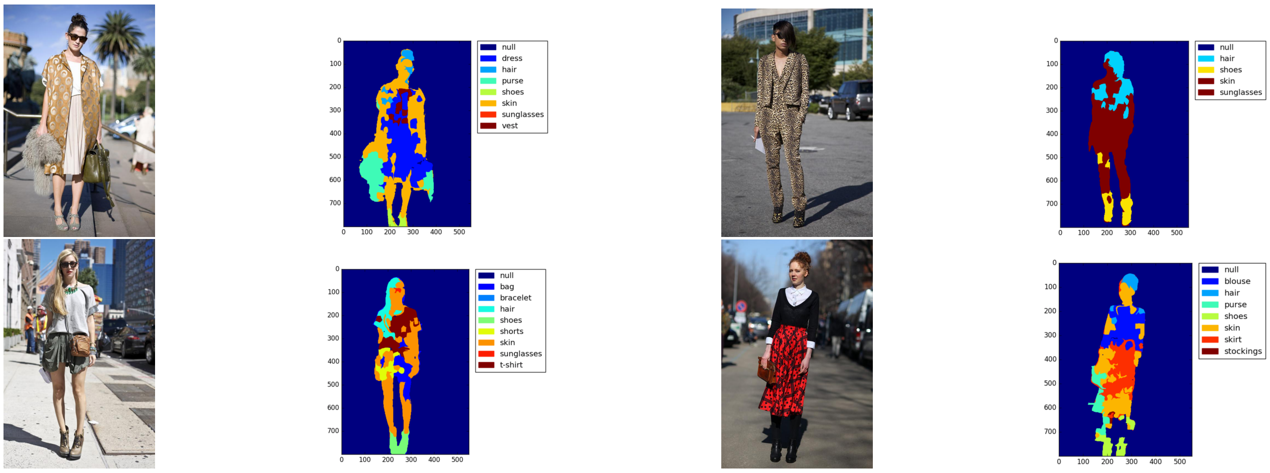


Fig. 3. Some unsuccessful results.

an image cannot be segmented well (the boundaries between them are similar).

7. FUTURE WORK

We mainly have two targets in the next step:

- First, we plan to use a more general unary function, in order to get rid of the limitation of position. Thus to decline the dependence of pose detection.
- Secondly, we would like to improve the function to more complex environment and texture, aiming at increasing the accuracy of segmentation and labeling.

8. AUTHOR ROLES

Daitao Xing is responsible for feature extraction, unary training, the *MRF* model and optimization algorithm. Jun Chen is responsible for the *BOW* model and attempts on bank filter response and Tamura texture descriptors. Jin Huang is responsible for human pose estimators, and improvement on evaluation results.

9. REFERENCES

[1] Wei Yang, Ping Luo, and Liang Lin, "Clothing co-parsing by joint image segmentation and labeling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3182–3189.

[2] Kota Yamaguchi, M Hadi Kiapour, Luis E Ortiz, and Tamara L Berg, "Parsing clothing in fashion photographs," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3570–3577.

[3] Hong Chen, Zi Jian Xu, Zi Qiang Liu, and Song Chun Zhu, "Composite templates for cloth modeling and sketching," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*. IEEE, 2006, vol. 1, pp. 943–950.

[4] Nan Wang and Haizhou Ai, "Who blocks who: Simultaneous clothing segmentation for grouping images," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 1535–1542.

[5] Yihang Bo and Charless C Fowlkes, "Shape-based pedestrian parsing," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 2265–2272.

[6] Agnés Borras, Francesc Tous, Josep Lladós, and Maria Vanrell, "High-level clothes description based on colour-texture and structural features," in *Iberian Conference on Pattern Recognition and Image Analysis*. Springer, 2003, pp. 108–116.

[7] Peng Guan, Oren Freifeld, and Michael J Black, "A 2d human body model dressed in eigen clothing," in *European conference on computer vision*. Springer, 2010, pp. 285–298.

[8] Yi Yang and Deva Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2878–2890, 2013.

[9] Brian Fulkerson, Andrea Vedaldi, Stefano Soatto, et al., "Class segmentation and object localization with superpixel neighborhoods.," in *ICCV*. Citeseer, 2009, vol. 9, pp. 670–677.

[10] Edgar Simo-Serra, Sanja Fidler, Francesc Moreno-Noguer, and Raquel Urtasun, "A high performance crf

model for clothes parsing," in *Asian conference on computer vision*. Springer, 2014, pp. 64–81.