Mix and Match: Joint Model for Clothing and Attribute Recognition

Kota Yamaguchi[1]
http://vision.is.tohoku.ac.jp/~kyamagu
Takayuki Okatani[1]
okatani@vision.is.tohoku.ac.jp
Kyoko Sudo[2]
sudo.kyoko@lab.ntt.co.jp
Kazuhiko Murasaki[2]
murasaki.kazuhiko@lab.ntt.co.jp
Yukinobu Taniguchi[3]
ytaniguti@ms.kagu.tus.ac.jp

[1] Tohoku University
Sendai, Japan
[2] NTT
Yokosuka, Japan
[3] Tokyo University of Science
Tokyo, Japan

Abstract

This paper studies clothing and attribute recognition in the fashion domain. Specifically, in this paper, we turn our attention to the compatibility of clothing items and attributes. For example, people do not wear a skirt and a dress at the same time, yet a jacket and a shirt are a preferred combination. We consider such inter-object or inter-attribute compatibility in the recognition problem, and formulate a Conditional Random Field (CRF) that seeks the most probable combination in the given picture. The model takes into account the location-specific appearance with respect to a human body and the semantic correlation between clothing items and attributes, which we learn using the max-margin framework. We evaluate our model using two datasets that resemble realistic application scenarios: on-line social networks and shopping sites. The empirical evaluation shows that our model effectively improves the recognition performance over baselines including the state-of-the-art feature designed exclusively for clothing recognition. The results also suggest that our model generalizes well to different fashion-related applications.

1 Introduction

Clothing recognition is recently getting more and more attention in the vision community perhaps due to its usefulness in real-world applications, such as on-line social networking [24], e-commerce [3, 4, 8, 16, 21], trend analysis [20], or personal fashion recommender [12, 18]. This paper studies clothing detection, which is an essential component to the above applications or more advanced clothing analysis such as parsing [5, 11, 14, 17, 25], style understanding [9, 10], and attribute recognition [1, 2, 30].

In this paper, we specifically aim at answering if the following hypothesis is correct: *Taking into account multiple clothing items or attributes together improves the detection*

Pages 51.1-51.12
DOI: https://dx.doi.org/10.5244/C.29.51

Figure 1: Some clothing pairs are compatible while others are not. In this paper, we aim to take advantage of such relationship in clothing recognition.

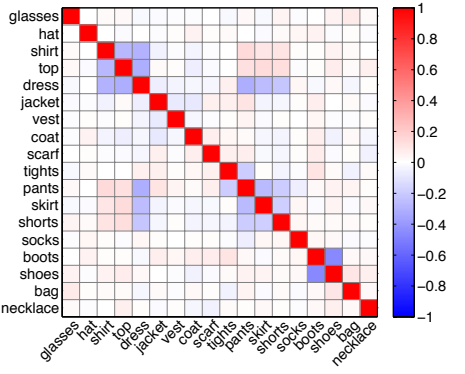


Figure 2: Pearson correlation between clothing items in Chictopia dataset. Notice exclusive blocks, e.g., *shirt*, *top*, and *dress*.

accuracy over detecting individual items independently. The intuition is that people do not wear a skirt and a dress at the same time, and thus there is an exclusive relation between these items. We illustrate such clothing-compatibility in Fig 1. Our detection framework explicitly considers such inter-object or inter-attribute relationship.

Inter-object relationship was previously utilized in clothing parsing as an appearance-level compatibility between adjacent pixels or regions [8, 14, 23, 25]. Our study is distinct from clothing parsing in that we are rather considering the compatibility between clothing items at semantics-level; i.e., people do not wear *dress* and *skirt* together not because they are visually distinct, but because of their functionality.

Our approach is the second-order joint model based on the Conditional Random Field (CRF) over the combination of clothing items or attributes. Given an image, we consider the probability distribution over clothing items, and output the *maximum a posteriori* (MAP) assignment as a detection result. In the unary term, we also take advantage of strong contextual relationship between the location of clothing and human body. Correlation between items are explicitly modeled in the second-order term in our model (See Fig 2). We empirically show that our approach is outperforming the independent baseline including the approach based on the state-of-the-art feature for clothing recognition [25].

Our model is similar to part-aligned attribute detectors [27, 28, 29] in that we take advantage of pose-aligned feature to recognize detailed attributes. However, our main focus in this paper is the joint modeling of inter-object or inter-attribute relationships. Our model has the same spirit with Wang's work [22] in that both aim at jointly modeling inter-label relationships, but this paper focuses more on empirical studies in realistic scenarios with real-world data. Also, by taking advantage of relatively stable pose-variation in fashion pictures, we apply a simple-yet-effective deterministic approach to compute localized image-features based on Convolutional Neural Networks [7].

To study clothing detection in a realistic scenario, we use two datasets each with different application in mind: Chictopia dataset [24] for clothing detection in fashion blogs and Dress dataset for attribute recognition. The successful empirical results suggest that our model generalizes well to two different fashion applications.

(a) Chictopia dataset (b) Dress dataset

Figure 3: Our datasets consider two scenarios; a) Chictopia dataset considers automatic clothing tagging in fashion blogs, and b) Dress dataset considers attribute recognition in e-commerce.

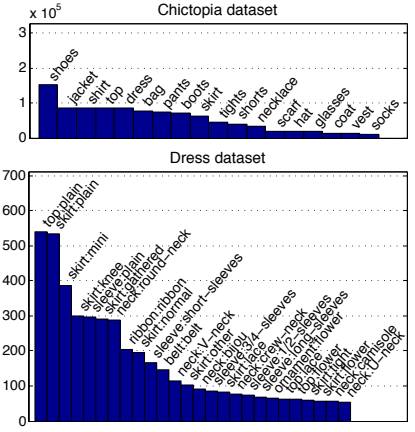


Figure 4: Label distribution.

We summarize our contribution in the following.

- CRF-based detection that takes into account inter-label correlation
- Simple yet effective deterministic approach to localize region of interests in fashion problems
- Empirical evaluation using Chictopia and Dress datasets that confirms the generalizability of the proposed model in different fashion-recognition problems

2 Dataset

Chictopia dataset Chictopia dataset is a collection of blog posts from Chictopia, an online social networking site specialized for fashion. Using the publicly available data [24], we first applied a human-pose detector [25, 26] and only kept images with a standing person. The detected bounding-boxes around the body are later used to compute image features. Next, we searched images with at least 2 clothing-keywords under 18 categories (See Fig 7) in metadata, and identified 26,8124 usable images. Fig 3(a) shows an example picture.

Note that clothing tags are noisy and far from perfect. It is common to observe missing items, especially for minor items such as *necklace* or *socks*. Also, sometimes there are conflicting items appearing together, such as *shoes* and *boots*, due to the tagging-format in Chictopia where users can associate clothing keywords together with a free-form text. However, we did not apply any manual transformation to these noisy data for scalability in the realistic scenario. Tag statistics are shown in Fig 4.

We also use the publicly available Fashionista dataset [23] to learn the human-body detector and the spatial prior about clothing, which is also collected from Chictopia.

Dress dataset Dress dataset consists of 712 images of dress products we collected on an e-commerce site. For each of the dress images, we manually gave bounding-box annotations for 8 parts of dress (*top*, *skirt*, *neck*, *sleeve*, *ornament*, *ribbon*, *belt*, and *pocket*). Only *top* and *skirt* appear at every picture, and other parts might not be present. We manually annotated

each part with detailed binary attributes, such as *round-neck* or *V-neck* for neck part. In the initial annotation, we had in total 58 attributes. Out of 58 attributes, we chose 26 for evaluation and removed infrequent attributes that occurred less than 4% in the dataset.

The bounding-box annotations are used to learn our localized image feature we discuss in Sec 4. Fig 3(b) shows an example picture from Dress dataset. In Dress dataset, the product image does not always contain a person, and sometimes both frontal and rear views appear side-by-side. We first apply a human-body detector trained on Fashionista dataset and calculate features from only one of the views. Although the bounding box does not perfectly align when a person is not present, we could obtain reasonable bounding boxes around the dress region[1].

3 Joint detection model

Let us denote a set of labels by $Y \equiv \{y_i\}$, $y_i \in \{0,1\}$, where i is one of the clothing items or attributes, such as *shirt* or *skirt:plain*. Given a feature $X \equiv \{\mathbf{x}_i\}$, we define our joint probability distribution over labels by a log-linear model:

$$\ln P(Y|X) \equiv \sum_i w_i \phi(\mathbf{x}_i, y_i) + \sum_{i,j \in V} w_{i,j} \psi(y_i, y_j) - \ln Z, \tag{1}$$

where we denote the model parameters by $\mathbf{w} \equiv [w_i, w_{i,j}], \forall i, j$, the normalization constant by Z, and the set of label-pairs by V. Our model consists of the unary term $\phi(\mathbf{x}_i, y_i)$ that considers the likelihood of assigning a label given a feature, and the binary term $\psi(y_i, y_j)$ that considers the inter-label relationships.

3.1 Data likelihood

In this paper, we use logistic regression of each label, expressed by:

$$\phi(\mathbf{x}_i, y_i) \equiv \ln p(y_i|\mathbf{x}_i), \tag{2}$$

$$p(y_i = 1|\mathbf{x}_i) \equiv \sigma(\mathbf{a}_i^{\mathrm{T}} \mathbf{x}_i + b_i), \tag{3}$$

where $\mathbf{a}_i$ and b_i are the regression parameters for each item. We learn the logistic regression [6] from the training examples. The unary term can be thought of a regular appearance-based detector. Our joint model augments the prediction of the unary term by inter-label correlation.

Note that it is possible to directly use $\mathbf{x}_i$ for the potential term in Eq 1. We did not choose to do so to include additional non-linearity in the model and also to make learning computationally tractable.

3.2 Inter-label correlation

We use the normalized Pearson correlation for the binary term:

$$\psi(y_i, y_j) \equiv \begin{cases} \ln \frac{1}{2}(1 + c_{i,j}), & \text{if } y_i = y_j \\ \ln \frac{1}{2}(1 - c_{i,j}), & \text{otherwise} \end{cases} \tag{4}$$

[1] We have also tried to learn a dress detector based on region proposals, but we could not observe any better result than the human-body detector perhaps due to the lack of training data. In this paper, we chose to use a human-body detector.

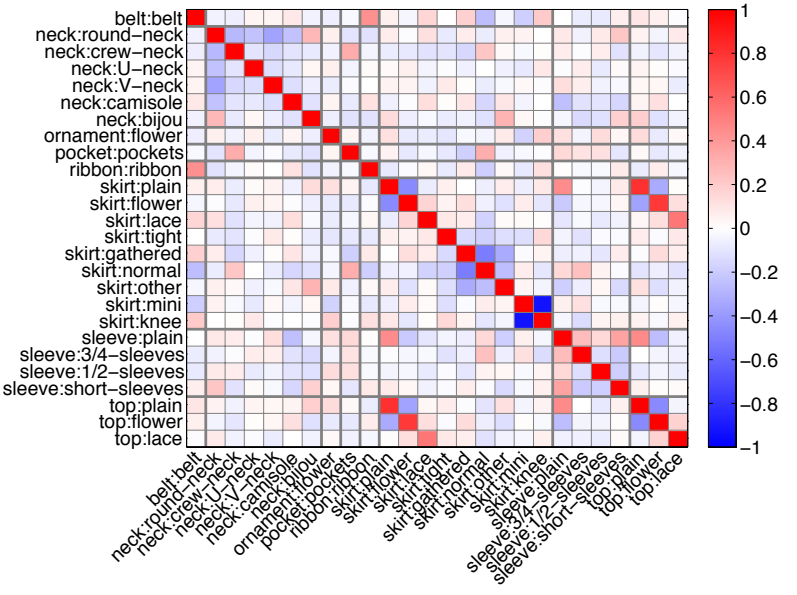


Figure 5: Pearson correlation between dress attributes in Dress dataset. Some attributes appear across parts (e.g., plain pattern) and give a strong positive correlation. Notice negative correlation within a part-block indicating some attributes are exclusive each other (e.g., neck shapes).

where $c_{i,j}$ is the Pearson correlation between item i and j in the training examples.

The binary term enforces the inter-label correlation in the assignment. We visualize in Fig 2 and 5 the correlation between annotations in Chictopia and Dress datasets. In Chictopia, clearly we can observe exclusiveness in upper-body (*shirt*, *top*, *dress*), lower-body (*dress*, *pants*, *skirt*, *shorts*), or footwear (*boots*, *shoes*). Also, the positive correlation between tops and bottoms indicates they are likely worn together. In Dress, we can observe exclusive groups (e.g, neck shapes, skirt length). Some attributes are positively correlated (e.g., top and skirt have the same pattern, flower and lace are likely to appear together). Our joint model encompasses such second-order information.

3.3 MAP inference

Given a feature X, we can detect clothing items by MAP inference over the joint model:

$$Y^* \in \arg\max_Y P(Y|X). \tag{5}$$

We use the loopy belief propagation [15] to approximately solve Eq 5.

3.4 Max-margin learning

We learn the model parameters $\mathbf{w}$ with the Structural SVM framework [19]. Let us denote the concatenation of potential functions by $\Psi(X,Y)$ so that Eq 1 is expressed in the linear

form: $\ln P(Y|X) = \mathbf{w}^{\mathrm{T}}\Psi(X,Y) - \ln Z$. Using the margin-rescaling formulation, our learning problem can be expressed by the following optimization:

$$\begin{aligned} &\min_{\mathbf{w},\xi} \quad \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{N}\sum_{k}^{N}\xi_k \\ &\text{s.t.} \\ &\forall k, \forall Y, \quad \mathbf{w}^{\mathrm{T}}\left(\Psi(X_k,Y_k) - \Psi(X_k,Y)\right) \geq \Delta(Y_k,Y) - \xi_k, \end{aligned} \tag{6}$$

where we denote the slack variables by $\xi \equiv \{\xi_k\}$, the loss function by $\Delta(Y_k,Y)$, and the number of training examples by N. C is the constant parameter to balance between the regularization and the loss term. The intuition of this objective is that we constrain $\mathbf{w}$ such that for any training example k, the true assignment Y_k produces the maximum log-linear score with a margin against any other incorrect assignment Y. We solve Eq 6 using the cutting-plane algorithm for a general loss [19].

In this paper, we propose to use the class-weighted zero-one loss:

$$\Delta(Y_k,Y) \equiv \frac{1}{2}\sum_i \delta_k(y_{k,i},y_i), \tag{7}$$

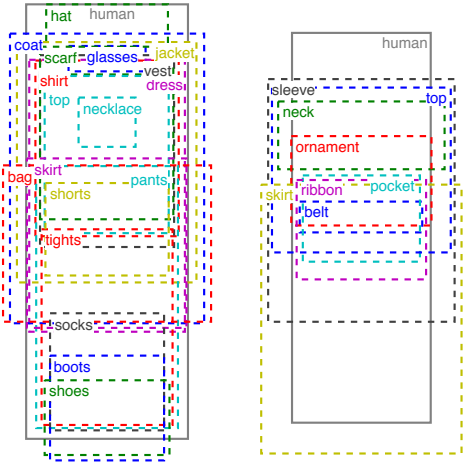
$$\delta_k(y_{k,i},y_i) \equiv \begin{cases} \frac{N}{N-N_i}, & \text{if } y_{k,i}=0, y_i=1 \\ \frac{N}{N_i}, & \text{if } y_{k,i}=1, y_i=0 \\ 0, & \text{otherwise,} \end{cases} \tag{8}$$

where N_i is the number of positive examples for label i in the training set. The purpose is to penalize more for making an incorrect prediction to infrequent label i. Note that both of our datasets form a long-tailed distribution; The majority of labels does not appear frequently, while the rest appears more often (Fig 4). Without considering the class balance in the loss, the learned model tends to find a trivial parameter that always assigns 0 to rare class, such as *hat* or *vest*.

4 Localized image feature

Most of the existing work in clothing recognition requires an accurate pose estimation beforehand [13, 23, 25], because clothing items are worn on a specific body part. However, that approach has a drawback in that the failure in part localization can easily lead to incorrect recognition results. In this paper, instead of relying on the pose estimation of every body-part, we relax this localization requirement to only a bounding-box around the human-body, and use rather a deterministic approach to define a region of interest. Given a bounding-box around the full human body, we extract an image patch from a specific location relative to the body bounding-box. This simple approach yields a surprisingly good result as we show in the experiment.

We learn the relative locations based on training data. We show the relative location of bounding-boxes used for Chictopia and Dress in Fig 6. For Chictopia, we first calculate the average tight bounding-box from pixel-wise annotations in the Fashionista dataset, make the boxes symmetric, and enlarge the box size by 40%. We apply the same procedure for Dress dataset except that there is no pixel-wise annotation there.



(a) Chictopia clothing (b) Dress attributes

Figure 6: Relative location of part bounding boxes.

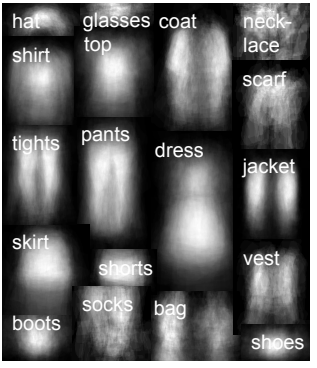


Figure 7: Average shape of each item within the relative bounding-box in Fashionista dataset [23]. Our deterministic localization covers sufficient regions of interest.

The major concern for this approach is the recall under pose-variance. However, we observed from data that in fashion pictures people do not make a significant pose-change and we are able to extract sufficient coverage of the part-regions by this simple deterministic localization. Using pixel-wise annotation in Fashionista dataset, we show the average shape of clothing of Chictopia at the relative locations in Fig 7, which confirms that our simple localization sufficiently covers the appearance of target item. A by-product of our deterministic approach is that we can localize regions in a constant time.

Once the region of interest is determined, we extract the CNN feature (`fc7` of AlexNet, 4096 dimensions) learned from ImageNet [7], and use as a feature-input to the logistic regression in Eq 2.

5 Experimental results

We compare the following methods in the experiments.

Style Descriptor: The hand-crafted image-feature based on pose estimation proposed in the state-of-the-art clothing parsing [25]. We learn logistic regression for prediction.

CNN Global: We predict labels by logistic regression from the CNN feature calculated from the full-body bounding box without part localization.

CNN Local: We predict by logistic regression using the localized CNN feature. This is equivalent to removing the second-order term in Eq 1.

CNN Local CRF: Our joint detection model described in Eq 1.

The comparison to Style Descriptor constitutes the relative performance of our CNN-based detection over the state-of-the-art. We can measure how much the localized image feature or the joint model improves detection performance from the comparison between the CNN Global, the CNN Local, and the CRF models.

Our experimental protocol is based on 10-fold cross validation. From Chictopia dataset, we randomly sample 9,000 training examples and 1,000 testing examples each with at least 2 clothing tags. From Dress dataset, we make a 90% train / 10% test split. We measure

Table 1: Total performance evaluation.

(a) Clothing-detection in Chictopia dataset

Method	Accuracy	Precision	Recall	F1
Style-Descriptor	0.690 ± 0.002	0.345 ± 0.003	**0.646** ± 0.009	0.450 ± 0.004
CNN-Global	0.740 ± 0.003	0.390 ± 0.004	0.573 ± 0.009	0.464 ± 0.005
CNN-Local	0.768 ± 0.004	0.436 ± 0.006	0.622 ± 0.007	0.512 ± 0.005
CNN-Local-CRF	**0.782** ± 0.003	**0.456** ± 0.005	0.595 ± 0.008	**0.516** ± 0.004

(b) Attribute-prediction in Dress dataset

Method	Accuracy	Precision	Recall	F1
Style-Descriptor	0.781 ± 0.015	0.526 ± 0.035	0.661 ± 0.021	0.585 ± 0.028
CNN-Global	0.837 ± 0.006	0.632 ± 0.021	0.723 ± 0.013	0.674 ± 0.015
CNN-Local	0.840 ± 0.007	0.638 ± 0.020	**0.727** ± 0.020	**0.680** ± 0.018
CNN-Local-CRF	**0.843** ± 0.007	**0.652** ± 0.023	0.708 ± 0.013	0.678 ± 0.016

the recognition performance in terms of accuracy, precision, recall, and F1. We repeat this procedure for 10 times and report the average with standard deviation.

The performance is summarized in Table 1. We also show the precision of individual items in Fig 8. We first observe that the CNN-based feature is outperforming the Style Descriptor designed for clothing recognition.

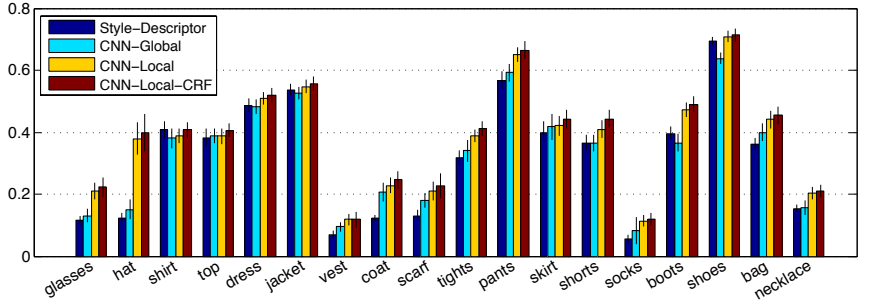
Using the localized image feature (CNN Local), we further boost the performance over the full-body feature (CNN Global). This effect is apparent in smaller items appearing at the end of body, such as *glasses*, *hat*, or *boots* in Chictopia. Note that the Style Descriptor is performing better than CNN Global for footwear because of its dependency to precise human-pose estimation. However, our CNN-based feature beats this baseline with a simple localization approach. Compared to Chictopia, we have observed less significant improvement in Dress dataset, perhaps due to the lack of training examples (e.g., *neck:camisole*).

Our CRF model makes improvement over CNN Local in accuracy, precision, with a small loss in recall. This can be explained by the effect of inter-label correlation in our model successfully *s*uppressing conflicting prediction (e.g., *tights* and *pants*, *round-neck* and *v-neck*) made in the independent model (CNN Local). Precision is typically more valuable than recall in the detection problem, since we can easily improve recall by predicting everything as positive. Our model successfully augments the independent prediction by CNN-based features towards the higher-precision regime.

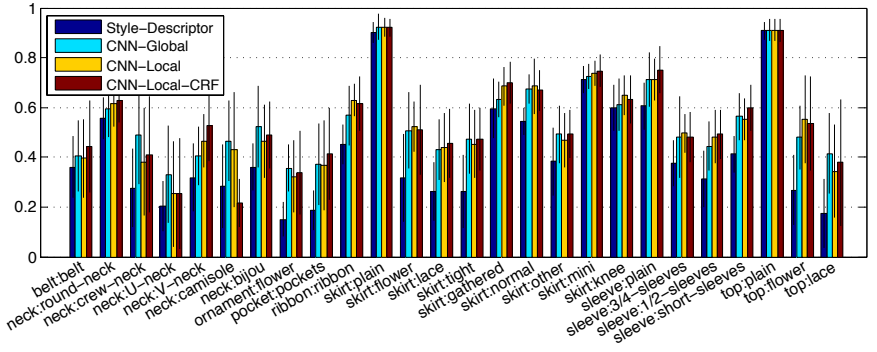
Fig 5 shows a few qualitative results. There is often noise in the annotation (e.g., *jacket* and *vest* in the second row). Detecting by the global feature tends to predict false positives for small items (e.g., *hat*). Our CRF model produces mostly similar results to CNN Local, except that the model lowers the detection confidence of incompatible labels (e.g., *top* vs. *dress*, *short-sleeves* vs *1/2-sleeves*).

Discussion Our joint model successfully improves the precision of predictions, but the drawback is that we are more likely to miss less frequent combination, such as a layered combination of *shirt* and *top*. Perhaps for an application where higher recall is important, independent prediction of items would be still useful since joint prediction would suppress such cases in the long tail. Predicting a combination in the long tail is inherently a difficult problem.

In this work, we did not fine-tune the CNN feature mainly due to the noise in data (Chictopia dataset) and the lack of data size (Dress dataset). CNN is known to perform excellent



(a) Clothing-detection precision in Chictopia dataset



(b) Attribute-prediction precision in Dress dataset

Figure 8: Precision for individual items or attributes.

when there is a high-quality, supervised dataset, but such dataset does not exist in a new application problem such as fashion. In that sense, this data-bottleneck issue is always a challenge for applying deep models in a new domain.

6 Conclusion and future work

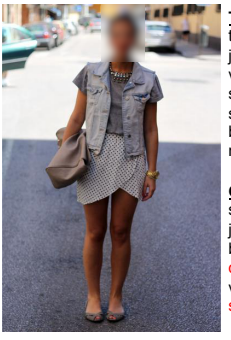
We proposed a joint clothing detection model that considers inter-label correlation of items. The model also takes advantage of the spatial prior of clothing with respect to human-body. The empirical study using the two realistic datasets reveals that our model performs the best among the baseline approaches including the state-of-the-art feature. Also, our model successfully augments independent predictions by logistic regression to higher-precision regime, using second-order relationships between labels.

Though we specifically studied the clothing detection in the paper, our CRF framework can be applied to other category-specific applications such as product catalogs in e-commerce sites or in used-car markets. We hope to extend our work to different applications.

Also, it is our future work to study the precise effect of pose variation in fashion applications, and if any, to improve our model [28]. We would like to see how an optimized CNN architecture [11] trained from the large amount of fashion images affects the clothing and attribute detection performance.

Truth	Style-Desc	CNN-Global
glasses	bag	dress
dress	dress	shoes
shoes	scarf	vest
bag	shoes	necklace
necklace	glasses	scarf
	hat	jacket
	vest	
	coat	

CNN-Local	CNN-Local-CRF
dress	dress
shoes	shoes
bag	bag
glasses	glasses
scarf	scarf
necklace	necklace
top	



Truth	Style-Desc	CNN-Global
top	shoes	hat
jacket	bag	dress
vest	shorts	bag
skirt		shorts
shoes		vest
bag		shoes
necklace		glasses

CNN-Local	CNN-Local-CRF
shoes	shoes
jacket	vest
bag	jacket
dress	bag
vest	dress
shirt	

(a) Clothing detection

Truth	Style-Descriptor	CNN-Global	CNN-Local	CNN-Local-CRF
neck:V-neck	top:plain	sleeve:plain	skirt:normal	top:plain
ornament:flower	skirt:mini	skirt:normal	sleeve:plain	skirt:normal
skirt:plain	skirt:flower	top:plain	top:plain	skirt:plain
skirt:normal	skirt:plain	skirt:plain	skirt:plain	sleeve:plain
skirt:knee	top:lace	neck:V-neck	neck:V-neck	neck:V-neck
sleeve:plain	top:flower	pocket:pockets	ornament:flower	ornament:flower
top:plain	skirt:other	skirt:knee	top:lace	

Truth	Style-Descriptor	CNN-Global	CNN-Local	CNN-Local-CRF
neck:round-neck	skirt:plain	sleeve:plain	sleeve:plain	sleeve:plain
ornament:flower	top:plain	sleeve:short-sleeves	skirt:plain	skirt:normal
skirt:plain	neck:round-neck	neck:round-neck	skirt:normal	skirt:plain
skirt:normal	skirt:mini	top:plain	sleeve:short-sleeves	sleeve:short-sleeves
skirt:mini	ornament:flower	skirt:plain	skirt:knee	skirt:knee
sleeve:plain	sleeve:plain	skirt:normal	sleeve:1/2-sleeves	top:plain
sleeve:short	neck:V-neck	skirt:knee	ornament:flower	neck:round-neck
top:plain	sleeve:short-sleeves	sleeve:1/2-sleeves	neck:round-neck	
	skirt:gathered	neck:U-neck		
	neck:bijou	skirt:mini		

(b) Attribute prediction

Figure 9: Qualitative results. False positives are marked red. Items are ordered by detection confidence. The CRF model sorts items by marginal probability.

References

[1] Lukas Bossard, Matthias Dantone, Christian Leistner, Christian Wengert, Till Quack, and Luc Van Gool. Apparel classification with style. *ACCV*, pages 1–14, 2012.

[2] Huizhong Chen, Andrew Gallagher, and Bernd Girod. Describing clothing by semantic attributes. In *ECCV*, pages 609–623. 2012.

[3] George A Cushen and Mark S Nixon. Mobile visual clothing search. In *ICME Workshops*, pages 1–6. IEEE, 2013.

[4] Wei Di, Catherine Wah, Anurag Bhardwaj, Robinson Piramuthu, and Neel Sundaresan. Style finder: Fine-grained clothing style detection and retrieval. In *CVPR Workshops*, pages 8–13, 2013.

[5] Jian Dong, Qiang Chen, Wei Xia, Zhongyang Huang, and Shuicheng Yan. A deformable mixture parsing model with parselets. *ICCV*, 2013.

[6] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *J Machine Learning Research*, 9:1871–1874, 2008.

[7] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM Multimedia*, pages 675–678, 2014.

[8] Yannis Kalantidis, Lyndon Kennedy, and Li-Jia Li. Getting the look: clothing recognition and segmentation for automatic product suggestions in everyday photos. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, pages 105–112. ACM, 2013.

[9] M. Hadi Kiapour, Kota Yamaguchi, Alexander C. Berg, and Tamara L. Berg. Hipster wars: Discovering elements of fashion styles. *ECCV*, 2014.

[10] Iljung S Kwak, Ana C Murillo, Peter N Belhumeur, David Kriegman, and Serge Belongie. From bikers to surfers: Visual recognition of urban tribes. *BMVC*, 2013.

[11] Xiaodan Liang, Si Liu, Xiaohui Shen, Jianchao Yang, Luoqi Liu, Liang Lin, and Shuicheng Yan. Deep human parsing with active template regression. *TPAMI*, 2015.

[12] Si Liu, Jiashi Feng, Zheng Song, Tianzhu Zhang, Hanqing Lu, Changsheng Xu, and Shuicheng Yan. Hi, magic closet, tell me what to wear! In *ACM Multimedia*, pages 619–628, 2012.

[13] Si Liu, Zheng Song, Guangcan Liu, Changsheng Xu, Hanqing Lu, and Shuicheng Yan. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *CVPR*, pages 3330–3337, 2012.

[14] Si Liu, Jiashi Feng, Csaba Domokos, Hui Xu, Junshi Huang, Zhenzhen Hu, and Shuicheng Yan. Fashion parsing with weak color-category labels. *IEEE Transactions on Multimedia*, 16(1), January 2014.

[15] Joris M. Mooij. libDAI: A free and open source C++ library for discrete approximate inference in graphical models. *J Machine Learning Research*, 11:2169–2173, August 2010.

[16] Rasmus Rothe, Marko Ristin, Matthias Dantone, and Luc Van Gool. Discriminative learning of apparel features. *Machine Vision Applications*, 2015.

[17] Edgar Simo-Serra, Sanja Fidler, Francesc Moreno-Noguer, and Raquel Urtasun. A high performance crf model for clothes parsing. *ACCV*, 2014.

[18] Edgar Simo-Serra, Sanja Fidler, Francesc Moreno-Noguer, and Raquel Urtasun. Neuroaesthetics in fashion: Modeling the perception of fashionability. *CVPR*, 2015.

[19] Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. Large margin methods for structured and interdependent output variables. In *J Machine Learning Research*, pages 1453–1484, 2005.

[20] Sirion Vittayakorn, Kota Yamaguchi, Alexander C Berg, and Tamara L Berg. Runway to realway: Visual analysis of fashion. *WACV*, 2015.

[21] Xianwang Wang and Tong Zhang. Clothes search in consumer photos via color matching and attribute learning. *ACM Multimedia*, 2011.

[22] Yang Wang and Greg Mori. A discriminative latent model of object classes and attributes. *ECCV*, 2010.

[23] Kota Yamaguchi, M Hadi Kiapour, Luis E Ortiz, and Tamara L Berg. Parsing clothing in fashion photographs. In *CVPR*, pages 3570–3577, 2012.

[24] Kota Yamaguchi, Tamara L Berg, and Luis E Ortiz. Chic or social: Visual popularity analysis in online fashion networks. *ACM Multimedia*, 2014.

[25] Kota Yamaguchi, M Hadi Kiapour, Luis E Ortiz, and Tamara L Berg. Retrieving similar styles to parse clothing. *TPAMI*, 2014.

[26] Yi Yang and Deva Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, pages 1385–1392, 2011.

[27] Ning Zhang, Ryan Farrell, Forrest Iandola, and Trevor Darrell. Deformable part descriptors for fine-grained recognition and attribute prediction. In *ICCV*, pages 729–736, 2013.

[28] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell. Part-based r-cnns for fine-grained category detection. In *ECCV*, pages 834–849. 2014.

[29] Ning Zhang, Manohar Paluri, Marc'Aurelio Ranzato, Trevor Darrell, and Lubomir Bourdev. Panda: Pose aligned networks for deep attribute modeling. In *CVPR*, pages 1637–1644. IEEE, 2014.

[30] Weipeng Zhang, Jie Shen, Guangcan Liu, and Yong Yu. A latent clothing attribute approach for human pose estimation. *ACCV*, 2014.