Attributes-oriented Clothing Description and Retrieval with Multi-task Convolutional Neural Network

Yizhang Xia*† , Baitong Chen†, Wenjin Lu†, Frans Coenen* and Bailing Zhang†
*Department of Computer Science, University of Liverpool, Liverpool, United Kingdom
†Department of Computer Science and Software Engineering, Xi'an Jiaotong-Liverpool University, Suzhou, China

***Abstract*—This paper seek answer to question how to search clothing when consumer pays attention to a part of clothing. A novel framework is proposed to solve above problem by attributes. First of all, Fast-RCNN detects person from complex background. Then a Convolutional Neural Network (CNN) is combined with Multi-Task Learning (MTL) to extract features related to attributes. Next Principal Component Analysis (PCA) reduce dimensionality of feature from CNN. Finally, Locality Sensitive Hashing (LSH) searches similar samples in the gallery. Extensive experiments were done on the clothing attribute dataset, experimental results proves this framework is effective.**

***Index Terms*—Fashion Description, Fashion Retrieval, Multi-Task Learning (MTL), Convolutional Neural Network (CNN).**

I. Introduction

Clothing image retrieval is a frequent requirement for online shopping. For example, consumers may want to find their favorite clothing from online photos. However, due to the geometric variations in the shape of the clothes, the different lighting and the complex scene, clothing image retrieval is very difficult.

This paper study how to search clothing when consumers are keen on a part of clothing. For example, in Fig. 1 consumers like clothing that resemble the left T-shirt neckline type, but do not care about its color. The clothing on the right in Fig. 1 is the search result they want.

Many researchers have been trying to solve these problems. [9] An earlier practice is to combine a variety of hand-designed features, such as PHoG, HoW, VSSIM, retrieve clothing from phone [1]. The bottleneck of this method is hand-designed features. Then, Grana [2] realized above problem and introduced Bag of Words (BoW) to improve the clothing retrieval performance. However, his improvement is not obvious. Next, Wang [3] first introduced attribute into clothes retrieval. Wang [3] used attributes to rearrange the retrieval results that have been arranged by color code-book construction and obtained higher performance. Then, Di [4] and Chen [5] tried to employ fine-grained visual classification for clothing retrieval. Di [4] combined hand-designed features and BoW. On the other hand, Chen [5] employed Convolutional Neural Network (CNN). At the same time, Chen [5] pointed out that attribute-based clothing retrieval is one of application. Next, Kiapour [6] presented a more complex question how to match street clothing photo in an online store. He also used attributes to retrieve clothing. [6] Meanwhile, Huang [7] proposed a dual attribute-aware ranking network for retrieving similar clothing from an online shop, with conclusion that CNN outperforms hand-designed features. [7] Subsequently, Xiong [8] discussed how to share features between street photos and online shopping photos in CNN, and proposed partial-sharing CNN architecture. In a most recent study, Liu [9] published the DeepFashion dataset and considered multiple tasks in CNN through Multi-Tasking Learning (MTL).

Fig. 1. Partial attribute-based clothing retrieval. Customer likes the category, the pattern and the neckline, but is indifferent to the color in the left photo. The right clothing are the desired retrieval result.

In short, CNN and attributes are the mainstream in clothing retrieval at present. This paper employs MTL [11] with CNN and builds an appropriate attributes tree for clothing attributes. Then, a customer's favorite attribute-related features are connected for partial attribute clothing retrieval.

The remaining of the paper is constructed as follows. In section II, whole system is overviewed, and person detection, attribute-based feature extraction and similarity measurement are explained in details, respectively. Experiments and results are illustrated in section III. Following conclusion is given in the section IV.

II. System Overview

The whole system is illustrated as Fig. 2. Firstly, pedestrian detection crops person from complex background photo. Then CNN predicts attributes and extracts attributes-related features. Finally, a LSH searches the first N most similar photos.

A. Pedestrian Detection

Huang [7] verified that clothing detection can improve clothing retrieval performance. Following the practices, this paper employs Fast RCNN [10] to detect person from complex background photos, Fig. 3.

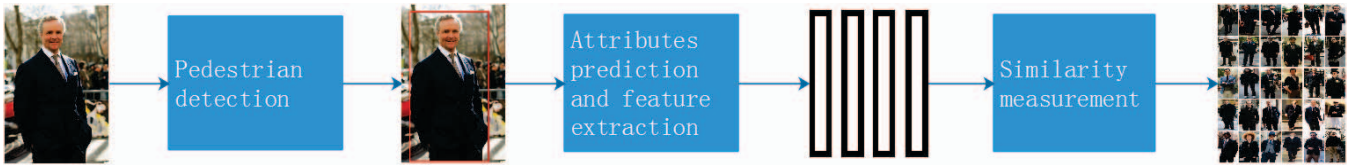

Fig. 2. System overview. All of blue rectangles are processors.

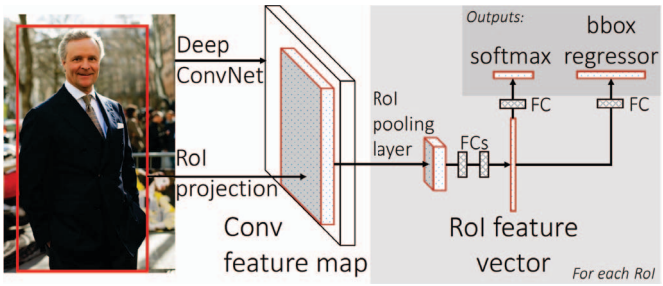

Fig. 3. Fast RCNN [10].

Fast RCNN [10] is a real-time object detection framework as it uses Spatial Pyramid Pooling (SPP) layer [13] and MTL. SPP layer avoids calculating the repetitive area of input image several times. And MTL avoids saving CNN features into hard disk for object location task since it prompt object classification task and object location task to share CNN feature. Though its accuracy is not higher than RCNN [14], this paper selects fast RCNN [10] since it is much quicker than RCNN [14].

B. Attribute-based Feature Extraction

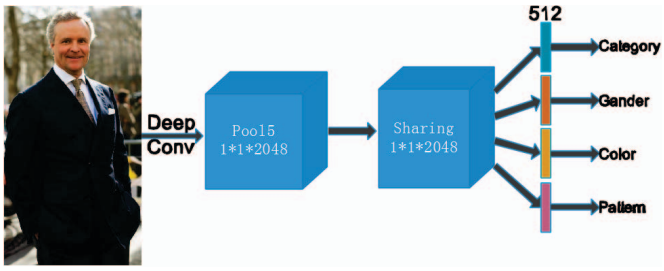

Fig. 4. Multi-task CNN.

This paper combines CNN and MTL to predict clothing attributes and extract attributes-related features, Fig. 4. In order to minimize intra-class distance and maximize inter-class distance, we created an attribute tree. All of the clothing attributes are divided into four groups, namely, category, gender, color and pattern. The priori knowledge is introduced into this attribute tree with the relationship among the attributes. Some attributes are not easy to describe for customer. For example, customers can not describe the color on a clothing, but they definitely understand what they like is a kind of color. In this situation, the feature from upper branch of all color can be used to retrieve clothing, and the experiments are analysed in section III-C.

In the Fig. 4, a sharing layer follows the antepenultimate layer of pre-train CNN model. It fuses all of attributes' information by convolution. Then, sharing layer is bifurcated into four branches. In the feed-forward process, the sharing layer assigns the copy to each branch, as given in Eqn 1. In the back Propagation process, sharing layer accumulates each branch's gradient, Eqn 2.

$$InB_i = Out_S \tag{1}$$

$$G_S = \sum_{i=1}^{4} GB_i \tag{2}$$

In Eqn 1 and Eqn 2, InB_i and Out_S are the feed forward inputs of branch i and output of sharing layer. G_S and GB_i are gradient of sharing layer and branch i in back propagation.

C. Similarity Measurement

The similarity measurement between clothing images is implemented by Principal Component Analysis (PCA) and Locality Sensitive Hashing (LSH) [15]. PCA reduces the dimension of feature, maintaining the largest variance.

Similarity measurement is an important issue in computer vision. Abundant feature information is usually represented by high dimensional vectors. An ideal similarity measurement generally needs to satisfy the following four conditions: high accuracy, low spatial complexity, low time complexity and supporting high dimensionality. LSH hashes similar objects into the same bucket. That is, LSH can classify and cluster high dimensional data, such as image. According to the characteristics of LSH, it is used to retrieve similar clothing from train set for test set.

III. Experiment

The proposed method is evaluated on the Clothing Attributes Dataset [12], section III-A. Attributes prediction is showed in section III-B, and attributes-based clothing retrieval is displayed in section III-C. The program is based on MatConvNet [16]. And the important training parameters of CNN follows from [17], i.e., stochastic gradient decent (SGD) with a batch size of 16 examples and learning rate of 0.00001. Each epoch of training takes about 15 minutes on an NVIDIA TITAN GPU, and the network around converges in 50 epoches.

Fig. 5. Examples of the Clothing Attribute Dataset.

A. Clothing Attribute Dataset

The Clothing Attributes Dataset [12] was built by Chen and collected images from Sartorialist 1 and Flickr. It contains 1856 RGB images that include clothed people. Samples are shown in Fig. 5. Then, six workers from Amazon Mechanical Turk (AMT) were employed to annotate ground truth of twenty six clothing attributes. Labels that have 5 or more consents are determined as the true label, otherwise, it is annotated with unknown.

B. Clothing Attributes Classification

As Section II-B explained, multi-task CNN is used to predict clothing attributes. In the Fig. 4, $Pool5$ layer is followed by sharing layer with $1 \times 1 \times 2048$ filter size. All attributes are then divided into four groups, category, gender, color and pattern.Each branch includes a convolution layer, $1 \times 1 \times 512$, a batch normalization layer and a ReLu layer. Finally, each branch will be bifurcated into single attribute task that includes one full connection layer and one soft-max layer.

The neuron number of each branch, 512, is a empirical parameter. At the same time, the L2 regularization item is employed in our experiments. these two tricks guarantee abundant feature for subsequent classification and retrieval.

The clothing attributes prediction is evaluated on the Clothing Attributes Dataset [12] and the accuracy is listed in the Table I. The dataset is randomly divided into three parts, training set, validating set and testing set at the ratio of $7:1:2$. Convergent ResNet152 [18] or ResNet50[18] is loaded as initialization state of multi-task CNN, and other layers of multi-task CNN are initialized randomly.

Seeing from Table I, the combination of ResNet152 and MTL obtains highest average accuracy, 91.66%. It much outperforms the method in [12], 9.28%. It indicates that the multi-task CNN framework is better than the framework with fusion of multiple hand-designed features.

It is a pity that the average accuracy of Rest152 is just higher a litter than ResNet50, 0.36%. Whereas, the time complexity and space complexity of ResNet152 are larger than ResNet50, Table II. The test time of ResNet152 and MTL is 45.6 millisecond per image. This means that it benefits the real-time aspect of clothing description and retrieval.

This network also can be evaluated on DeepFashion dataset [9], since the attributes of DeepFashion dataset is divided into different branches, such as texture, shape and so on. This article is not discussed because space is limited.

C. Clothing Retrieval

As Section II-C and Section III-B explained, The output features of each branch from the multi-task CNN are used

	ResNet152+ MTL	ResNet50+ MTL	Before CRF[12]	After CRF[12]
Category	62.50	61.00	48.46	54.91
Gender	88.73	87.28	82.29	81.07
Black	85.51	83.77	84.91	84.07
Blue	94.78	93.04	90.23	90.7
Brown	92.46	91.01	85.93	87.15
Cyan	93.62	93.91	85.28	86.12
Gray	81.74	81.45	79.11	78.93
Green	96.23	95.36	88.64	89.49
Purple	95.94	95.65	80.42	82.85
Red	96.52	96.23	89.95	92.29
White	82.90	83.19	75.56	76.4
Yellow	96.81	96.52	84.35	83.41
>2 Color	91.30	92.46	77.9	79.02
Floral	97.36	97.95	66.78	84.72
Graphics	99.14	98.85	87.43	92.10
Plaid	96.17	96.76	68.08	76.21
Solid	95.33	96.00	82.76	90.70
Spot	99.12	97.64	67.43	76.87
Stripe	95.30	96.55	71.07	78.27
Average	91.66	91.30	78.77	82.38

TABLE I
ACCURACY OF CLOTHING ATTRIBUTES PREDICTION

	ResNet152+MTL (ms/f)	ResNet50+MTL (ms/f)
Training	322.3	82.1
Testing	45.6	18.1

TABLE II
TIME COMPLEXITY OF MULTI-TASK CNN

Fig. 6. Example of clothing retrieval result. The order is from the upper left corner to the lower right corner, and the following figure is the same.

Fig. 7. Partial attributes-based clothing retrieval.

	Category	Category+ Gender	Category+ Gender+ Color	Category+ Gender+ Color+ Pattern
Category	30.67	30.04	27.88	27.92
Gender	72.33	77.18	74.08	73.3
Black	67.71	67.45	71.6	70.7
Blue	86.29	86.52	87.54	87.2
Brown	84.77	84.63	85.61	85.51
Cyan	89.01	89.05	89.08	89.02
Gray	67.85	68.41	71.42	70.96
Green	90.24	90.24	90.24	90.27
Purple	90.17	90.19	90.24	90.15
Red	91.12	91.17	91.3	91.18
White	69.47	68.52	72.39	71.92
Yellow	91.06	91.17	91.28	91.32
>2 colors	84.67	84.55	85.04	84.96
Floral	90.45	90.16	90.04	90.47
Graphics	93.42	93.25	92.88	93.29
Plaid	89.32	89.3	89.18	89.48
Solid	75.24	75.13	74.78	76.19
Spot	89.67	89.78	89.68	90.31
Stripe	82.36	82.67	82.48	82.89
Average	80.83	81.02	81.41	81.42

TABLE III
ACCURACY OF ATTRIBUTES-BASED CLOTHING RETRIEVAL

	Category	Category+ Gender	Category+ Gender+ Color	Category+ Gender+ Color+ Pattern
Category	30.67	30.04	27.88	27.92
Gender	72.33	77.18	74.08	73.30
Color	82.94	82.90	84.16	83.93
Pattern	86.74	86.72	86.51	87.11

TABLE IV
ACCURACY OF ATTRIBUTES-BASED CLOTHING RETRIEVAL WITH DIFFERENT BRANCHES

to retrieve similar clothing based on a group of attributes. A retrieval example is showed in Fig. 6. First of all, these features are concatenated. Then they are normalized by Z-score Normalization. Next they are reduce half dimension by PCA. Finally, LSH is used to search the top 30 most similar clothing.

PCA reduces half dimension of concatenated features from CNN. According to our experiments, it decreases performance of LSH tinily, but, it double the speed of LSH.

The accuracy of attributes-based clothing retrieval is listed in Table III. Compared to all combinations of branches, the combination of all branches achieves the highest average accuracy of 81.42%. It shows richer features to achieve higher performance. However, the growth is small. When only a single branch feature is used for retrieval, attributes from other branches also get comparative performance since all attribute tasks share information through MTL.

The average retrieval accuracy of each branch attributes is listed in Table IV. The red diagonal of Table IV shows the highest average precision in each attributes branch. The above phenomenon proves that the feature of this branch are beneficial to the attributes belonging to this branch. In each attribute branch the brown upper right corner is smaller

	Fast RCNN	Multi-task CNN	LSH	Total
Time ms/f	74.6	18.1	8.8	101.5

TABLE V
TIME COMPLEXITY OF THE WHOLE SYSTEM

than the red diagonal since more other branch features will affect the performance of attributes that belong to the previous branch. Nevertheless, brown upper right corner is larger than blue lower left corner since adding is better than not adding.

A detailed example is showed in Fig. 7. When the category branch feature is used for retrieval, the twenty-forth image is a female. However, when the tandem of the category and gender branches features are used for clothes retrieval, There are no females in the first 30 most similar images. And then the search range is expanded to the first 100 images. In the case of category and gender branches, the first female image appeared in the seventy-ninth. But it is the only female image. On the other hand, there are eleven female image in the case of only category branches.

The table V lists the runtime of each module for the entire system. The total time is 101.5 millisecond per image when the system is running on the CPU, $Intel(R)Xeon(R)E5-1650v3$. The network structures of Fast RCNN and multi-task CNN are $fast-rcnn-caffenet-pascal07-dagnn$ and $ResNet50$, respectively. This means that the system is real-time.

IV. Conclusion

In this paper, a novel framework is proposed for attributes-based clothing description and retrieval. Firstly, person is cropped by Fast RCNN. Then multi-task CNN predicts clothing attributes and extracts attributes-based features. Next, the half dimension of the concatenated feature is reduced by PCA. Finally, LSH is used to quickly search for the most similar clothing. This system is real-time and can be used for big data. Extensive experimental results show that this system is valid. It can be used for automatic clothing annotation and automatic clothing retrieval.

Acknowledgment

The first author would like to thank Rongqiang Qian for his valuable help.

References

[1] A. Nodari, M. Ghiringhelli, A. Zamberletti, M. Vanetti, S. Albertini and I. Gallo, "A mobile visual search application for content based image retrieval in the fashion domain", 2012 10th International Workshop on Content-Based Multimedia Indexing (CBMI), Annecy, 2012, pp. 1-6.

[2] C. Grana, D. Borghesani and R. Cucchiara, "Class-Based Color Bag of Words for Fashion Retrieval", 2012 IEEE International Conference on Multimedia and Expo, Melbourne, VIC, 2012, pp. 444-449.

[3] X. Wang, T. Zhang, D. R. Tretter and Q. Lin, "Personal Clothing Retrieval on Photo Collections by Color and Attributes", in IEEE Transactions on Multimedia, vol. 15, no. 8, Dec. 2013, pp. 2035-2045.

[4] W. Di, C. Wah, A. Bhardwaj, R. Piramuthu and N. Sundaresan, "Style Finder: Fine-Grained Clothing Style Detection and Retrieval", 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops, Portland, OR, 2013, pp. 8-13.

[5] Q. Chen, J. Huang, R. Feris, L. M. Brown, J. Dong and S. Yan, "Deep domain adaptation for describing people based on fine-grained clothing attributes", 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, 2015, pp. 5315-5324.

[6] M. H. Kiapour, X. Han, S. Lazebnik, A. C. Berg and T. L. Berg, "Where to Buy It: Matching Street Clothing Photos in Online Shops", 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, 2015, pp. 3343-3351.

[7] J. Huang, R. Feris, Q. Chen and S. Yan, "Cross-Domain Image Retrieval with a Dual Attribute-Aware Ranking Network", 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, 2015, pp. 1062-1070.

[8] Y. Xiong, N. Liu, Z. Xu and Y. Zhang, "A parameter partial-sharing CNN architecture for cross-domain clothing retrieval", 2016 Visual Communications and Image Processing (VCIP), Chengdu, 2016, pp. 1-4.

[9] Z. Liu, P. Luo, S. Qiu, X. Wang and X. Tang, "DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations", 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 1096-1104.

[10] R. Girshick, "Fast R-CNN", 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, 2015, pp. 1440-1448.

[11] Y. Xia, B. Zhang and F. Coenen, "Face occlusion detection based on multi-task convolution neural network", 2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), Zhangjiajie, 2015, pp. 375-379.

[12] H. Chen, G. Andrew, and G. Bernd, "Describing clothing by semantic attributes", European Conference on Computer Vision, 2012, pp. 609-623.

[13] K. He, X. Zhang, S. Ren and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 9, pp. 1904-1916, 2015.

[14] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Region-Based Convolutional Networks for Accurate Object Detection and Segmentation", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 1, pp. 142-158, 2016.

[15] M. Datar, N. Immorlica, P. Indyk, and V. Mirrokni, "Locality-sensitive hashing scheme based on p-stable distributions", DIMACS Workshop on Streaming Data Analysis and Mining, 2003.

[16] A. Vedaldi and K. Lenc, "MatConvNet: Convolutional Neural Networks for MATLAB", acm multimedia, 2015, pp. 689-692.

[17] C. Zhang and Z. Zhang, "Improving multiview face detection with multi-task deep convolutional neural networks", IEEE Winter Conference on Applications of Computer Vision (WACV), 2014, pp. 1036–1041.

[18] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition", 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 770-778.