Interactive Image Search for Clothing Recommendation

Zhengzhong Zhou, Yifei Xu, Jingjin Zhou and Liqing Zhang
MOE-Microsoft Key Laboratory for Intelligent Computing and Intelligent Systems
Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China
{tczhouzz, fei960922, zhoujingjin}@sjtu.edu.cn, zhang-lq@cs.sjtu.edu.cn

ABSTRACT

This demo delivers a novel retrieval system meeting users' multi-dimensional requirements in clothing image search. In this system, users are able to use both image and text descriptors as query inputs. We employ the color, texture, shape and attributes as additional descriptors to further refine the requirements. We propose the Hybrid Topic (HT) model, a probabilistic network integrating the multi-channel descriptors into a unified framework, to learn the intricate semantic representation of the above descriptors. The proposed model provides an effective multi-modal representation of clothes. Our experiments show that the HT method significantly outperforms the CNN-based deep search methods.

Keywords

Content-based Image Retrieval, Clothing Image Retrieval, Topic Model

1. INTRODUCTION

Using image retrieval technology to search for clothing products provides great convenience in apparel e-commerce [6, 4, 5]. When a user gravitates to someone's attire in a photo, he can simply submit the photo to the retrieval system to get a group of garments similar to the photo for selection.

However, there still exist two major problems. The first one is "semantic gap". Due to the intra-class variation and inter-class ambiguity in image space, low-level visual features could not describe the clothing semantic like category or style explicitly. The second one is "multi-dimensional requirements". Since the appropriate query images are not always available, users tend to further refine their demands by modifying the color, texture, shape and attribute descriptors of the original queries.

In this demo, we present a novel interactive retrieval system (see Fig. 1) to tackle both problems utilizing topic model [1, 9, 8]. Our feature extraction procedure is described as a two-level abstraction. For the first level abstraction, we extract clothing descriptors including color, texture, shape and text around images and convert them into the bag of words (BOW) formats. The BOW format enables us to use topic model for semantic analyzing. Meanwhile, it benefits us to modify the local details of image descriptions.

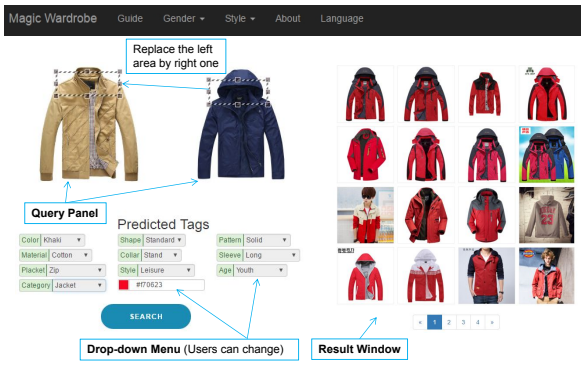

Figure 1: Illustration of System Interface.

For the second level abstraction, we present a probabilistic network to combine the multi-channel descriptors into a unified framework, *i.e.* Hybrid Topic (HT) model. In this model, each image is depicted as a linear combination of hybrid topics, while each topic is described as the distributions over various codewords. The HT model learns the intricate relationship among different descriptors, which can derive an effective representation of clothing images.

We develop an interactive retrieval system on the server with 2 Intel Xeon 2.6GHz processors and 128GB memory. It contains about 120,000 upper-body clothing images crawled from online shopping platform JD.com. Experiments show that our system not only achieves better performance than the CNN-based deep search method [5], but provides a more personalized interaction for online customers as well. The average retrieval time for a query is in 1 second.

2. SYSTEM FRAMEWORK

Fig. 2 shows the flowchart of our system. Given a query image and corresponding search conditions, the region of clothes is cropped using object detection algorithm as a preprocessing stage. We extract multiple visual descriptors from the detected region and quantize them into BOW representation. Furthermore, we adjust the codeword weights of the modified feature to meet users' requirement.

MM '16 October 15-19, 2016, Amsterdam, Netherlands

ACM ISBN 978-1-4503-3603-1/16/10.

DOI: http://dx.doi.org/10.1145/2964284.2973834

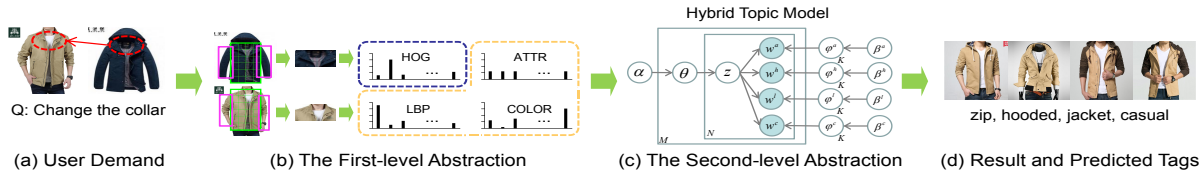

Figure 2: The Framework of Our Two-level Abstraction.

Using the pre-trained hybrid topic model, we integrate the above descriptors into effective retrieval features and infer the attributes of query image to improve the user interaction. Finally, we calculate the similarity between query image and images in database. Those top ranked images are returned as the retrieval result.

The First-level Abstraction. In the pre-processing stage, we adopt object detection [10, 3, 7] to localize the torso and sleeve regions in clothing images. We resize the torso region to 192×128 and use two slide windows (128×64 and 64×128) to sample it (stride 16). In this way, we split the image into 17 local regions (2 sleeve regions, 1 torso region and 14 overlapping regions cropped from the torso region). For each local region, we extract 3 types of visual descriptors: The histogram of oriented gradients (HOG) [2], Local binary patterns (LBP), and color histogram (COL). We also extract the keywords from the text description as attribute descriptor (ATTR). Using sparse coding, we convert all these descriptors into BOW formats, so that we can simply modify the codeword weights of descriptors instead of image contents.

HT Model for the Second-level Abstraction. As noted, we present a novel topic model, *i.e.* Hybrid Topic (HT) to integrate both visual and text information of clothing images. Fig. 2(c) shows the graphical model of HT. Images are represented as a list of "phrases". A phrase is a set of codewords w^j in different descriptor channels ($j \in \{a(ATTR), h(HOG), l(LBP), c(COL)\}$). For each image, we draw a topic mixture proportion θ to describe its high-level concept. For each phrase, we assign a topic z to determine its meaning. We assume that w^js and z follow the Multinomial distribution with prior φ^js and θ. While φ^j and θ follow the Dirichlet distribution with prior β^js and α. The above parameters of HT model are estimated by Gibbs sampling. To achieve an effective image representation, we train 17 HT models (one for each local region) on the dataset, and denote the topic proportion θs as local descriptions of images. Thus, we concatenate the θs extracted from 17 local regions of a clothing image as its retrieval feature.

Image annotation. To further facilitate the users, we automatically add tags to the query image in our retrieval system. The HT model is able to evaluate the missing values of text descriptors by the Bayesian reasoning. Meanwhile, users could modify these recommended tags to confirm their requirements.

3. EXPERIMENTS

To evaluate the performance of our approach, we realize a CNN based method called Deep Search [5] as the baseline. We invite 10 subjects to perform 100 retrieval tasks in our retrieval system. For each task, we upload a clothing image and modify its color, shape, texture or attributes to refine the requirement. We measure the performance by $NDCG@k = \sum_{j=1}^{k} \frac{2^{r(j)}-1}{\log(j+1)}$, where $r(j) = 1$ iff the j^{th} returned image satisfies the subject's demand, otherwise 0.

Table 1: The Retrieval Accuracy of HT.

	$NDCG@20$
Deep Search	0.42
Hybrid Topic	0.66

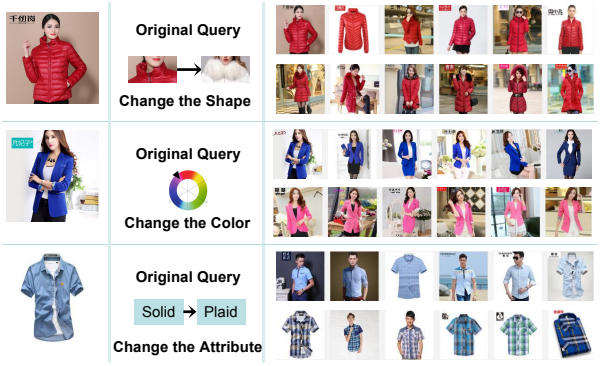

Figure 3: Several Groups of User Requirements and Corresponding Results.

Our experiment result is shown in Table 1. As we see, the hybrid topic method improves the precision by 24% compared with the Deep Search. It integrates the demand of modification into image representation, which reduces the distance between a query and user desired images in feature space. Fig. 3 shows some examples of our retrieved results.

4. CONCLUSIONS

In this demo, we present a novel interactive clothing image retrieval system to meet users' multi-dimensional requirements which can be refined through editing both visual features and attributes. It offers a two-level abstraction for clothes and constructs hybrid topic models to link the relationship among visual and attribute descriptors. Based on our framework, more diversified retrieval applications can be designed in further research.

5. ACKNOWLEDGMENTS

The work was supported by the National Basic Research Program of China (2015CB856004), and the National Natural Science Foundation of China (61272251).

6. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision & Pattern Recognition*, pages 886–893, 2013.

[3] R. Girshick. Fast r-cnn. *Computer Science*, 2015.

[4] J. Huang, R. Feris, Q. Chen, and S. Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In *IEEE International Conference on Computer Vision*, pages 1062–1070, 2015.

[5] J. Huang, W. Xia, and S. Yan. Deep search with attribute-aware deep network. pages 731–732, 2014.

[6] S. Liu, Z. Song, M. Wang, C. Xu, H. Lu, and S. Yan. Street-to-shop: cross-scenario clothing retrieval via parts alignment and auxiliary set. In *ACM International Conference on Multimedia*, pages 3330–3337, 2012.

[7] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, pages 1–1, 2016.

[8] Z. Sun, C. Wang, L. Zhang, and L. Zhang. Query-adaptive shape topic mining for hand-drawn sketch recognition. In *ACM International Conference on Multimedia*, pages 519–528, 2012.

[9] C. Wang, D. Blei, and F. F. Li. Simultaneous image classification and annotation. In *IEEE Conference on Computer Vision & Pattern Recognition*, pages 1903–1910, 2009.

[10] C. Wang, L. Zhao, S. Liang, and L. Zhang. Object proposal by multi-branch hierarchical segmentation. In *IEEE Conference on Computer Vision & Pattern Recognition*, pages 3873–3881, 2015.