Product Image Search with Deep Attribute Mining and Re-ranking

Xin Zhou[1], Yuqi Zhang[1], Xiuxiu Bai[1], Jihua Zhu[1], Li Zhu[1], and Xueming Qian[1,2(✉)]

[1] School of Software Engineering, Xi'an Jiaotong University, Xi'an 710049, China
{zhouxin0209,zhangyuqi}@stu.xjtu.edu.cn,
{xiuxiubai89,zhujh,zhuli,qianxm}@mail.xjtu.edu.cn

[2] The Ministry of Education Key Laboratory for Intelligent Networks and Network Security,
Xi'an Jiaotong University, Xi'an 710049, China

Abstract. With the high-growing of e-commerce, more and more users have changed to buy from websites rather than in stores. To deal with mass products, the traditional text-based product search has become incompetent to meet use's requirement. In this paper, we explore deep learning with convolutional neural networks (CNN) to resolve query's classification, and propose an efficient approach for product image search. For a query image, we first train a CNN model of a large database containing various product images to discriminate the query's category. Then we search similar products from the established category and utilize these visual results to parse the query with attribute. Finally we use the extracted attribute tags to finish the textual re-ranking and obtain the most relevant retrieved product list. Experimental evaluation shows that our approach significantly outperforms state of art in product image search.

Keywords: Product image search · Visual results · Attribute tags · Textual re-ranking

1 Introduction

With the rapid development of the internet technology, e-commerce websites are becoming increasingly popular as the convenient shopping experience they provide for users. Many shopping search engines are designed to have numerous personalized services to meet user's various demands, like Google, Amazon, and Taobao. For a query, the item retrieval list can be located by category, price, brand and etc.

Nowadays, the shopping search engines are mainly based on textual query. Users can search the item they desired to by inputting several key words, such as "black, tight, dress". However, it is difficult to clearly describe an item by only several brief and rough keywords to query about. Meanwhile, facing a great deal of diverse items, users usually spend a lot of time to select one by one in the tediously item retrieval list. Furthermore, consider such a scene: when reading a magazine or viewing an album, users may find some items they are interested in and take photos intending to have a similar one. In traditional textual search [16], there is no problem if the interested item is a book, electron, or other item with a concrete name. But it turns out to be troublesome when the

E. Chen et al. (Eds.): PCM 2016, Part II, LNCS 9917, pp. 561–570, 2016.
DOI: 10.1007/978-3-319-48896-7_55

items are clothing, bag, or shoes. It is difficult for us to give a clear description from the photo with concrete style details to find out their real needs. Once reaching the wrong text label, the retrieval list may become terrible and useless.

Obviously the traditional textual search is not adequate to deal with the above situation while pictures often provide more important cues for product search. Content-based image retrieval (CBIR) techniques [11–13] have become classical as they provide an effective way to mine semantic information from images. But CBIR was mainly applied to search rigid images (such as building and landscapes). These algorithms are unable to deal with most of product images as they are non-rigid. Especially for clothing, it is apt to folding and geometric deformations. In addition, the style detail of a product often plays an important role in searching, while the traditional algorithms may omit it. SIFT (scare-invariant feature transform) is one of the most robust registration method of correspondence matching [11–15]. And it performs well in our experiment on most categories.

As the query's category is unknown, the search will surely take a long time and make an inaccurate result within multiple categories products. So, we employ CNN to infer query's category. As we all know, recently deep learning has been proposed to solve data learning and analysis problem well for the Big Data.

In this work, we first collect a large dataset of a variety of categories products from the famous e-commerce websites, such as Amazon. Our dataset combines product images and detailed description texts. And we utilize the description information to automatically label each product images with their style attributes. Beyond colors and patterns, users mainly care the styles of the products.

For a query image, we first use the CNN model trained offline to determine the query's category, and consider the visual similarity to retrieve the similar product images in the determined category. Then we use the top examples of initial product list to orderly predict style attributes of the query. Finally we compute the similarity between the predicted tags and each product tags in visual retrieved results to re-rank the initial product list, providing the most similar products to query for users. The main contributions of our work are as follows:

(1) We propose an effective product search with attribute re-ranking approach. This approach uses the visual search results and the textual descriptions for results re-ranking.
(2) We propose an effective approach to estimate the category of the input product image by deep neural networks. This approach makes full use of the excellent discrimination characters of deep learning tools, such as convolution networks, to predict the category of the query.
(3) We propose category constrained search by utilizing visual words of SIFT feature to represent the local feature of the query image, to get the visual similar images. The category refined images is far less than the original product dataset, which can not only guarantee the search performances but also speed up the searching process.

2 Related Works

Product Search Based on Attributes: Nowadays, many e-commerce searching engines mainly depend on textual retrieval and design a fine-grained and all-sided description system for products. The product searching platforms, such as Amazon, Taobao, and 360buy, have defined various attributes (style, color, pattern, etc.) for users to search. With the rapid increase of the product database scale, more refined descriptions for product were created, especially in garment domain. Some work [4, 5] has primarily focused on analyzing the relationships between general clothing attributes, with respect to human activeness and occasion. Bossard et al. [6] presented a classification pipeline for classifying upper-body clothing by visual attributes and related them to occasion, season and lifestyle. In [7], a style-related vocabulary presenting clothing compositions was built for a new garment dataset to improve the search performance. Note that the current work is almost only for clothing, we expand the attributes to wider domains, not only for clothing, but also for shoes, bag, accessory and other product categories.

Product Image Search: Although there are some achievements on general image retrieval, researches on product image search are still limited. Recently some search engines for product image search have been developed. Google Googles and Amazon Flow are well-known commercial mobile product image search engines, but working robustly only for a few near-planar, textual object categories. And the search results are always inaccurate to meet users' real needs. Recently, researchers have devoted some efforts to product image retrieval. Tseng et al. [8] proposed an efficient garment visual search based on shape categories and styles, which did well in solving garment matching on non-rigid geometric deformations. In [9], Mizuochi et al. focused on local similarity of clothing retrieval with multiple images, which allowed user to circle one part of clothing they require while inputting the query image and provide a comprehensive result. Consider that query image always contains body part, Yamaguchi et al. [10] utilized pose estimation to analyze the clothing components in a fashion photograph, and use the analysis results to search similar clothing. As we can see, research on product image search has always been a challenge to cover a wide range of product categories while providing accurate results. In this paper, we consider a simple but efficient approach to make some efforts in this domain.

Deep Learning: Deep learning is a hierarchical model designed to emulate the learning process of the sensorial areas of the neocortex in the human brain. The algorithm can learn low-level features to obtain complex data representations through a hierarchical learning. Moreover, some companies like Google, Microsoft, and IBM have used Deep learning for the project of analyzing the big data of users. To handle the high-dimensional data with Deep learning, CNN are proposed. CNN have achieved impressive accuracy improvements in many areas of computer vision, especially in image parsing. In [1], Huang et al. proposed a Dual Attribute-aware Ranking Network (DARN) for cross-domain clothing image retrieval, which used CNN to recognize the semantic attributes of input clothing. Chen et al. [2] used CNN to resolve clothing style classification and retrieval task. Lin et al. [3] designed a clothing image retrieval framework by combining deep learning with Hash codes. In our work, we use the collected multiple categories product images, not only one clothing category, to train a CNN model for query's category recognition, so that we can finish the subsequent work better.

3 Approach Overview

For a query image, our approach consists of three steps: (1) Recognize query's category through the pre-trained CNN model. (2) Retrieve similar products by visual-based retrieval from the established category. (3) Use the visual results to parse the query's attributes and measure the similarity to re-rank the initial results.

Figure 1 depicts the overall retrieval pipeline. We firstly input the query to the pre-trained CNN model to determine which category the query belongs to. Then we apply the image retrieval approach based on hierarchical vocabulary tree [11] to obtain the initial visual retrieved results from the established category. Note that the visual retrieved results might contain a lot of noise, then we use some textual methods to improve the initial results. The process can be divided into the following steps: (1) K-NN tag prediction to parse query's attributes, (2) Attribute similarity to re-rank the visual results. The detail was discussed in Sects. 3 and 4. For a query image, we recognize its category and compare its visual feature and textual feature to other products in the established category. Therefore, we retrieve an image list ranked by style similarity to the query image.

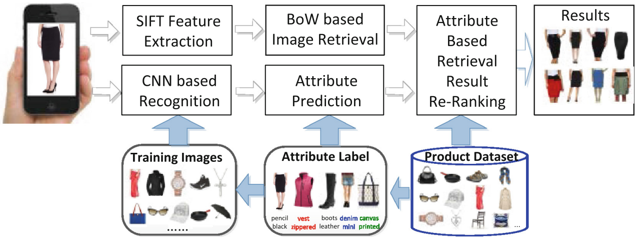


Fig. 1. The retrieval pipeline of our product searching approach

3.1 CNN Based Recognition

We use the CNN method to solve the problem of query's category recognition. We use the collected product images to train and fine-tune the pre-trained network with latent layers. In the training process, we first select multiple categories product images as training set containing 123700 images, and they are fed into the pre-trained network with the initial training parameters we give. Then we use another test set as input to measure and supervise the accuracy of the classified results. Meanwhile, we fine-tune the training parameters and compare the test results and finally choose the best. After obtaining the best fine-tune network, the query is fed into the network and the outputs of the last layer become the classification results.

Note that the feature vectors generated from the latent layers contains the classification information. Each element value in the vector is the probability of which category the query is determined to. Experiment demonstrates that our trained network works well in classification. The detail of experiment result was discussed in Sect. 4.

3.2 Visual Retrieval

We utilize the image retrieval approach based on BoW [11] to preliminarily obtain a ranking result based on visual similarity, which can be helpful for parsing query item subsequently.

(1) Visual feature descriptor
We consider SIFT feature as visual feature descriptor for style retrieval which is useful for finding styles with similar appearance. SIFT was developed for image feature generation in object recognition applications. The invariant features extracted from images can be used to perform reliable matching between different views of an object or scene. So that we extract the 128-D SIFT feature to describe each image in our database, and carry our hierarchical quantization to obtain the visual word representation [12–16]. In this paper, we experimentally choose depth = 7 and breadth = 10 for hierarchical clustering, and experiment result show that the final number of visual word is 590518.

(2) BoW based retrieval
After the hierarchical clustering and quantization, we utilize the BoW (bag-of-word) model [11] to present an image. The BoW model is built as a visual word histogram by statistics of the number of features of different visual words in an image. By this time, in the established category, we present an image i as a N dimensional BoW histogram named h(k), k = 1, 2,… N. To consider a query image q, we present it as the corresponding dimensional BoW histogram named q(k) in the same way. We use the Euclidean distance to measure the similarity between a pair of tags.

$$D(i, q) = \sum_{k=1}^{N} (h(k) - q(k))^2 \tag{1}$$

where D(i, q) is the similarity score between two images. We rank the results by the score in descending order. Now we obtain a ranked item list with similar appearance.

3.3 Textual Re-ranking

Following the visual retrieval, we start to re-rank the visual results using the textual approach we proposed.

(1) Attribute prediction
Since the visual retrieval returns a result with visual similarity, we then aim to find the most similar items in details in the visual result. It is widely accepted that style may be the most representative attribute for an item. To analyze our dataset, we extract the

useful description information as style tags for each item. Then we use K-NN to predict the query's attributes. For a Top K prediction, we record the times each tag appeared in the K re-ranked samples and compute its frequency. In a K samples statistic, the frequency is defined as follow:

$$f = T/K \tag{2}$$

where T is the appearance times of the corresponding tag in K samples. In this paper, we experimentally selected K = 6. We also define N_f as the number of the tags whose frequency is f. Then we experimentally make a rule to select the most relevant tags for query:

a. If $f > 0.5$ and $N_f > 5$, the 5 nearest item tags are selected; otherwise $N_f \leq 5$, all the tags that satisfies $f > 0.5$ are selected.

b. If all $f = 0.5$, the nearest sample's tags are selected.

(2) Attribute similarity to re-rank

After finishing parsing the query image in details, we represent a query with several attribute tags. We compute the similarity between tags using *Normalized Google Distance (NGD)*. The Normalized Google Distance is a semantic similarity measure derived from the number of hits returned by the Google search engine for a given set of keywords. Keywords with the same or similar meanings in a natural language sense tend to be "close" in units of *NGD*, while words with dissimilar meanings tend to be farther apart. Experiments demonstrate it works well in our paper. For a pair of tags, the similarity score can be compute as follow:

$$u_{i,j} = \exp\left\{-\frac{\max\left[\log n(t_i), \log n(t_j)\right] - \log n(t_i, t_j)}{N - \min\left[\log n(t_i), \log n(t_j)\right]}\right\} \tag{3}$$

where $n(t_i)$ is the number of images which contain t_i, $n(t_j)$ is the number of images which contain t_j, $n(t_i, t_j)$ is the number of images which contain both t_i and t_j.

We compute the score of each sample tag with all the query tags using *NGD*, and select the highest score as final score of the sample tag. Then we can compute average *NGD* score of an sample item:

$$Score_{NGD}(n, q) = \sum_{k=1}^{N_{tags}} u_{i,j}/N_{tags} \tag{4}$$

where N_{tags} is the number of tags of the sample item.

Consider that the visual similarity plays an important role in the retrieval, we add the exponentiate visual score *D(n, q)* to the final relevance score. We also define a weight λ of the textual score, so that the weight of visual score is 1 – λ. The whole score between a query image *q* and a sample image *n* is defined as follow:

$$Score(n) = \lambda Score_{NGD}(n, q) + (1 - \lambda)\exp(-D(n, q)) \tag{5}$$

In this paper, the value of λ is experimentally discussed to be 0.5. We re-rank the results above according to the *score(n)* in descending order so as to get the final retrieved item list.

4 Experiment

4.1 Dataset Construction

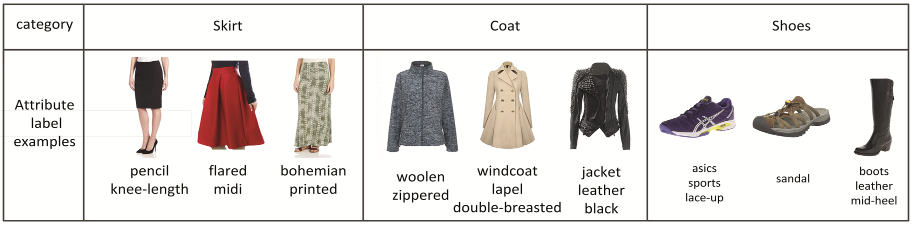
We collect the product dataset from the famous online shopping website, Amazon.com. In order to enrich the diversity of products, we totally harvested 123700 products with 20 categories. The dataset mainly contains three fields: product images with different visual angles, detailed description information, and the whole users' reviews from the corresponding product.

We note that the product description statements always contain the key words to well present style of the product. So that we consider an automatically tag labeled method to create style tags for each product. For a category (such as skirt), we firstly compute the word frequency from the description statements of all products and sort the words according to the frequency. Then we manually select the most critical descriptive words with high frequency as a benchmark (e.g. pencil, mini, knee-length, etc.) for the category. For a product, we extract the words appearing in the benchmark from the description statements as style tags (e.g. pencil, knee-length). Particularly, different benchmarks were assigned to different labeling tasks so as to reflect the unique characteristics of different category. Table 1 shows some benchmarks of different categories. As thus, we complete the whole dataset labeling (see examples of the labeling interface in Table 2).

Table 1. Benchmark examples of different categories

Category	Attribute benchmarks
Skirt	pencil, maxi, mini&short, A-line, printed, denim, midi, striped,…
Coat	pocket, hood, zip, double-breasted, leather, denim, vest, peacoat,..
Shoes	slippers, sports, boots, canvas, slingback, ballet, martens, leather…

Table 2. Attribute label examples of different categories



category	Skirt			Coat			Shoes		
Attribute label examples	pencil knee-length	flared midi	bohemian printed	woolen zippered	windcoat lapel double-breasted	jacket leather black	asics sports lace-up	sandal	boots leather mid-heel

4.2 CNN Recognition Precision

CNN recognition plays a very important role in our approach as it firstly constraint the category of query to search so as to guarantee our search performance and also speed up the searching process. Figure 2 shows the test results. We can see that the total classification accuracy of the top 1 category achieves 91.3%, and the top 2 categories achieve 96.4%, and the top 3 categories achieve 97.7%. In addition, most categories achieved more than 95%. As for the wrong results, their probabilities of top1 category are less than 0.9, but the right category appears at top2 or top3. So we make a rule for the subsequent search. For a query's classification result:

a. If the top1 category accuracy > 0.9, the subsequent search will be done only in the category.
b. If the top1 category accuracy < 0.9, the subsequent search will be done in the top three categories.

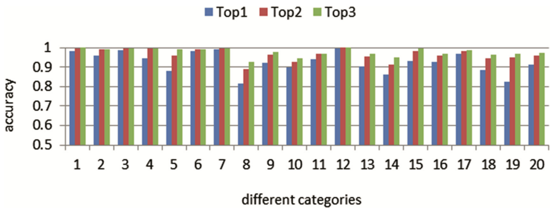


Fig. 2. Category classification accuracy of different categories

4.3 Product Image Search Precision

In order to evaluate search performance, we selected some product images as queries and manually labeled the ground-truth of them in advance. We measure the performance of our approach by computing the Precision (PR) of Top N retrieved product datum and the total Average Precision (AP) [13–16]. The PR calculation is extended to multiple categories products. We select 600 products from 20 different typical categories as queries to experiment. The whole process of search is based on the whole dataset which contains 123700 products with 20 categories.

To analyze the search performance of our approach, the two methods that SIFT feature and the low-level feature: color and texture (CT), based visual search without textual re-ranking (TR) are selected as traditional baselines. In addition, we compare the low-level feature combined with textual re-ranking (CT+TR) method, and the subsequent textual re-ranking is same to our approach. Figure 3 shows the AP of the above methods. We can observe that our approach is demonstrated obviously superior than others. CT achieves the lowest AP, which may be due to less discriminative to deal with clothing with transformation. The top1–top20 retrieval accuracy of CT+TR improves about 5% AP when compared with CT, demonstrating the contribution of textual re-ranking method we proposed. By comparing SIFT and our approach SIFT +TR, our approach achieves it is clearly that the results turned to be improved a lot after we add to the TR method.

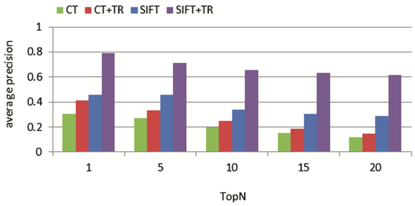


Fig. 3. Average precision on different methods

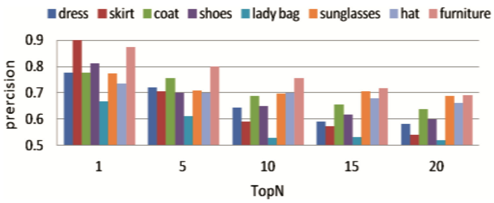


Fig. 4. Precision on different categories

Figure 4 shows the PR of different categories on our approach. We can observe that the PR relatively achieves high score on most categories, and even over 80% on some categories, such as skirt, furniture and shoes. As for other categories, the minimum is higher than 50%, and the Top1 and Top5 perform better than others. The different performance between different categories may be due to the irregularity recognition on categories. For example, the skirt is easier to recognize than lady bag, as bags usually contain more details.

4.4 Examples of Search Results

Figure 5 provides six examples search results. Overall, our approach performs well in the most experimental category. No matter product image is only about the entity or contains human model, there are few influences on the search results. Different from retrieving the same products to query, our approach provide more choices not only based on the same colors. As we can see, our approach is also sensitive to the figure and pattern of products showing the efficient consideration of product details.

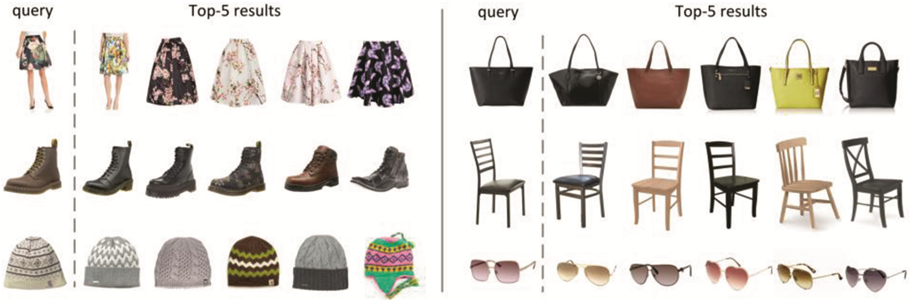


Fig. 5. Top 5 retrieved examples.

5 Conclusion

We have proposed an efficient product image search with deep attribute mining and re-ranking. Different from previous approaches, we concentrate on multiple various categories to search. Meanwhile, we constrain category search by CNN classification to guarantee the search performance and speed up the searching process. We innovatively consider the semantic features to refine the search results by using the attribute tags we mined textually. Experimental evaluation shows successful results on most categories, demonstrating a significant boost over previous work.

References

1. Huang, J., et al.: Cross-domain image retrieval with a dual attribute-aware ranking network. In: IEEE International Conference on Computer Vision, pp. 1062–1070. IEEE (2015)
2. Chen, J.C., Liu, C.F.: Visual-based deep learning for clothing from large database. In: ASE Bigdata & Socialinformatics. ACM (2015)
3. Lin, K., et al.: Rapid clothing retrieval via deep learning of binary codes and hierarchical search. In: ACM on International Conference on Multimedia Retrieval, pp. 499–502. ACM (2015)
4. Liu, S., et al.: Hi, magic closet, tell me what to wear! In: Proceedings of the 20th ACM International Conference on Multimedia, pp. 1333–1334. ACM (2012)
5. Nguyen, T.V., et al.: Sense beauty via face, dressing, and/or voice. In: Proceedings of the 20th ACM International Conference on Multimedia, pp. 239–248. ACM (2012)
6. Bossard, L., Dantone, M., Leistner, C., Wengert, C., Quack, T., Gool, L.V.: Apparel classification with style. In: ACCV (2012)
7. Wei, D., Catherine, W., Anurag, B., Robinson, P., Neel, S.: Style finder: fine-grained clothing style recognition and retrieval. In: CVPRW (2013)
8. Tseng, C.H., Hung, S.S., Tsay, J.J.: An efficient garment visual search based on shape context. In: Proceedings of the 9th WSEAS International Conference on Multimedia Systems and Signal Processing. World Scientific and Engineering Academy and Society (WSEAS) (2009)
9. Mizuochi, M., Kanezaki, A., Harada, T.: Clothing retrieval based on local similarity with multiple images. In: Proceedings of the ACM International Conference on Multimedia, pp. 1165–1168. ACM (2014)
10. Yamaguchi, K.: Parsing clothing in fashion photographs. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3570–3577 (2012)
11. Qian, X., et al.: Landmark summarization with diverse viewpoints. IEEE Trans. Circuits Syst. Video Technol. **25**(11), 1857–1869 (2015)
12. Qian, X., Tan, X., Zhang, Y., Hong, R., Wang, M.: Enhancing sketch-based image retrieval by re-ranking and relevance feedback. IEEE Trans. Image Process. **25**(1), 195–208 (2016)
13. Qian, X., Zhao, Y., Han, J.: Image location estimation by salient region matching. IEEE Trans. Image Process. **24**(6), 4348–4358 (2015)
14. Yang, X., Qian, X., Xue, Y.: Scalable mobile image retrieval by exploring contextual saliency. IEEE Trans. Image Process. **24**(6), 1709–1721 (2015)
15. Yang, X., Qian, X., Mei, T.: Learning salient visual word for scalable mobile image retrieval. Pattern Recogn. **48**(10), 3093–3101 (2015)
16. Lu, D., Liu, X., Qian, X.: Tag based image search by social re-ranking. IEEE Trans. Multimedia (2016). doi:10.1109/TMM.2016.2568099