Efficient street clothing recognition and retrieval with fusion of deep convolutional features

Ji Wei
Industrial Technology Research Institute, Zhengzhou University
Zhengzhou, China
lfl_ji@163.com

Liu Chang
Industrial Technology Research Institute, Zhengzhou University
Zhengzhou, China
525495574@qq.com

Jiao Ziao
School of Software and Applications Technologies, Zhengzhou University
Zhengzhou, China
1054568970@qq.com

Abstract*——*Traditional Content-Based Image Retrieval (CBIR) methods use only histogram of gradient or other approaches to extract shallow features of an image. In this paper, we propose a new CBIR method, which uses two convolutional neural networks to extract feature vector and a linear combination of cosine similarity and Euclidean distance to measure how close the image is to the query one. Before performing retrieval, we design a neural network based on VGG16 to predict the class that the query image belongs to. Experiments conducted on the dataset provided by JD Fashion challenge demonstrate the accuracy of our method.

***Index Terms*—Image retrieval, feature extraction, similarity**

I. INTRODUCTION (*HEADING 1*)

Image retrieval is to browse, search and retrieve images from a large amount of digital images. Most existing image retrieval approaches require meta data including captioning, keywords, or descriptions to images to perform retrieval over the manual annotated words. However, the manual annotation process is not only labor intense but also expensive. To address this, there have been many efforts done on automatic image retrieval, such as Content-Based Image Retrieval (CBIR) methods [1].

A CBIR method always consists of two major steps: the feature extraction and the similarity calculation. Feature extraction is to perform some calculations on an image to obtain a feature vector which represents the image. Feature of images in the database were extracted before the retrieval started, and the feature vectors usually stored into a search engine. The feature extraction will also be performed on the target image when the retrieval begins. The feature vectors of target image and images in the database are further used to measure the similarity such that images close to the target one could be found.

The quality of feature extraction has an importance to the accuracy of a CBIR method [2]. There has a large amount of research done on feature extraction, such as Histogram of Oriented Gradients (HOG), Local Binary Pattern (LBP) and Speeded Up Robust Features (SURF). HOG decomposes the target image into small cells and calculates a histogram of oriented gradients. The features consist of normalized histogram of each cell. Due to its operations on local regions, the HOG maintains desirable invariance to image geometry and optical deformation. Besides, local shape information can be well described by HOG, since its features represent the structural composition, such as edges. However, the computation cost of HOG is relatively high. Similar to HOG, the LBP also operates on local regions, which also named cells. For each pixel in a cell, an 8-digit binary number is obtained by comparing the center pixel's value to that of its 8 neighbors. Then a histogram of each number occurring is computed over all of the cells. The feature vector is the concatenation of histograms of all cells. LBP has significant advantages, such as rotational invariance and gray invariance. However, the performance of LBP may decrease due to the changes in lighting and reflection. Moreover, its time complexity rises significantly with bigger image resolution. SURF uses a blob detector based on the Hessian Matrix to find points of interest. The feature vector is based on the sum of the Harr wavelet response around these points. SURF is time efficient and robust against different image transformation. But it relies too heavily on the gradient direction of pixels in local regions. If a small deviation of the direction of gradient occurs, the result of feature matching may be much different, leading to mismatch of images. In practical, a pre-processing is often required before performing the above feature extraction algorithms. And the features extracted by these algorithms can be regarded as the shallow features of the image, which may not well represent the image.

The outperformance of Convolutional Neural Network (CNN), a popular deep learning method, shows its great potential in image retrieval [3]. In most existing CBIR methods, a certain type of an image is generally used as the basis for the similarity judgment to complete the final comparison and retrieval process. Since only one feature of an image is used, these methods are also called image retrieval based on a single feature. The single features of an image can be roughly divided into three categories: color features, texture features, and shape features. The above-mentioned image of a certain type of image can be extracted to give the visual display of the people's visual effects, but for the computer, these intuitive visual effects cannot be recognized, and the computer needs to generate the corresponding feature vector from the extracted features, and then Compare with the feature vectors pre-stored in the database to achieve the purpose of retrieval. In order to make up for the shortcomings of the single feature image retrieval method, the combination of multiple features is a

trend in the field of content-based image retrieval [4]. The method of this paper uses two CNN models to extract features and combines these features to perform retrieval.

In computer vision, as the depth of the neural network deepens, the features extracted by the neural network are more abstract, which is also an important reason for the deep network effect. However, the appearance of gradient explosions and gradient disappearance makes it difficult to train deeper networks, and their emergence may lead to problems that cannot be converged. When the deep network can converge, it will be found that as the depth of the network increases, the correct rate begins to saturate or even fall, which is called the degradation problem of the network. To solve the above problem, the residual network does not allow each layer of the network to directly fit the output of the previous layer while adding some shortcut connections to the forward network [5]. These connections skip some layers and pass the raw data directly to subsequent builds. These new shortcuts do not increase the parameters and complexity of the model. The entire model is still trained using an end-to-end approach and does not increase the difficulty of implementation. The VGG network was first proposed in 2014 and won the first and second results in the positioning group and the classification group respectively in the ImageNet competition that year. The model also has good results in generalization to other data sets. The VGG16 network uses a 3x3 convolution kernel to convolve the image, which can greatly preserve the image details, and finally classify the image after passing through three fully connected layers [6].

In this paper, we propose a novel CBIR method which uses convolutional layers of VGG16 and ResNet50 to extract image features, and combines the cosine similarity and the Euclidean distance to measure the similarity between two images. The outputs of VGG16 and ResNet50 are linearly combined to give the feature vector. In contrast to other CBIR methods, the advantage of our method is that the feature vector is not a single image feature description, and it is possible to preserve the content of the image.

II. The Method

In this paper, we propose a deep convolutional neural network based CBIR method which integrates two advanced CNN networks that are VGG16 and ResNet50 to extract features vectors. Besides, we propose a combination of cosine similarity and Euclidean distance to compute the similarity between the feature vectors of two images.

A. Overall framework

Our CBIR method consists of three major steps as shown in Fig.1. We firstly modified the fully connected layers of VGG16 to build a classification model which judges which class (shoe, cloth, or backpack) the target image belongs to. Then, the convolutional layers of VGG16 and ResNet50 networks are integrated to extract image feature. Finally, the similarity between the target and images in the database are computed via the proposed combinations of cosine similarity and Euclidean distance. During the similarity calculation, only images belonging to the same class in the database will be compared to the target image, which saves computation and improves accuracy.

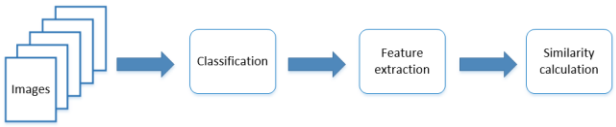

Fig. 1 Flow chart of the proposed CBIR method

B. Feature extraction and dimensionality reduction

In computer vision tasks, the output of convolutional layers is always used as an input to train a prediction or classification model. So, it is a feasible way to use convolutional layers instead of SIFT or other algorithms to extract feature vector. In our CBIR method, the trained convolutional layers of VGG16 and ResNet50 residual network are combined to obtain feature vector. It is worth noting that the output of VGG16 is of 2048 dimensions, while that of ResNet50 is of 512 dimensions. To address the unpredictable effect of unbalanced length of vectors, Primary Component Analysis (PCA) is applied to the output of VGG16 such that the lengths of outputs of both networks are equal. The feature vector of the target image is the linear combination of the outputs of two networks, as shown in Fig.2.

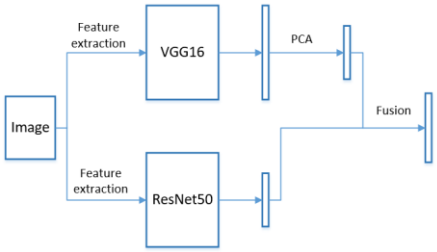

Fig. 2 Feature extraction: the final feature vector is the linear combination of the output of ResNet50 and post processed output of VGG16

C. Feature vector matching

In this section, we propose a measurement of similarity which combines cosine similarity and Euclidean distance, which are defined as follows.

$$dist_C(X,Y) = \frac{\sum_{i=1}^{n} x_i \times y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \times \sqrt{\sum_{i=1}^{n} y_i^2}} \quad (1)$$

$$dist_E(X,Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (2)$$

where $X=(x_1, x_2, \ldots, x_n)$ and $Y=(y_1, y_2, \ldots, y_n)$ are the feature vectors of two images to be matched, and $dist_C$ and $dist_E$ compute cosine similarity and Euclidean distance, respectively. From their definitions, it can be easily got that the bigger the $dist_C$ is the closer distance the two vectors is. And a smaller the $dist_E$ implies that the two images are more similar. However, the goal of the feature vector matching is to simultaneously obtain bigger $dist_C$ and smaller $dist_E$. To address this, we define the following similarity metric.

$$dist(X,Y) = \frac{1}{2}\left(dist_C + \frac{1}{1+dist_E}\right) \quad (3)$$

From Eq.(3), it can be seen that both the two components of the similarity metric vary from 0 to 1. So, the value of our similarity metric also falls in the ranges (0, 1), and the smaller the metric is, the more similar the two images is.

III. EXPERIMENTS

In this section, the proposed CBIR method is validated on the dataset provided by JD AI Fashion challenge. The dataset contains about 12,000 coupled digital images for training, and 150,000 images for querying. We test our method from the following two aspects: the classification and retrieval accuracy.

A. Classification accuracy

The first step of our method is to judge which class the images to be retrieved belongs to. Therefore, it is essential to test the classification accuracy. To obtain a better performance on classification, we modified the fully connected layers of VGG16. We also tested the classification accuracy of original VGG16 and ResNet50, as shown in Table 1. It can be concluded from the table that our modified network obtains the best accuracy, reaching 97%.

Table 1 CLASSIFICATION ACCURACY OF COMPARED MODELS

Model	**Classification Accuracy**
VGG16	0.85
ResNet50	0.91
VGG16 after training	0.97

B. Retrieval accuracy

To demonstrate the superiority of fusion feature over single feature, the proposed method is compared to CBIR methods using VGG16 or ResNet50 to extract feature vector and Eq.(3) to compute similarity. Table 2 depicts the retrieval and classification results of the compared methods. It is worth noting that the classification accuracy is the averaged ratio of retrieved images that belong to the same class of the target one. And the retrieval accuracy is the ratio that the most similar image which is retrieved is the same as the coupled image in the training set. From the Table, it can be easily got that the fusion feature is better than a single one in both retrieval accuracy and classification accuracy.

Table 2 Retrieval Accuracy

Model	Retrieval Accuracy	Classification Accuracy
VGG16	0.21	0.93
ResNet50	0.27	0.92
Our method	0.29	0.95

In the following, we test our method without classification before retrieval, as shown in Fig.3. The first column shows images to be retrieved; the rest columns present top five retrieved images. From the figure, it can be seen that images belonging to difference class from the image to be retrieved occur in the list, which may reduce the retrieval accuracy.

(a) (b)

Fig. 3 Retrieval results without classification before searching. (a) the two images to be retrieved. (b) the five retrieval images with best similarity to the targets. It is worth noting that there exist images which belong to different class of the target.

To demonstrate the effect of the classification process, we perform retrieval with the same input as Fig.3, as shown in Fig.4. From the figure, it can be seen that all the retrieved images fall in the same class as the query image. Moreover, the retrieval results after classification are also more accurate than the retrieval results without classification.

(a) (b)

Fig. 4 Retrieval results with classification before searching. (a) the two images to be retrieved. (b) the five retrieval images with best similarity to the targets.

IV. CONCLUSIONS

In this paper, we propose a content-based image retrieval method, which uses two popular convolutional neural networks to extract feature vector and combines cosine similarity with Euclidean distance to measure the similarity between two images. Experiments conducted on JD AI Fashion challenge data set show that the linear combination of features obtains better performance on both retrieval and classification than single feature. And the classification before retrieval also improves the performance of our method. Future works will focus on the improvement of feature extraction and similarity evaluation.

REFERENCES

[1] Rangkuti, Abdul Haris, Z. E. Rasjid, and D. J. Santoso. "Batik Image Classification Using Treeval and Treefit as Decision Tree Function in Optimizing Content Based Batik Image Retrieval". *Procedia Computer Science*, 59(2015), pp. 577-583.

[2] Yandex, Artem Babenko, and V. Lempitsky. "Aggregating Local Deep Features for Image Retrieval". IEEE International Conference on Computer Vision, 2016, pp.1269-1277.

[3] Salvador, Amaia, et al. "Faster R-CNN Features for Instance Search". IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016, pp. 394-401.

[4] Liu, Peizhong, et al. "Fusion of Deep Learning and Compressed Domain features for Content Based Image Retrieval". IEEE Transactions on Image Processing, pp.99(2017), pp. 1-1.

[5] He, Kaiming, et al. "Deep Residual Learning for Image Recognition". IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770-778.

[6] Simonyan, Karen, and A. Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". *Computer Science* (2014).