Differential Learning: A Powerful Tool for Interactive Content-Based Image Retrieval

Qinghe Zheng, Xinyu Tian, Mingqiang Yang and Hongjun Wang

***Abstract*—Faced with the huge image data in the context of big data era, how to effectively manage, describe, and retrieve them has become a hotspot issue in academia and industry. In this paper, we propose an end-to-end image retrieval system based on deep convolutional neural network and differential learning method. Compared with the traditional method of using the deep convolutional activation features as the feature vector to match the image, we simplify the retrieval process of the method and decrease the problem of "semantic gap" in the content- based image retrieval system. We first build an image matching dataset based on the gravitational field model, that is to add the similarity score label for each image in the dataset manufacturing stage. Then we train the improved deep learning model and verify the effectiveness of the algorithm on three common image matching datasets (*i.e.*, Caltech-101, Holidays, and Oxford Paris). Finally, the experimental results show that our improved deep learning model with differential learning method that used for image retrieval system has state-of-the-art image matching performance. The overall retrieval accuracy in Caltech-101, Holidays, and Oxford Paris datasets are 88.5%, 94.1%, and 96.2%, respectively. As the number of returned image increases, the retrieval accuracy of the system decreases slightly and eventually becomes stable at a high value. And the differential learning based retrieval method is superior to many traditional algorithms in terms of image matching accuracy and single image processing speed.**

***Index Terms*—image retrieval system; differential learning; deep convolutional neural network; gravitational field model; end-to-end training**

I. Introduction

With the rapid development of computer technology and multimedia technology, a large number of images are generated for information expression and transmission. Thus, How to find a specific image effectively in a massive image dataset requires the full use of image retrieval techniques.

Manuscript received September 12, 2018; revised November 29, 2018. This work was supported by the Fundamental Research Funds of Shandong University (Grant 2018JC040), the National Natural Science Foundation of China (Grant 61571275), and the Shandong Provincial Natural Science Foundation (Grant ZR2014FM030 and ZR2014FM010).

Qinghe Zheng is with School of Information Science and Engineering, Shandong University, Jimo, Qingdao 266237, China (e-mail: 15005414319-@163.com).

Xinyu Tian is with College of Mechanical and Electrical Engineering, Shandong Management University, Changqing, Jinan 250357, Shandong China (e-mail: 18769796159@163.com).

Mingqiang Yang is with School of Information Science and Engineering, Shandong University, Jimo, Qingdao 266237, China (corresponding author, e-mail: yangmq@sdu.edu.cn).

Hongjun Wang is the School of Information Science and Engineering, Shandong University, Qingdao 266237, China (e-mail: imageinstitute@-outlook.com).

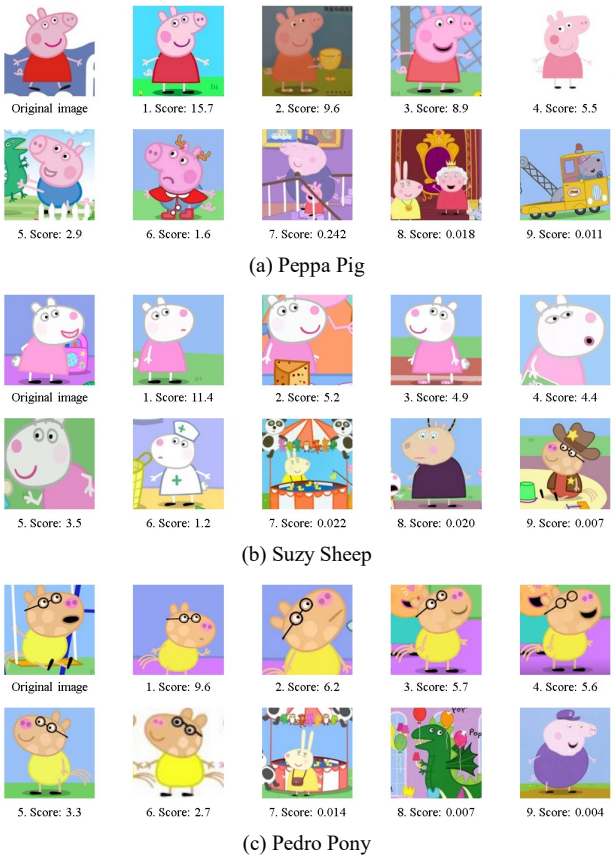


Fig. 1. The output results of the content-based image retrieval system. The retrieval results include the approximate image obtained by the search and their matching score.

Image retrieval has been a research hotspot in the field of computer vision and information retrieval in recent years. Content-based image retrieval (CBIR), which originated in the 1990s, is a mainstream method for early image retrieval. This method is based on the similarity of image visual features such as local features [1] or global features [4] to achieve the purpose of image searching, as shown in Fig. 1. However, content-based image retrieval system faces the semantic gap challenge [18], that is, the underlying features of the image are difficult to reflect the rich semantics of objects, events and scenes that contained in the image. Specifically, the CBIR system mainly includes the following difficulties:

- Selection of image descriptors.
- Determination of distance measurement method.
- Interference of noise.
- Lack of effective training samples.

The traditional content-based image descriptors [19], [20],

[21] [46] [47] [48] are difficult to describe the rich semantic information contained in the image, and it is difficult to use the ability to express the complete semantic meaning in the text. Therefore, how to effectively establish effective distance relationship between text semantics and image feature space is the key to solve the problem. On the other hand, the large amount of noises that introduced in the process of image establishment and transmission brings a lot of interference to the accurate extraction of image descriptors, which makes it difficult for them to establish the relationship between the image content and the image accurately. The lack of effective image samples that used for training also limits the retrieval performance of many algorithms and makes it difficult to fully exploit their abilities.

In response to these problems, many algorithms have been proposed to improve the performance of the CBIR system. In the image retrieval system, the quality of image features is very important to the retrieval performance. Accurate features can greatly improve image retrieval performance, and the loss of retrieval performance due to dimensionality reduction can also be reduced. Therefore, the feature extraction, as a very important part of image retrieval, has been studied more and more widely. The traditional content-based image retrieval system mainly uses low-level visual features, such as color, shape, texture, etc., or through the matching of feature points (*e.g.*, SIFT [1], SURF [2] and so on) to complete the image retrieval. Sheikholeslami [3] uses wavelet analysis to extract image texture features for geographic image retrieval. Lu [4] proposed a method of weighted color and texture feature distance for aerial remote sensing image retrieval. One of the biggest problems with these algorithms is that they cannot handle the semantic gap [5] [49] (that is, the huge difference between the similarity that machine get from low-level visual features and the similarity that people get from advanced semantic features).

In order to eliminate the influence of semantic gap, Zhou [6] proposed the relevance feedback algorithm. Because of the unfriendly and troublesome human-computer interaction, the relevance feedback is more applicable to the case of limited sample. Mojsilovic and Rogowitz [7] completed the image classification and retrieval simultaneously by searching the relationship between low-level visual features and accomplish image semantics through multi-level clustering analysis. Zhao [8] proposed a BOVW model based on multi-scale central circle structure, and proved that this method is effective in the retrieval of high resolution images. However, these methods have obvious manual intervention and they depend heavily on the classification or clustering algorithms, and therefore have poor adaptability in practical applications.

At present, deep learning technology is one of the most promising technology to solve the problem of semantic gap. The development of deep learning based model provides a new idea for the acquisition of image features. In the big data background, through the training of large-scale image data, more complex and powerful deep learning based models can fundamentally reveal the complex and abundant information in large scale images, and excite out a large number of representative features. Lin [9] proposed an effective deep learning framework to generate binary hash codes for fast image retrieval, and demonstrated the excellent performance of the algorithm on CIFAR-10 and MNIST data sets. Ji [10] used the features learned by the convolutional neural network for image retrieval, evaluates the performance of the output of different hidden layers of the neural network for image retrieval. However, the algorithm is not an end-to-end training and testing process. Therefore, we use the excellent feature extraction and classification capabilities of VGGNet [11] to redesign it into an end-to-end training and testing network for image retrieval. Through designing a novel inter-sample loss function, the training strategy based on matching accuracy is established for the neural network.

The rest of the paper is organized as follows. In section II, we introduce the related work of CBIR methods. In section III, we introduce an improved deep convolutional neural network structure that used for image matching. In section IV, we describe the differential learning process of the CBIR system based on gravitational field model in detail. In section V, we demonstrate the effectiveness of the algorithm by showing the retrieval results over three different datasets. Finally, we discuss what we learned, our conclusion and future works in section VI.

II. Related Works

How to improve computer's learning of image semantics? How to make image semantic information play an important role in the image retrieval system to improve the retrieval performance? In recent years, scholars have turned their research focus to large vocabulary learning (thousands to tens of thousands) based on large-scale datasets, and combine the features and semantic text information through information fusion technology to improve the image retrieval performance. The main technique of this type of CBIR system include the following aspects: digital image feature extraction, image semantic learning, and the fusion of image information. In this part, we intend to describe the development and status quo of these three technologies, and analyze the development trend of CBIR system, pointing out several research hotspots and development directions in the future.

A. Image representation

The early content-based image retrieval system compares the image similarity according to the visual features of the image. The digital features of the image include chromaticity [22], texture [23], shape [24] and space location [25], which can be used to describe the global feature of the whole image content or to describe the local feature of the object area in the image.

A detailed discussion of various typical image digital feature extraction algorithms is presented by Liu [26]. The chrominance feature [27] is the most widely used image digital feature. Commonly used chromaticity features include color histograms, color moments, color aggregation vector, and color covariance matrices. However, these chromaticity features cannot directly related to the high-level semantics. Therefore, Wu [28] defines a "domain color" of the HSV as the regional chromaticity feature. Experiments show that this method is more suitable for semantic-based image retrieval. Image texture feature extraction algorithm is divided into two types: spatial domain and frequency domain. Tamura texture [29] is a typical spatial domain texture feature, and the most

(a) fingerprint (b) shoe print

Fig. 2. Image examples. The feature descriptors used to describe images in different scenarios are different.

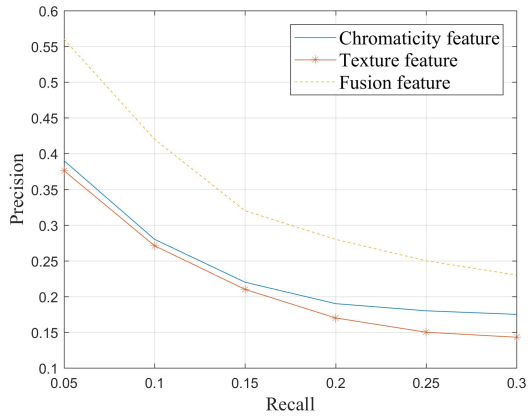


Fig. 3. The role of feature fusion in content-based image retrieval system.

commonly used frequency domain texture features are wavelet feature [30] and Gabor feature [31]. Shape features include aspect ratio, shape invariant moments, Fourier shape descriptor, circularity, etc. The spatial location feature is for the target object in the image. The simplest is the absolute spatial position such as "upper, middle and lower", and some algorithms use the central coordinate point of the object area as its position feature. Relative spatial location is usually more important than absolute location for image semantic learning.

Quellec [32] defines a novel object area feature called color-size feature, which includs color-size histograms and invariant moments. This feature combines color with the distribution information of the size of the region. The color distribution information is used to describe the content of the image, and the size of the area is used as the weighting factor when the image is matched. Yue [33] proposes a color feature extraction method based on color index correlation statistics. It combines numerical statistics to obtain the feature vectors with certain scale, translation and rotation invariance, so as to express the color and texture characteristics of the image more effectively. Choi [34] proposes a feature extraction algorithm which is robust to the change of image scale, image rotation and image transformation, and applies it to network image retrieval. The algorithm proposed by Bian [35] is based on the multi-scale space theory to extract the multi-scale information of the image and use the histogram projection to obtain the multi-scale phase features of the image. These algorithms provide information describing the content of the image from different perspectives.

Image retrieval systems should make full use of various image features and organically combine them. For certain types of images, global features can better describe their content (such as the fingerprint in Fig. 2 (a)), while for some images the local object area is more important (such as the On the shoe print in wooden floor in Fig. 2 (b)). In addition, extensive experiments [36] [50] have also confirmed that the combination of texture features and chromaticity features is better than the two features used alone, as shown in Fig. 3.

B. Semantic learning

In order to make image retrieval more close to human semantics and facilitate users to query directly with keywords, CBIR system introduces image semantic learning technology. In general, people are more interested in the object in the image than the whole image. Therefore, most image semantic learning algorithms are based on object regions.

Lai [36] analyzed and discussed various aspects of image semantic learning technology, and compared the advantages and disadvantages of different algorithms. Image semantic learning algorithms can be divided into five major classes: using relevance feedback, using common sense and ontology knowledge, using image templates, using machine learning techniques, and fusion of image visual features and text information [37]. The most commonly used technique is to obtain image semantics through machine learning algorithms, including support vector machine, neural network, K-nearest neighbor algorithm and so on. Cheng [38] proposed that the user selects the region of interest, then extracts the various features of the selected area to construct the corresponding classifier, and finally integrates the results of each classifier to get the retrieval results. Carvalho [39] combined with image template and decision tree technology to learn 19 kinds of natural scenery images. Lowanshi [40] used KNN method to classify a group of images. He [41] combined the regional significance analysis and the user participation feedback, and learned the query target concept that the user is interested in through the manifold sorting algorithm. Koch [42] though that due to the semantic gap between image features and semantic information, the integration of the two kinds of information of image semantics and visual features should be asymmetric. Based on this idea, the author introduced an improved late fusion technology. Wysoski [43] used the latent semantic kernels to integrate visual information and text information into a new description space, which defined the potential concepts.

Although these algorithms have achieved certain effects, there is a general problem in applying these techniques to actual image dataset retrieval: the vocabulary of semantic learning (category of images) is limited, far from meeting the needs of actual image retrieval. As the type of image increases, the accuracy of the search drops rapidly. For example, the algorithm in [42] has a learning accuracy of 68.2% for 10 vocabularies and 40.3% for 20 vocabularies. Most image semantic learning techniques can only analyze dozens of image categories. In practical applications, human semantics includes about 5 000 to 30 000 vocabulary concepts [43].

In general, it is an effective way to solve the semantic learning of large vocabulary by using the semantically rich large scale image dataset as the training set, and using the hierarchical semantic architecture and the decision tree technique. However, the practical application of the image semantic learning in massive image retrieval still needs further study.

III. Deep Convolutional Neural Network for Image Retrieval System

In this section, we introduce the image retrieval system based on an improved deep convolutional neural network. The new CNN architecture and corresponding loss function that used for image matching system are presented in detail. Moreover, the scoring mechanism of image retrieval system are designed to improve the performance of content-based image retrieval system .

A. Structure of Deep CNN for Image Matching

Deep convolution neural network achieves better results in feature recognition tasks than traditional methods. It has strong ability of feature extraction and image representation. Therefore, CNN is often used for image recognition and speech recognition. We summarize some of the advantages of deep learning, including:

- Big data support;
- No feature engineering required;
- Strong adaptability;

Compared with the traditional machine learning algorithm, deep convolutional neural network performs better with more training data. In practice, the best advice for improving accuracy through deep networks is to use more data. On the other hand, the traditional machine learning algorithm usually requires complex feature engineering. First, deep exploratory data analysis is performed on the dataset, followed by a simple reduction of sample dimension. Finally, we must carefully select the best function to pass them to the machine learning algorithm. When deep networks are used, data can be delivered directly to the network and good performance are achieved directly. Moreover, deep learning technology can be easily adapted to different fields and applications. Transfer learning enables pre training deep networks to be effective for different applications in the same field.

As shown in Fig. 4, we first construct the feature extraction module of the image matching system. We use the classical VGG-16 model, a very deep convolutional neural network for image classification. It consists of 13 convolutional layers, 3 fully connected layers, and 5 max-pooling layers. The size and number of convolution kernels in convolutional layers between two max-pooling layers are the same. The final classification results are output by the softmax layer. The convolutional layer reduces the number of training parameters (weights and bias) between the layers of the neural network by parameter sharing technique. Then, the down-sampling layer reduces the size of the next input data by sampling the output of the last convolutional layer, which generally uses the 2×2 window to do sampling operations, thus reducing the data by about 75%. There are two kinds of common used pooling methods: max-pooling method and mean-pooling method, where the mean pooling can reduce the error caused by limited neighborhood size. It can retain more background information of images, and is propitious to image retrieval. Therefore, the model introduced in this paper adopts the mean-pooling method.

In the training process, the convolutional stride is fixed to 1 pixel; the spatial padding of the input of convolutional layer is such that the spatial resolution is preserved after convolution, *i.e.*, the padding is 1 pixel for 3×3 convolution layers. Spatial

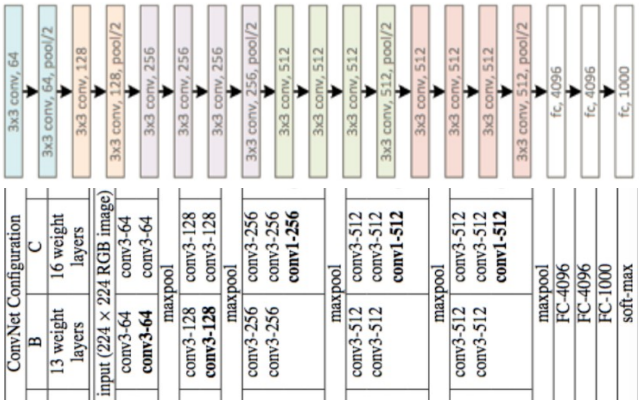


Fig. 4. Architecture of VGG-16 model for feature extraction.

pooling is carried out by 5 max-pooling layers, which follow some of the convolution layers (not all the convolution layers are followed by max-pooling). Max-pooling is performed over a 2×2 pixel window.

The input of the network is RGB image of 224×224 size. The only pre-processing operation is to subtract the mean value of RGB from each pixel in the training set. In order to augment the dataset, we need to randomly cut each re-scaled image in each iteration process. The cropped image also needs to be randomly flipped horizontally and RGB color shifted.

In the image retrieval system, the network VGG-16 with excellent feature extraction ability can be used to extract the high-level semantic information and can be prepared for the subsequent image matching. But in order to make the network to achieve image matching function, we remove the last one used to output the classification results of the fully connected layer and softmax layer.

B. Corresponding Loss Function

The performance of image matching system is not only dependent on the extracted image features, but also on the similarity measure function used in the image matching process. The similarity measure function directly determines the efficiency and the accuracy of image retrieval. Image retrieval system based on deep neural network generates feature vectors from low-level to high-level training. Finally, the corresponding image sample is represented based on feature vectors. Therefore, we first define the adjusted cosine similarity of the output of the last layer as a similarity measure function, as shown in (1).

$$sim(i,j)=\frac{\sum_{\mu\in U}(V_{\mu,i}-\overline{V_{\mu}})(V_{\mu,j}-\overline{V_{\mu}})}{\sqrt{\sum_{\mu\in U}(V_{\mu,i}-\overline{V_{\mu}})^2}\sqrt{\sum_{\mu\in U}(V_{\mu,j}-\overline{V_{\mu}})^2}} \quad (1)$$

where $V_{\mu,i}$ and $V_{\mu,j}$ represent the feature vectors extracted by DCNN of the *i-th* image and the *j-th* image, respectively. U represents the number of image samples in a training batch. $\overline{V_{\mu}}$ is the mean value of all feature vectors in a training batch. Cosine distance is used to measure the difference between two individuals by using the cosine value of the two vectors. Compared with Euclidean distance (Fig. 5), cosine distance pays more attention to the difference of directions between

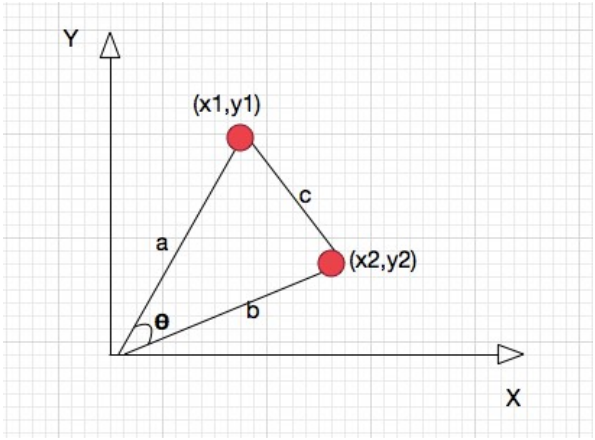


Fig. 5. Euclidean distance measure and cosine similarity measure.

two vectors. Euclidean distance is the absolute distance of each point of space, which is directly related to the position coordinates of each point, and the cosine distance is the angle of the space vector, which is more reflected in the direction difference rather than the position. Euclidean distance can reflect the absolute difference of individual numerical characteristics, so it is more used to analyze the differences from the dimensions of the dimensions, such as the use of user behavior indicators to analyze the similarity or difference of user values. The cosine distance is more distinguishable from the direction, but not sensitive to the absolute value. More used to use the user to distinguish the similarity and difference of interest by the content score. At the same time, the problem (because cosine distance is not sensitive to the absolute value) is corrected.

Then we design the loss function of the network according to (2).

$$J_i(\boldsymbol{W},\boldsymbol{b}) = sim(i,j) + \beta\sum_{j=1}^{s} KL(\rho \,\|\, \hat{\rho}) \tag{2}$$

where

$$KL(\rho \,\|\, \hat{\rho}) = \rho \log\frac{\rho}{\hat{\rho}} + (1-\rho)\log\frac{(1-\rho)}{(1-\hat{\rho})} \tag{3}$$

where

$$\hat{\rho} = \frac{1}{s}\sum_{j=1}^{s} a_i^2(x^j) \tag{4}$$

where J_i represent the loss function of image i. $\boldsymbol{W}$ and $\boldsymbol{b}$ represent the weight and bias of the DCNN, respectively. ρ is the sparse parameter and $\hat{\rho}$ is the mean value of ρ. β is a manually set hyper-parameter. s is the number of neurons in the last fully connected layer. Kullback-Leibler (KL) distance represents the relative entropy between two Bernoulli random variables, measures the difference between two probability distributions in the same probability space. a_j^2 indicates the activation extent of neuron i in the hidden layer in the case that the given input is x^j.

If softmax is directly trained, the features obtained do not necessarily have clustering characteristics actually. Instead, they will try to fill the entire sample space. An additional penalty, which defines a trainable center for each category, requiring each class to be close to its center. Therefore, in general, the first term is responsible for opening the distance between different classes, and the second term is responsible for reducing the distance between the same class.

C. Activation Function

The two most commonly used activation functions in the traditional deep convolutional neural network, the sigmoid systems (logistic-sigmoid, tanh-sigmoid) are regarded as the core of the neural network. From the mathematical point of view, the nonlinear sigmoid function has a large signal gain to the central region, and the signal gain of the two sides is small. It has a good effect on the feature space mapping of the signal. The non-linear combination of re-weighted inputs is used to produce nonlinear decision boundary. From the neuroscience point of view, the central region resembles the excitatory state of the neuron, and the two regions resemble the inhibitory state of the neuron. Therefore, the key features can be pushed to the central area in the neural network learning, and the non key features can be pushed to the two sides.

At present, a large number of activation functions have been developed for researchers to use. For example, arctan function, bent identity function, softplus function, softsign function and tanh function are defined as follows.

- Arctan function:

$$\arctan(x) = \tan^{-1}(x) \tag{5}$$

- Bent identity function:

$$f(x) = \frac{\sqrt{x^2+1}-1}{2} + x \tag{6}$$

- Softplus function:

$$softplus(x) = \ln(1+e^x) \tag{7}$$

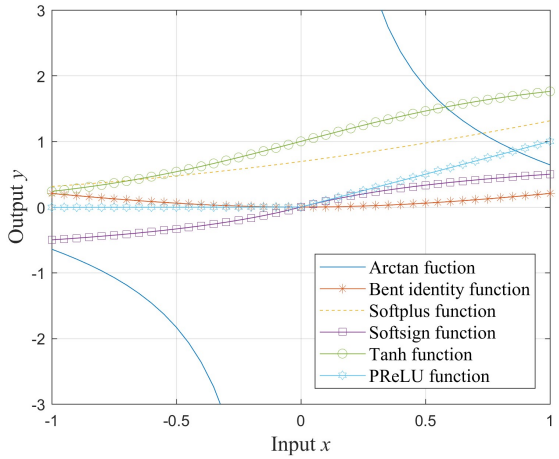
- Softsign function:

$$softsign(x) = \frac{x}{1+|x|} \tag{8}$$

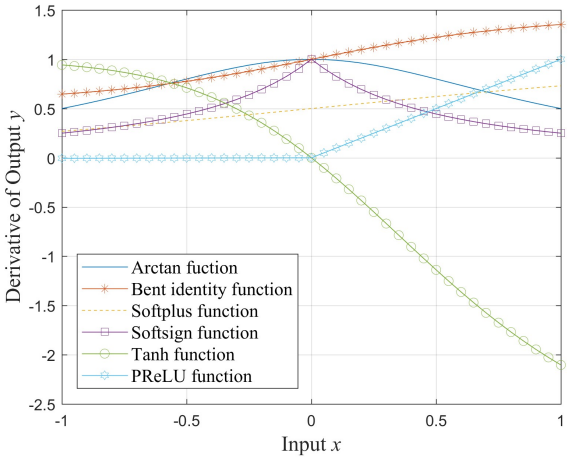
- Tanh function:

$$\tanh(x) = \frac{2}{1+e^{-}} \tag{9}$$

We draw the graphs of these activation functions and their derivatives to observe mathematical properties and analyze their role in the learning process of the deep CNN, as shown in Fig. 6. Suitable activation functions mainly include the



(a) activation functions



(b) derivatives

Fig. 6. Relationships between inputs and the outputs of frequently-used activation functions and their corresponding derivatives.

following attributes: nonlinear, continuous differentiable, monotonicity, smoothness, and approximate identity near the origin.

For the selection of the activation function, the Parametric Rectified Linear Unit (PReLU) function [12] as in (5) is used after each layer.

$$\text{PReLu}(x) = \begin{cases} x, & \text{if } x > 0 \\ \alpha x, & \text{if } x \le 0 \end{cases} \tag{10}$$

where α is a small fixed value (*e.g.*, 0.005). It avoids the gradient vanishing problem like sigmoid function. At the same time, it is faster and more efficient in large and complex data than sigmoid function. Compared with ReLU [13], it can avoid neuronal necrosis phenomenon, that is, once the neuron output is zero, it can no longer be trained. When the activation function is non-linear, a two-layer shallow neural network can be proved to be the general approximation function. The identity activation function does not satisfy this property. When a multi-layer neural network uses the identity activation function, the entire network is equivalent to a single-layer model. The continuously differentiable property is necessary for gradient-based optimization methods. On the other hand, The binary activation function has no differentiability at point zero, it can be zero at all other values, and the gradient based optimization method has no effect on it. The monotonicity of the activation function guarantees that the error surface associated with the single layer model is convex. Smoothing functions with monotonic derivatives have been proved to be better generalizing in some cases. These properties indicate that these activation functions are more consistent with Occam's razor. If the activation function has an attribute that approximates identity near the origin, the neural network will learn effectively if the weight is initialized for a small random value. Moreover, when the activation function does not have this property, a special case must be used in the initialization weight.

During the experiment, we tried a variety of activation functions to observe their validity. The results show that the PReLU function performs best in the content-based image indexing system.

D. Similarity Score

Similarity measure is a useful method that comprehensively judges the similarity between two things. The closer the two things are, the greater the similarity of them is, and the more distant the two samples are, the smaller their similarity measures. There are many kinds of methods for similarity measurement, which are usually selected according to specific practical problems. The common degree of similarity is the correlation coefficient (the degree of proximity between variables), the similarity coefficient (the degree of proximity between the samples). If the sample is given the qualitative data, the degree of proximity between different samples is measured, the matching coefficient and the consistency of the samples can be utilized. There are many ways to measure similarity. Some are used in specialized fields, and others are suitable for specific types of data. How to choose similarity measures is a rather complicated problem.

The image retrieval database is the object of the retrieval system, and the retrieval library mainly stores the feature vectors of each image through neural network calculation. The similarity score *SS* based on feature vectors is defined by (11).

$$SS(i,j) = \frac{1}{1 + sim(i,j)} \tag{11}$$

where $SS(i, j)$ is the similarity score between image $\mathbf{x}_i$ and image $\mathbf{x}_j$. $sim(i, j)$ represent the similarity degree between image i and image j, as shown in (12).

$$sim(i,j) = \frac{\mathbf{x}_i \cdot \mathbf{x}_j}{||\mathbf{x}_i|| * ||\mathbf{x}_j||} \tag{12}$$

Cosine similarity measures the angle between two vectors, and the result is represented by the cosine of the angle. It tends to give better results when the data is not very standard. Compared to distance metrics, cosine similarity pays more attention to the difference in direction between two vectors, rather than distance or length. The cosine distance is more to distinguish the difference from the direction, but not the absolute value, more used to distinguish the similarity and

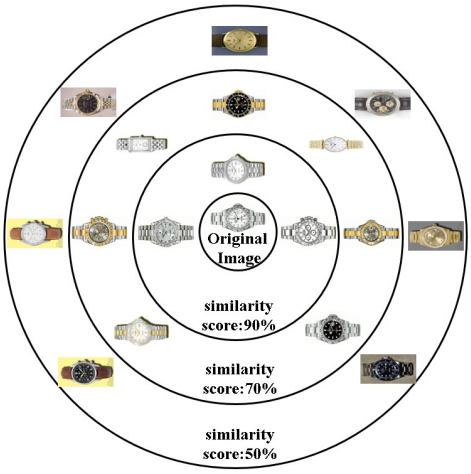


Fig. 7. Image matching dataset based on gravitational field model.

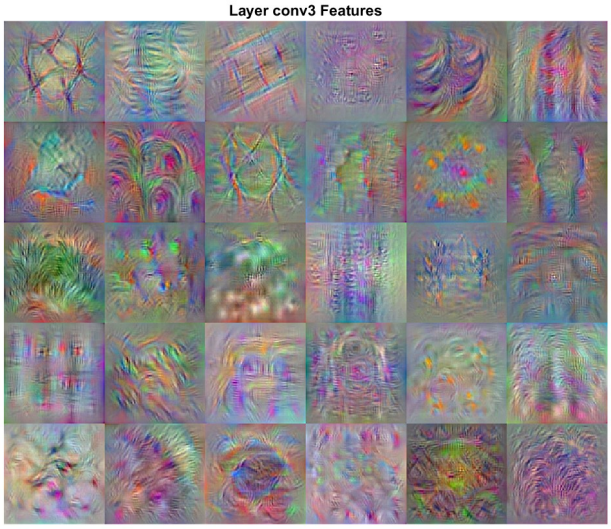


Fig. 8. Visualization of kernels in the third convolution layer.

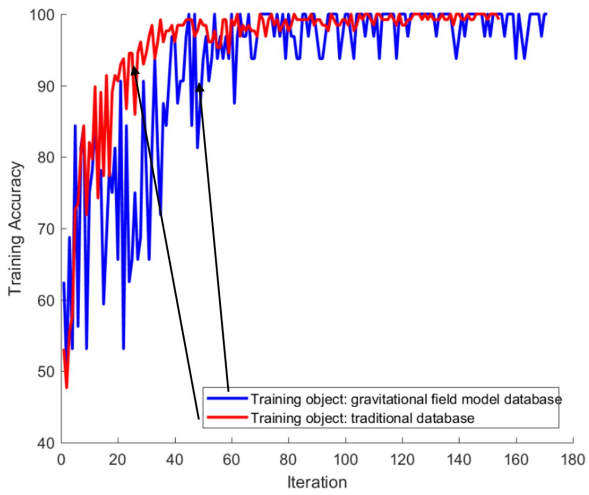


Fig. 9. Convergence curves of networks under two training methods.

difference using the representation of the feature on the content, and correct the metrics that may exist between the features are not uniform. Problem (because the cosine distance is not sensitive to absolute values).

It can be seen that when the two images are exactly the same, the similarity score is 100%. And the greater the similarity distance, the lower the score. The most important factor affecting the similarity calculation is the feature vector corresponding to image, that is, whether the trained DCNN model is accurate. The more accurate the DCNN model is, the smaller the differences between two instances with the same dimension will be, and the smaller the total distance and the higher the image matching score. Finally, the retrieval system sorts the images from large to small according to the similarity score and returns the search results in turn.

E. Discussions

The deep learning based models are just a series of simple and continuous geometric transformations of mapping vector space. Suppose there is a continuous transformation that can be learned and can be mapped from **X** to **Y**, and there are enough dense samples of **X**:**Y** as training data, and it can only map one data manifold X to another manifold **Y**. For most tasks, it may not be learned regardless of the size of the deep neural network, which means the corresponding geometric transformation may be too complex, or there is no appropriate data for learning.

The method of improving deep learning by increasing the number of layers and increasing the amount of training data is insufficient, and cannot solve the extreme limited essential problem that the deep learning model shows. Moreover, most of the applications and programs we want to learn in deep learning cannot be represented as geometric transformation of the continuous data manifold.

Deep learning allows the network model to directly learn the representations of the image, which greatly reduces the error caused by manually extracting the features of the image. The convolutional neural network in deep learning is the most commonly used technology because it is relatively simple and efficient, and the extracted features are more accurate than the traditional feature extraction algorithms. The use of deep CNN not only allows the model to learn the representation of the training sample data through the hidden layer, but also allows the model to learn the representation of the image. By adding a hidden attribute to deep CNN model, this model can not only use domain-dependent image representation, but also learn a series of hash functions. This not only ensures the accuracy of image retrieval, but also improves the speed of image retrieval.

IV. Differential Learning Process Based on Gravitational Field Model

In this section, we first introduce the training dataset based on the gravitational field that used for the training process in content-based image retrieval system. Then we introduce the differential learning details of the algorithm.

A. Image Dataset Based on Gravitational Field

Traditional datasets only contain the category labels and corresponding feature points between images. We fuse the current image matching dataset and manually clean up the similarity score labels for each group of images, and finally build the gravitational field dataset model based on this label. The entire dataset consists of 20 000 images of 200 categories.

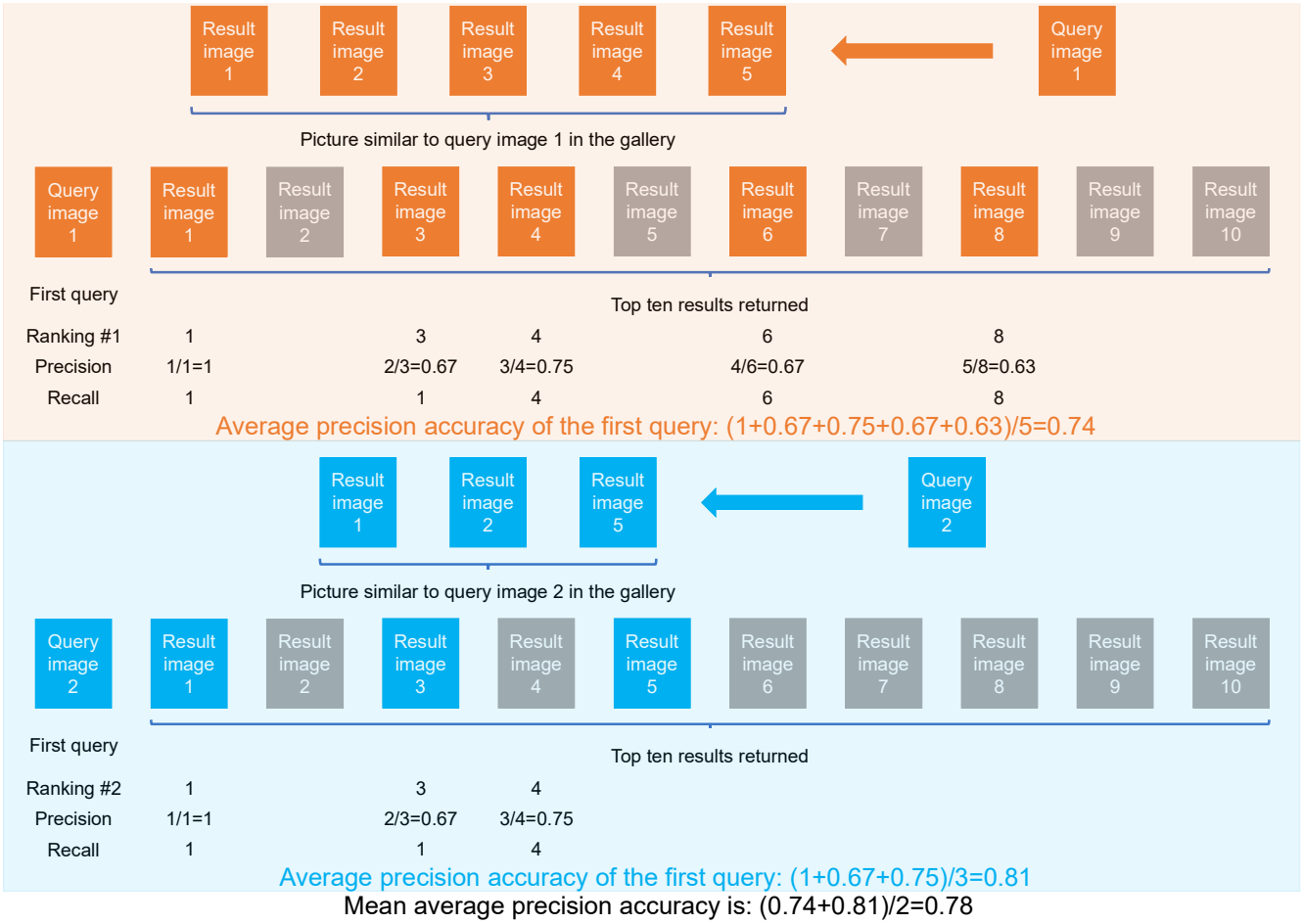


Fig. 10. Evaluation process of content-based image retrieval system.

As shown in Fig. 7, the loss function is adjusted according to the similarity score label, when the samples at each level enter the DCNN. The higher the similarity score is, the more effective the system can be trained. Two completely different images do not cause any help for the convergence of the system.

B. Training and Testing Process

In the image input phase, since we use the batch gradient descent method to train the whole network, each batch of images belong to the same category label. In the process of model training, the training time is relatively long due to the huge parameters of the neural network. So the performance of the machine has more obvious influence on the training time. In the experimental process, we use GPU parallel operation to accelerate it. We build the workstation with the NVIDIA GTX-1080, the processing speed is about 400 times of CPU I5-6600K. We observe the convergence of the network by visualizing some kernels in the third convolution layer after the completion of training, as shown in Fig. 8. It can be seen that the DCNN can extract the underlying edge, color features and high-level semantic features. At the same time, the addition of similarity score labels help to speed up the convergence of networks. As shown in Fig. 9, the DCNN trained on the gravitational field dataset converges early than network trained on general dataset by 20 batches. During the test phase, the image input does not require any label, and the final output is the similarity score according to (11) in Section III.

In order to reduce the overfitting problem caused by deep learning, we employ two data extension techniques. The first strategy is to rotate the image and revert the format to improve the robustness of the model. Another strategy is to increase the random number of pixels in the data set in order to achieve the invariance of illumination and color. The effect of the algorithm will be demonstrated in section V.

V. Experimental Results and Analysis

In this section, we first introduce the evaluation method and process, then we introduce an image retrieval database for validating the effectiveness of the algorithm and parameters setting in the training process. Extensive experiments are presented to demonstrate the image retrieval performance of the proposed algorithm, and a large number of analyses are used to illustrate the significance of the algorithm.

A. Evaluation Method and Process

We first explain the evaluating method of the content-based image retrieval system and the processing flow, as shown in Fig 10.

In the CRIR system, the system returns to the user a sorted list based on similarity according to the sorting results. The precision accuracy P_N of the first N results is usually adopted, which is defined as a search process for a sample image with $q_i \in R$ where R represents the certain image set with a specific semantic meaning. The purpose of the user submitting q_i is to retrieve R. The collection returns the top N results of the

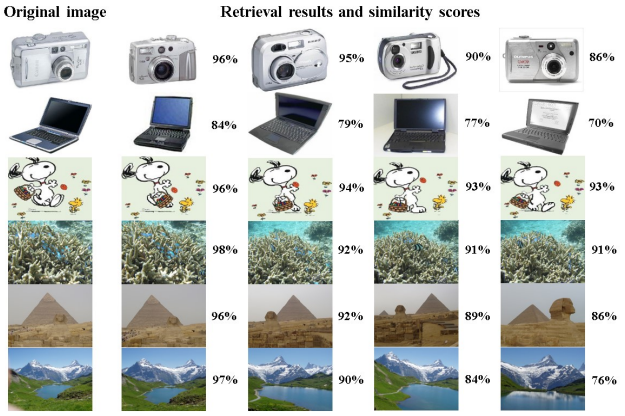


Fig. 11. Partial retrieval results on Caltech-101 and Holidays datasets. The output images are the retrieval results and the similarity score represents the matching degree.

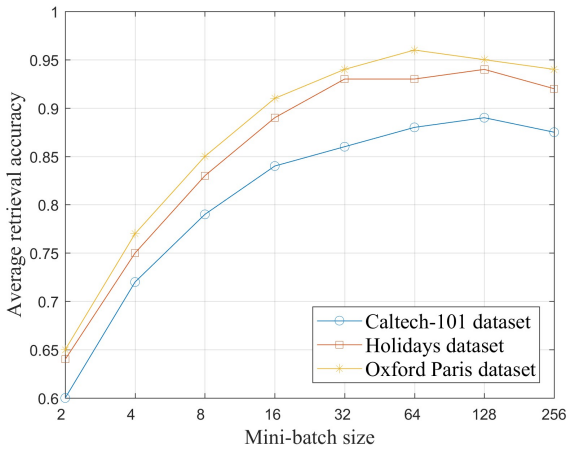


Fig. 12. Relationships between the retrieval accuracy and the mini-batch size on three image retrieval benchmarks.

TABLE I
RETRIEVAL ACCURACY ON CALTECH101 DATASET

Number	watch	pizza	brain	camera	cup	overall
10	0.92	1.00	0.96	0.79	0.97	0.915
20	0.90	0.99	0.96	0.76	0.96	0.908
30	0.90	0.99	0.96	0.76	0.96	0.901
40	0.89	0.99	0.94	0.75	0.94	0.893
50	0.86	0.97	0.90	0.75	0.93	0.885

TABLE II
RETRIEVAL ACCURACY ON HOLIDAYS DATASET

Number	sea	hill	pyramid	house	silva	overall
10	0.97	0.84	0.99	0.99	0.90	0.964
20	0.94	0.82	0.98	0.99	0.89	0.962
30	0.93	0.82	0.97	0.99	0.89	0.956
40	0.92	0.81	0.95	0.98	0.88	0.944
50	0.92	0.79	0.93	0.97	0.86	0.941

TABLE III
RETRIEVAL ACCURACY ON OXFORD PARIS DATASET

Landmarks	10	20	30	40	50
La Defense Paris	0.97	0.97	0.96	0.95	0.95
Eiffel Tower Paris	0.92	0.92	0.90	0.90	0.88
Louvre Paris	0.93	0.92	0.90	0.89	0.89
Pantheon Paris	0.88	0.87	0.87	0.85	0.84
Pompidou Paris	0.98	0.96	0.95	0.95	0.93
Overall	0.981	0.974	0.965	0.952	0.948

TABLE IV
COMPARISON RESULTS OF AVERAGE RETRIEVAL ACCURACY ON THREE DATASETS UNDER DIFFERENT ACTIVATION FUNCTION

Activation Function	Caltech-101	Holidays	Oxford Paris
Arctan function	0.84	0.88	0.88
Bent identity function	0.85	0.90	0.91
Softplus function	0.88	**0.94**	0.95
Softsign function	0.87	0.92	0.94
Tanh function	0.85	0.89	0.91
PReLU function	**0.89**	**0.94**	**0.96**

system, *i.e.*, $p_j, j = 1, 2, \ldots, N$. Then the precision accuracy is defined as

$$P_N(q_i) = \sum_{k=1}^{N} \frac{\psi(p_k, R)}{N} \tag{13}$$

where

$$\psi(x, Y) = \begin{cases} 1, & if\ x \in Y \\ 0, & if\ x \notin Y \end{cases}$$

Then, the average precision accuracy for all test samples is defined as

$$P_N = \sum_{i=1}^{Total\_Queary\_Count} \frac{(p_N(q_i))}{Total\_Queary\_Count} \tag{14}$$

The definition of the above precision accuracy simply shows how many of the top N results returned are correct.

We first select 10% of the total images to be queried as test samples, then use the model to search and output the top ten images as the retrieved results. Then we calculate the average precision accuracy for each time according to the method shown in Fig. 10, and finally get the mean average precision accuracy. A number of literatures [15] [22] [37] have shown that this avoids the interference caused by the training process and accurately evaluates the true performance of the model.

B. Image Datasets and Experimental Setup

To evaluate the effect of the training model, we tested the trained model on the Caltech-101 [14], Holidays [15], and Oxford Paris [51] datasets. The Caltech-101 dataset consists of 101 categories, and there are about 40 to 800 images per category. The Holidays dataset includes a set of images which mainly contains some of our personal holidays photos. The dataset contains 500 image groups, each of which represents a distinct scene or object. The Oxford Paris dataset consists of 6412 image samples collected from "Flickr" by searching for particular Paris landmarks. We randomly selected 10 classes in three datasets to observe the performance of content-based image retrieval system.

The improved deep convolutional neural network is trained by mini-batch stochastic gradient descent with a momentum (0.6) and weight decay (0.0001). The initial learning rate is set to 0.01 and is reduced by ten times every 50 epochs. Mini-batch size of Caltech-101, Holidays and Oxford Paris datasets are 128, 128 and 64, respectively. If the validation error is not reduced in the 200 epochs, the training process is then early stopped.

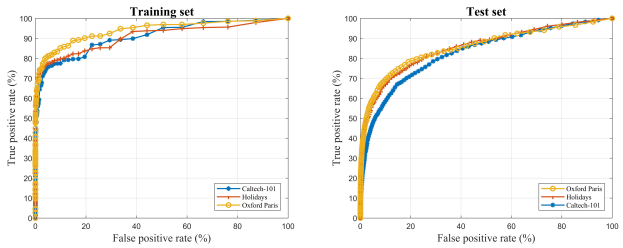


Fig. 13. ROC curves of the training set and test set of three image retrieval benchmarks.

The computer used in the experiments is configured as a six-core I7-8750CPU, a graphics card of the NVIDIA GTX 1050Ti, 16G memory, and a 1T solid state drive. And the experimental platform is MATLAB 2017a under windows10 system.

C. *Evaluation of Image Retrieval Results*

In this section, we count and compare the search results of multiple algorithms to illustrate the effectiveness of the algorithm. And from the perspective of time complexity, the computation performance of proposed algorithm in practical applications is observed.

The partial retrieval results and similarity scores obtained on Caltech-101 and Holidays datasets are presented in Fig. 11. The above three rows in Fig. 11 are the retrieving results of Caltech-101 dataset, while the bottom three rows represent the retrieving results of Holidays dataset. It can be seen from the figure that features extracted by DCNN not only retrieves images of specific objects such as boats, dolphins and lotus flowers, but also achieves excellent results in the retrieval of scene images.

In order to describe the accuracy of system retrieval, we define the retrieval accuracy *RA* in (15).

$$RA = \frac{RI}{TI} \tag{15}$$

where *RI* and *TI* represent the number of relevant images and total image number in an image retrieval process, respectively. We searched 10 classes of images separately, and retrieved 10, 20, 30, 40 and 50 images each time. The average accuracy of 20 retrieval results is taken as the retrieval accuracy of each kind. And the total retrieval results take the average value of the accuracy of 10 kinds of images. As shown in Table I, Table II and Table III, the overall retrieval accuracy are 88.5%, 94.1% and 96.2%, respectively, when the number of retrieved images is 50. A, B, C, D and E in Table I represent the five categories of watch, pizza, brain, camera and cup, respectively. A, B, C, D, and E in Table II represent the five categories of sea, mountains, pyramid, houses and forests, respectively. A, B, C, D, and E in Table III represent the five categories of La Defense Paris, Eiffel Tower Paris, Louvre Paris, Pantheon Paris, Pompidou Paris, respectively. From the view of overall effect, the increase of the number of returned images does not reduce the retrieval accuracy greatly, and the retrieval effect tends to be stabilized with the increase of the number of returned images.

To observe the impact of hyperparameters on algorithm performance, we observed the effect of batch size on retrieval accuracy on three databases. The experimental results are shown in Fig. 12. It can be seen that as the batch size increases, the retrieval accuracy is greatly improved. However, when the specific threshold is exceeded, the retrieval accuracy begins to stabilize or even to decrease. The results show that the selection of the appropriate batch size has an overall effect on the performance of the content-based image retrieval system, and a good batch size setting can significantly improve the performance of the retrieval system. At the same time, our algorithm has good robustness to large batch sizes and the algorithm performance is stable. When the sizes of mini-batch are 128, 128 and 64, the retrieval accuracy of Caltech-101, Holidays and Oxford Paris datasets are the highest, which shows that the mini-batch size of 64 or 128 is usually a good choice.

The choice of activation function is another reason that affects the performance of the algorithm. Different activation functions have different properties and play different roles in network training. Therefore, we compared the image retrieval accuracy of different activation functions, as shown in Table IV. The PReLU function has achieved the best performance on three datasets: 0.89 on Caltech-101, 0.94 on Holidays and 0.96 on Oxford Paris dataset. And comparing the results of multiple activation functions, we can see that the activation

TABLE V
COMPARISON RESULTS OF AVERAGE RETRIEVAL ACCURACY ON THREE DATASETS UNDER DIFFERENT ALGORITHM

Algorithm	Caltech-101	Holidays	Oxford Paris
Color based method [16]	0.75	0.78	0.79
Texture based method [17]	0.81	0.83	0.85
Fused feature [33]	0.82	0.85	0.86
Tag completion [28]	0.84	0.91	0.93
Hoi *et al.* [29]	0.79	0.85	0.86
Bian *et al.* [35]	0.85	0.92	0.92
Lai *et al.* [36]	0.84	0.90	0.91
Adaptive network[45]	0.84	0.88	0.90
Boosting framework [30]	0.87	0.92	0.95
Wavelet optimization [32]	0.81	0.84	0.88
Our method	**0.89**	**0.94**	**0.96**

TABLE VI
COMPARISON RESULTS OF QUERY TIME UNDER DIFFERENT METHODS

Method	Platform	Query time of 50 images (s)
Color based method [16]	C++	20.85
Texture based method [17]	C++	22.97
Fused feature [33]	MATLAB	38.54
Tag completion [28]	**C++**	**19.43**
Hoi *et al.* [29]	Python	22.71
Bian *et al.* [35]	Python	19.51
Lai *et al.* [36]	C++	23.37
Adaptive network [45]	MATLAB	40.54
Boosting framework [30]	MATLAB	42.91
Wavelet optimization [32]	MATLAB	34.44
Our method	MATLAB	30.27

TABLE VII
COMPARISON RESULTS OF AVERAGE RETRIEVAL ACCURACY ON THREE DATASETS UNDER DIFFERENT COMBINED ACTIVATION FUNCTION

Algorithms	Caltech-101	Holidays	Oxford Paris
Arctan+ Softplus	0.85	0.91	0.90
Bent identity+ Softplus	0.83	0.90	0.87
Softplus+ Softsign	0.83	0.88	0.91
Softsign+ Tanh	0.87	0.92	0.92
Tanh+ Bent identity	0.82	0.90	0.88
PReLU+ Softplus	**0.91**	**0.95**	**0.94**

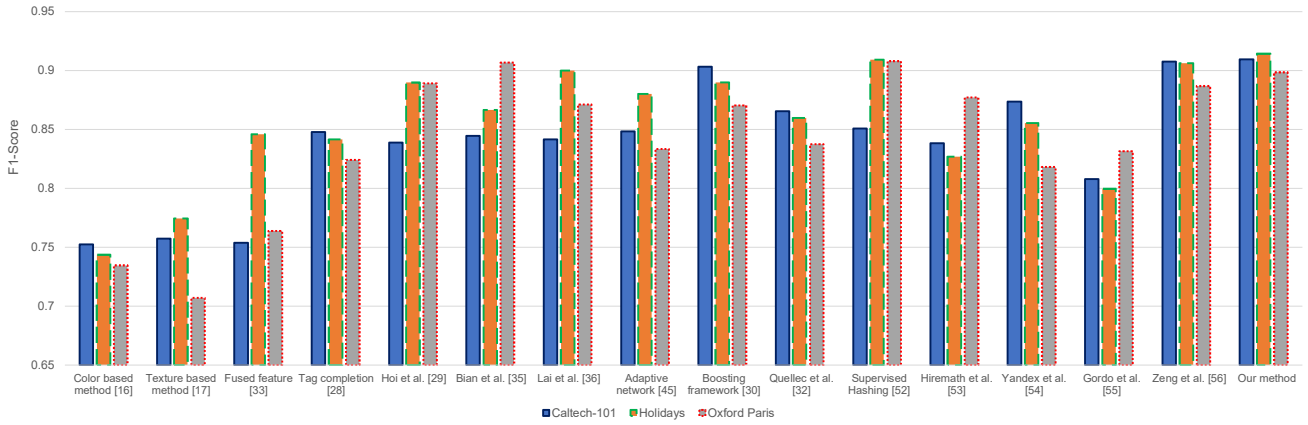


Fig. 14. Comparison of F_1-score of various algorithms on three image retrieval benchmarks. The first, second, and third column of algorithms represent the scores of Caltech-101, Holidays, and Oxford Paris, respectively.

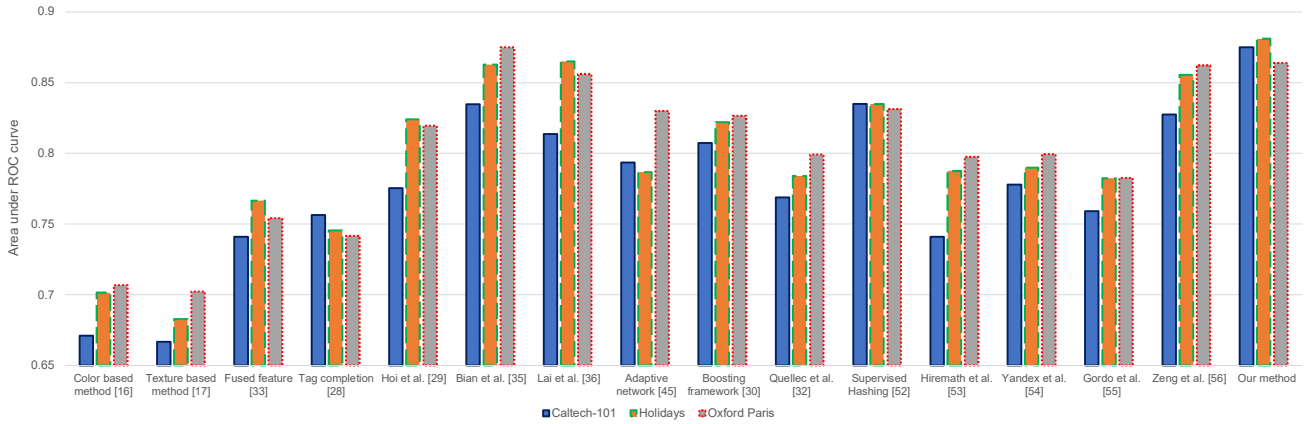


Fig. 15. Comparison of AUC values of various algorithms on three image retrieval benchmarks. The first, second, and third column of algorithms represent the AUC value of Caltech-101, Holidays, and Oxford Paris, respectively.

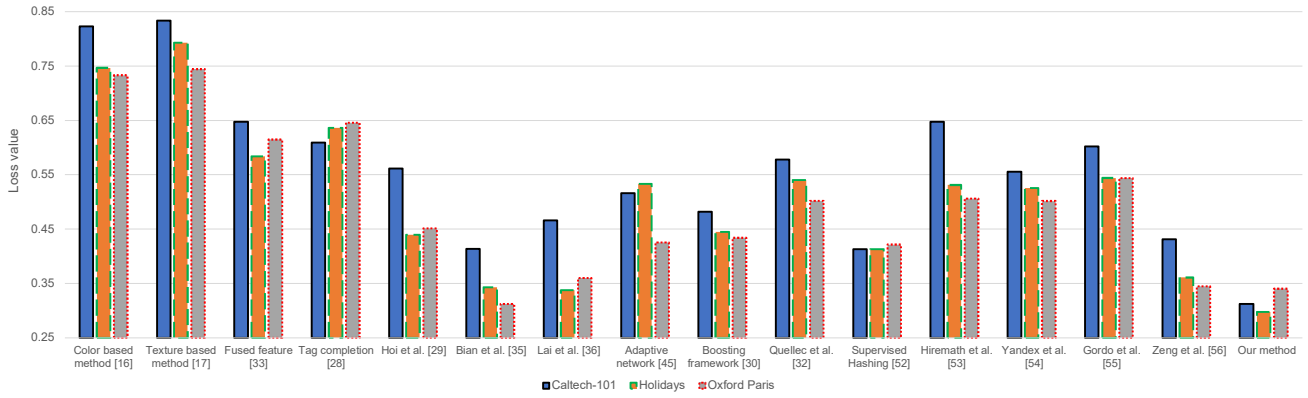


Fig. 16. Comparison of loss values of various algorithms on three image retrieval benchmarks. The first, second, and third column of algorithms represent the loss values of Caltech-101, Holidays, and Oxford Paris, respectively.

function has only little effect on the retrieval accuracy of the system. In the practical applications, researchers can choose corresponding activation functions according to the specific circumstances.

We compared the retrieval accuracy of each algorithm on three datasets, as shown in Table V. It can be seen that our method achieves the accuracy of 89%, 94% and 96% on the Caltech-101, Holidays and Oxford Paris, respectively. Then we compare these two image retrieval methods based on the two underlying features of color [16] and texture [17]. In the color based feature extraction process, the images are firstly transformed from RGB space into HSV space, and unequal interval quantization operation is performed. The three color components are then represented as one-dimensional vectors and eventually retrieved according to color based feature. The extraction of texture based features is based on the gray level co-occurrence matrix. The mean value and the standard deviations of the four parameters (energy, entropy, moment of inertia, correlation) are used as the final eight dimensional texture features. Table V shows the overall retrieval accuracy of our approach and these two methods over three datasets. It can be seen that our algorithm has the highest image retrieval accuracy. Our algorithm improves the retrieval accuracy rate by 14%, 16%, 17% and 7%, 11%, 11% on three datasets

compared to color feature based method and texture feature based method respectively. Fused features can improve the performance of above two methods and achieved the accuracy of 82%, 85%, and 86% on three benchmarks, respectively. Boosting framework showed the sub-optimal results with the average precision accuracy of 87%, 92%, and 95% on three benchmarks. By contrast, our approach has increased by 5%, 7% and 9% on three benchmarks, respectively.

In order to observe the performance of the algorithm in practical applications, we collect and compare the retrieval speed of different algorithms, as shown in Table VI. The implementation environment of the algorithm includes C++, Python and MATLAB. Due to the influence of environment, the retrieval speed of different algorithms cannot be directly used for comparison, but it can reflect the speed of algorithm operation. The retrieval speed of our proposed method is 30.3 seconds per fifty images. The fastest search method is the tag completion algorithm [28] in C++ environment, with a speed of 19.43 seconds per fifty images. It can be seen that the retrieval speed of the algorithm in the C++ environment is faster than that of the MATLAB environment, which is due to the characteristics of the compiler environment. In the future, we plan to transplant the algorithm to the C++ platform to complete image retrieval process and apply it to the actual environment.

To verify the validity and progressiveness of the combined activation functions used in the deep CNN model, we try to compare the retrieval performance of multiple combination of activation functions, as shown in Table VII. It can be clearly seen that the combination of PReLU function and Softplus function achieves the best average precision accuracy on both two image retrieval benchmarks (*i.e.*, 0.91 on Caltech-101, 0.95 on Holidays,), which outperforms the state-of-the-art performance with 0.02 and 0.01, respectively. However, it is worth noting that the combination of two different activation functions may has a adverse effect on the image retrieval, *e.g.*, the average precision accuracy on Oxford Paris is just 0.94, which is lower that the only usage of PReLU. In fact, PReLU only adds a very small number of parameters, which means that the computational complexity of the network and the risk of the over-fitting are only a little bit increased. In particular, when different channels use the same hyper-parameters, the number of weights are even smaller.

D. Comparison of ROC and AUC

In this section, the receiver operating characteristic (ROC) curves are obtained on the training and test set of three image retrieval benchmarks by varying the threshold, as shown in Fig. 13. We first calculate Precision rate P and Recall rate R, according to

$$P=\frac{TP}{TP+FP} \tag{16}$$

$$R=\frac{TP}{TP+FN} \tag{17}$$

where TP, FP and FN represent true positive, false positive and false negative samples. Then we can obtain the F-score, as given by

$$F=\frac{(a^2+1)P*R}{a^2(P+R)} \tag{18}$$

where a is the Harmonic parameter. F-score F is the weighted harmonic average of Precision and Recall. If $a = 1$, we can get the F_1-score to evaluate the performance of image retrieval system, as given by

$$F_1=\frac{2PR}{P+R} \tag{19}$$

F_1-score combines the results of P and R. When F_1-score is higher, the comparison shows that the experimental method is ideal. We compared the F_1-score of multiple algorithms, as shown in Fig. 14. It can be seen that our algorithm has the highest score on all three datasets, *i.e.*, the best performance. On the other hand, the F_1-score means that if the prediction probability exceeds the threshold, the image will be returned as one of the retrieval results. If the threshold is set to 0.5, the average precision accuracy of the training set are 0.97, 0.98, and 0.96 on the Caltech-101, Holidays, and Oxford Paris, respectively. And the corresponding classification accuracies of the test set of Caltech-101, Holidays, and Oxford Paris are 0.9, 0.93, and 0.94, respectively. The areas under the ROC curve (AUC) of training set are 0.94, 0.92, and 0.91 on three benchmarks. And the AUC values of all test sets are 0.88, 0.87, and 0.86, respectively. The comparison of various algorithms on AUC value and their loss are shown in Fig. 15 and Fig. 16, respectively. In fact, a higher AUC and a lower loss mean a more reliable algorithm. The AUC value of our proposed CBIR system outperforms the state-of-the-art performance with 3.7%, 2.3%, and 0.6%, respectively, which shows the better generalization ability of our algorithm on unseen cases.

E. Complexity Analysis

Suppose the class $\boldsymbol{S}_k$ has n_k vectors. And d is the dimension and $\boldsymbol{p}_i$ is the i-th vector. Category center is defined as

$$z_k=\left(\frac{1}{n_k}\sum_{i=0}^{n_k-1}p_{i1},\frac{1}{n_k}\sum_{i=0}^{n_k-1}p_{i2},\frac{1}{n_k}\sum_{i=0}^{n_k-1}p_{id}\right) \tag{20}$$

If $\boldsymbol{p}_{n+1}$ is a vector out of the category, then we can get

$$\begin{aligned}F_a(\boldsymbol{S}_k)&=\frac{1}{n_k}\sum_{i=0}^{n_k-1}\sum_{j=0}^{d-1}\left(p_{ij}^2+\frac{1}{n_k^2}\sum_{k=0}^{n_k-1}p_{kj}\right)^2-\frac{2}{n_k}p_{ij}\sum_{k=0}^{n_k-1}p_{kj}\\&=\frac{1}{n_k}\sum_{i=0}^{n_k-1}\sum_{j=0}^{d-1}p_{ij}^2+\frac{1}{n^2}\sum_{i=0}^{n_k-1}\sum_{j=0}^{d-1}\left(\sum_{k=1}^{n_k-1}p_{kj}\right)^2-\frac{2}{n_k^2}\sum_{i=0}^{n_k-1}\sum_{j=0}^{d-1}p_{ij}\sum_{k=1}^{n_k-1}p_{kj}\end{aligned} \tag{21}$$

$$\begin{aligned}F_a(\boldsymbol{S}_k\cup\boldsymbol{p}_{n_k+1})&=\frac{1}{n_k+1}\sum_{i=0}^{n_k}\sum_{j=0}^{d-1}p_{ij}^2+\frac{1}{(n_k+1)^2}\sum_{i=0}^{n_k-1}\sum_{j=0}^{d-1}\left(\sum_{k=0}^{n_k}p_{kj}\right)^2\\&\quad-\frac{2}{(n_k+1)^2}\sum_{i=0}^{n_k}\sum_{j=0}^{d-1}p_{ij}\sum_{k=0}^{n_k}p_{kj}\end{aligned} \tag{22}$$

If the time complexity of the contribution of a single vector in a certain class is calculated as O(d), then the time complexity of the contribution for all N vectors in each of the K clusters is O(NKd). This is the same as the time complexity of intra-class or inter-class dispersion for initializing and calculating one step, and the same complexity for one iteration of K-means. However, since the algorithm considers both intra-class and inter-class similarities, the total complexity is approximated by K-means because of fewer iterations.

VI. Conclusion

Faced with the huge image data in the context of big data era, how to effectively manage, describe, and retrieve them has become a hotspot issue in academia and industry. In this paper, we present an end-to-end image retrieval system based on gravitational field deep learning. The algorithm uses a gravitational field database with similarity score labels to train deep CNN for feature extraction, and finally outputs the image retrieval results. Experiments have tested the retrieval performance of the algorithm under three databsets, and the retrieval accuracy has reached 88.5%, 94.1% and 96.2%, respectively, which is superior to traditional algorithms, such as color feature and texture based image retrieval methods. Moreover, our proposed CBIR system outperforms a large number of state-of-the-art methods, including the AUC value and loss value. Furthermore, when the number of returned images increases, the retrieval accuracy decreases slowly and tends to be stable. And differential learning based retrieval method is the end-to-end system, which does not require the human participation and prior knowledge. Therefore, it is especially suitable for dealing with large amounts of data. The development of deep learning in image retrieval is in the ascendant, and there is great room for it in the future.

The current problem of the algorithm is that it is difficult to obtain the samples with similarity score, and the construction of gravitational field dataset requires a great deal of time and effort. In the future, we plan to try to use the semi-supervised approach to train the entire neural network. On the other hand, the processing speeds in the training and testing phases need to be further improved.

References

[1] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," International Journal of Computer Vision, vol. 60, no. 2, pp. 91-110, 2004.

[2] H. Bay, T. Tuytelaars, and L. V. Gool, "SURF: Speeded Up Robust Features," Computer Vision & Image Understanding, vol. 110, no.3, pp. 404-417, 2006.

[3] G. Sheikholeslami, A. Zhang, and L. Bian, "A Multi-Resolution Content-Based Retrieval Approach for Geographic Images," Geoinformatica, vol. 3, no. 2, pp. 109-139, 1999.

[4] L. Lu, R. Liu, and N. Liu, "Remote Sensing Image Retrieval Using Color and Texture Fused Features," Journal of Image & Graphics, vol. 9, no. 3, pp. 328-333, 2004.

[5] R. Zhao and W. I. Grosky, "Negotiating the Semantic Gap: from Feature Maps to Demantic Landscapes", Pattern Recognition, vol. 35, no. 3, pp. 593-600, 2002.

[6] X. Zhou and T. Huang, "Relevance Feedback in Image Retrieval: A Comprehensive Review," Multimedia Systems, vol. 8, no. 6, pp. 536-544, 2003.

[7] A. Mojsilovic and B. Rogowitz, "Capturing Image Semantics with Low-level Descriptors," IEEE International Conference on Image Processing, Thessaloniki, Greece, pp. 18-21, 2001.

[8] L. Zhao, P. Tang, and L. Huo, "Land-Use Scene Classification Using a Concentric Circle-Structured Multiscale Bag-of-Visual-Words Model," IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing, vol. 7, no. 12, pp. 4620-4631, 2015.

[9] K Lin, H. Yang, JH. Hsiao *et al.*, "Deep Learning of Binary Hash Codes for Fast Image Retrieval," IEEE Computer Vision and Pattern Recognition Workshops, Boston, USA, pp. 27-35, 2015.

[10] J. Wan, D. Wang, H. Hoi *et al.*, "Deep Learning for Content-Based Image Retrieval: A Comprehensive Study," in Proceedings of the 22nd ACM international conference on Multimedia, Florida, USA, pp. 157-166, 2014.

[11] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition", arXiv preprint, arXiv: 1409. 1556, 2014.

[12] B. Xu, N. Wang, T. Chen *et al.*, "Empirical Evaluation of Rectified Activations in Convolutional Network", arXiv preprint, arXiv: 1505. 00853, 2015.

[13] X. Glorot, A. Bordes, and Y. Bengio, "Deep Sparse Rectifier Neural Betworks," Journal of Machine Learning Research, vol. 12, pp. 315-323, 2011.

[14] T. Kinnunen, J. K. Kamarainen, L. Lensu *et al.*, "Making Visual Object Categorization More Challenging: Randomized Caltech-101 Data Set," IEEE International Conference on Pattern Recognition, Istanbul, Turkey, pp. 476-479, 2010.

[15] H. Jegou, M. Douze, and C. Schmid, "Hamming Embedding and Weak Geometric Consistency for Large Scale Image Search," European Conference on Computer Vision, Marseille, France, pp. 304-317, 2008.

[16] C. Liu, W. Hsaio, C. H. Lee *et al.*, "An HMM-Based Algorithm for Content Ranking and Coherence-Feature Extraction," IEEE Transactions on Systems Man & Cybernetics Part B, vol. 43, no. 2, pp. 440-450, 2013.

[17] R. M. Haralick, "Texture Features for Image Classification," IEEE Transactions on Systems Man & Cybernetics, vol. 3, no. 6, pp. 610-621, 1973.

[18] Y. Gong, S. Lazebnik, A. Gordo *et al.*, "Iterative Quantization: A Procrustean Approach to Learning Binary Codes for Large-scale Image Retrieval," IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 35, no. 12, pp. 2916-2929, 2013.

[19] Q. Zheng, M. Yang, Q. Zhang, and J. Yang, "A Bilinear Multi-Scale Convolutional Neural Network for Fine-grained Object Classification," IAENG International Journal of Computer Science, vol. 45, no. 2, pp. 340-352, 2018.

[20] Q. Zheng, M. Yang *et al.*, "Improvement of Generalization Ability of Deep CNN via Implicit Regularization in Two-Stage Training Process," IEEE Access, vol. 6, pp. 15844-15869, 2018.

[21] Q. Zheng, M. Yang *et al.*, "Understanding and Boosting of Deep Convolutional Neural Network Based on Sample Distribution," in IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference, Chengdu, China, pp. 823-827, 2017.

[22] S. Murala, R. P. Maheshwari, and R. Balasubramanian, "Local Tetra Patterns: A New Feature Descriptor for Content-Based Image Retrieval," IEEE Transactions on Image Processing, vol. 21, no. 5, pp. 2874-2886, 2012.

[23] C. B. Akgül, D. L. Rubin, S. Napel *et al.*, "Content-Based Image Retrieval in Radiology: Current Status and Future Directions," Journal of Digital Imaging, vol. 24, no. 2, pp. 208-222, 2011.

[24] M. Eitz, K. Hildebrand, T. Boubekeur *et al.*, "Sketch-Based Image Retrieval: Benchmark and Bag-of-Features Descriptors," IEEE Transactions on Visualization & Computer Graphics, vol. 17, no. 11, pp. 1624-1636, 2011.

[25] G. H. Liu, Z. Y. Li, L. Zhang *et al.*, "Image Retrieval Based on Micro-structure Descriptor," Pattern Recognition, vol. 44, no. 9, pp. 2123-2133, 2011.

[26] G. H. Liu, L. Zhang, Y. K. Hou *et al.*, "Image Retrieval Based on Multi-texton Histogram," Pattern Recognition, vol. 43, no. 7, pp. 2380-2389, 2010.

[27] C. H. Wei, Y. Li, W. Y. Chau *et al.*, "Trademark Image Retrieval Using Synthetic Features for Describing Global Shape and Interior Structure," Pattern Recognition, vol. 42, no. 3, pp. 386-394, 2010.

[28] L. Wu, R. Jin, and A. Jain, "Tag Completion for Image Retrieval," IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 35, no. 3, pp. 716-727, 2013.

[29] S. C. H. Hoi, W. Liu, and S. F. Chang, "Semi-supervised Distance Metric Learning for Collaborative Image Retrieval and Clustering," Acm Transactions on Multimedia Computing Communications & Applications, vol. 6, no. 3, pp. 1-26, 2010.

[30] L. Yang, R. Jin, L. Mummert *et al.*, "A Boosting Framework for Visuality-preserving Distance Metric Learning and Its Application to Medical Image Retrieval," IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 32, no. 1, pp. 30-44, 2010.

[31] G. H. Liu and J. Y. Yang, "Content-based Image Retrieval Using Color Difference Histogram," Pattern Recognition, vol. 46, no. 1, pp. 188-198, 2013.
[32] G. Quellec, M. Lamard, G. Cazuguel *et al.*, "Wavelet Optimization for Content-based Image Retrieval in Medical Databases," Medical Image Analysis, vol. 14, no. 2, pp. 227-241, 2010.
[33] J. Yue, Z. Li, L. Liu *et al.*, "Content-based Image Retrieval Using Color and Texture Fused Features," Mathematical & Computer Modelling, vol. 54, no. 3, pp. 1121-1127, 2011.
[34] Y. Choi and E. M. Rasmussen, "Searching for Images: The Analysis of Users' Queries for Image Retrieval in American History," Journal of the Association for Information Science & Technology, vol. 54, no. 6, pp. 498-511, 2014.
[35] W. Bian and D. Tao, "Biased Discriminant Euclidean Embedding for Content-Based Image Retrieval," IEEE Transactions on Image Processing, vol. 19, no. 2, pp. 545-554, 2010.
[36] C. C. Lai and Y. C. Chen, "A User-Oriented Image Retrieval System Based on Interactive Genetic Algorithm," IEEE Transactions on Instrumentation and Measurement, vol. 60, no. 10, pp. 3318-3325, 2011.
[37] M. Ferecatu, N. Boujemaa, and M. Crucianu, "Semantic Interactive Image Retrieval Combining Visual and Conceptual Content Description," Multimedia Systems, vol. 13, no. 5, pp. 309-322, 2008.
[38] M. Cheng, N. Mitra, X. Huang *et al.*, "Global Contrast Based Salient Region Detection," IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 37, no. 3, pp. 409-416, 2011.
[39] D. Carvalho and A. Freitas, "A hybrid decision tree/genetic algorithm method for data mining," Information Sciences, vol. 163, no. 1, pp. 13-35, 2013.
[40] V. K. Lowanshi, S. Shrivastava, V. Richhariya *et al.*, "An Efficient Approach for Content based Image Retrieval using SVM, KNN-GA as Multilayer Classifier," International Journal of Computer Applications, vol. 107, no. 21, pp. 43-48, 2014.
[41] J. He, M. Li, H. Zhang *et al.*, "Generalized Manifold-Ranking-Based Image Retrieval," IEEE Transactions on Image Processing, vol. 15, no. 10, pp. 3170-3177, 2006.
[42] Y. Liu, D. Zhang, G. Lu *et al.*, "A Survey of Content-based Image Retrieval with High-level Semantics," Pattern Recognition, vol. 40, no. 1, pp. 262-282, 2007.
[43] I. Biederman, "Recognition-by-components: A Theory of Human Image Understanding," Psychological Review, vol. 94, no. 2, pp. 115-147, 1987.
[44] M. Koch, M. Ring, F. Otto *et al.*, "Combination of Statistical and Semantic Data Sources for the Improvement of Software Engineering Courses," Acm Communications in Computer Algebra, vol. 44, no. 3, pp. 87-88, 2014.
[45] S. G. Wysoski, L. Benuskova, and N. Kasabov, "Fast and Adaptive Network of Spiking Neurons for Multi-view Visual Pattern Recognition," Neurocomputing, vol. 71, no. 13, pp. 2563-2575, 2008.
[46] W. Gao, L. Zhu, and Y. Guo, "Multi-dividing Infinite Push Ontology Algorithm," Engineering Letters, vol. 23, no. 3, pp. 132-139, 2015.
[47] T. Shi, X. Sun, Z. Xia, L. Chen, and J. Liu, "Fall Detection Algorithm Based on Triaxial Accelerometer and Magnetometer," Engineering Letters, vol. 24, no. 2, pp. 157-163, 2016.
[48] L. Wang and L. M. Wang, "Global Exponential Stabilization for Some Impulsive T-S Fuzzy Systems with Uncertainties," IAENG International Journal of Applied Mathematics, vol. 47, no. 4, pp. 425-430, 2017.
[49] Z. Xia, C. Yuan, X. Sun, D. Sun, and R. Lv, "Combining Wavelet Transform and LBP Related Features for Fingerprint Liveness Detection," IAENG International Journal of Computer Science, vol. 43, no. 3, pp. 290-298, 2016.
[50] M. Bahri, A. Kamal, and C. Lande, "The Quaternion Domain Fourier Transform and its Application in Mathematical Statistics," IAENG International Journal of Applied Mathematics, vol. 48, no. 2, pp. 184-190, 2018.
[51] J. Philbin, O. Chum, M. Isard *et al.*, "Lost in Quantization: Improving Particular Object Retrieval in Large Scale Image Databases," in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Alaska, USA, pp. 1-8, 2008.
[52] H. Liu *et al*., "Deep Supervised Hashing for Fast Image Retrieval," in IEEE Conference on Computer Vision and Pattern Recognition, pp. 2064-2072, 2016.
[53] P. S. Hiremath and J. Pujari, "Content Based Image Retrieval Using Color, Texture and Shape Features," in International Conference on Advanced Computing and Communications, pp. 872-876, 2016.
[54] A. B. Yandex and V. Lempitsky, "Aggregating Local Deep Features for Image Retrieval," in IEEE International Conference on Computer Vision, pp. 1269-1277, 2016.
[55] A. Gordo, J. Almazán, and J. Revaud, "Deep Image Retrieval: Learning Global Representations for Image Search," in European Conference on Computer Vision, pp. 241-257, 2016.
[56] S. Zeng *et al.*, "Image retrieval using spatiograms of colors quantized by Gaussian Mixture Models," Neurocomputing, 171(C), pp. 673-684, 2016.

Qinghe Zheng was born in Jining, Shandong, China in 1993. He received his B.S. degree from Xi'an University of Posts and Telecommunications in 2014 and M.S. degree from Shandong University in 2018. Currently, he is studying for his doctor's degree at Shandong University. His research interests include computer vision and machine learning.