IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received January 8, 2019, accepted January 24, 2019, date of publication February 21, 2019, date of current version April 1, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2898906

Fabric Image Retrieval System Using Hierarchical Search Based on Deep Convolutional Neural Network

JUN XIANG, NING ZHANG, RURU PAN, AND WEIDONG GAO
Key Laboratory of Eco-Textiles, Jiangnan University, Wuxi 214122, China

Corresponding authors: Ruru Pan (prrsw@163.com) and Weidong Gao (gaowd3@163.com)

This work was supported in part by the National Key R&D Program of China under Grant 2017YFB0309200 and in part by the Fundamental Research Funds for the Central Universities under Grant JUSRP41804.

ABSTRACT Fabric image retrieval is a meaningful issue, due to its potential values in many areas such as textile product design, e-commerce, and inventory management. Meanwhile, it is challenging because of the diversity of fabric appearance. Encourage by the recent breakthrough in the deep convolutional neural network (CNN), a deep learning framework is applied for fabric image retrieval. The idea of the proposed framework is that the binary code and feature for representing the image can be learning by a deep CNN when the data labels are available. The proposed framework employs a hierarchical search strategy that includes coarse-level retrieval and fine-level retrieval. Otherwise, a large-scale wool fabric image retrieval dataset named WFID with about 20 000 images are built to validate the proposed framework. The longitudinal comparison experiments for self-parameter optimization and horizontal comparison experiments for verifying the superiority of the algorithm are performed on this data set. The comparison experimental results indicate the superiority of the proposed framework.

INDEX TERMS Image retrieval, wool, fabric, feature extraction, machine learning, neural networks.

I. INTRODUCTION

In recent years, fabric image retrieval attracted the attention of more and more researchers due to its potential values in not only visual tasks, such as fabric identification, and recommendation system, but also in e-commerce, inventory management and textile product design. Traditional keyword-based image retrieval (KBIR) technology is now very mature and widely used in textile factory. The data in the KBIR database is generally marked by humans. This process couldn't be applied to the retrieval system with large-scale database. With the reason that the KBIR cannot solve the subjectivity of labeling personal content perception and description, it is more difficult to adapt to the emergence of a large amount of new data. This paper mainly studies fabric image retrieval based on image content, which is known as Content Based Image Retrieval (CBIR) [1].

General method of image retrieval (CBIR) involved two key components: 1) design a robust feature extraction algorithm for representing images; 2) choose a suitable distance or similarity calculation method for the image representation extracted by the algorithm. Traditional feature extraction method are usually use hand-crafted image descriptions, such as SIFT [2], GIST [3], Bag of Word [4], VALAD [5], and Fisher Vector(FV) [6], [7]. Although having achieved certain success in CBIR, these methods based on low-level feature depend heavily on feature extraction engineering which leads to their limitations.

Significant breakthrough has been achieved on image analysis by moving from the early low-level feature based algorithms to deep learning based end-to-end frameworks. Even through, the most challenging issues is still associating the low-level features based on pixel-level information to high-level semantic features from human perception, which is called "semantic gap". Despite some hand-crafted feature extraction algorithms based on low-level feature had been proposed to represent the image, the performance of these visual feature descriptors on image retrieval still has limitations until recent breakthroughs in deep learning in the field of image analysis. Recently, many methods were proposed to learn the image representation based on deep convolution neural network referred to as CNN. E. g, Krizhevsky *et al.* [8],

The associate editor coordinating the review of this manuscript and approving it for publication was Chang-Tsun Li.

proposed a method for image retrieval, where the feature extraction method used is based on a 7-layer convolutional neural network, and demonstrated a good performance on ImageNet [8]–[10]. However, the disadvantage of this method is that the dimension of the feature vector has a high dimension of 4096, which makes the calculation of similarity very large. Babenko *et al.* [11] proposed a method for compressing and reducing the features of CNN using PCA and achieved good performance.

To solve this problem, many methods using Approximate Nearest Neighbor referred to as ANN or HASH based techniques for speed-up [12]–[18] have been presented. The principle of acceleration of these methods is to first reduce the dimension of the feature vector and then binary code the feature vector. The similarity between binary codes can be expressed in Hamming distance, and the efficiency of calculating similarity is greatly increased.

Inspired by the recent research advancement in deep learning, a method based on deep convolutional neural network is proposed for fabric image retrieval. The deep CNN can learn the image representations and hash code when the labeled data are available. In other words, the proposed is based on supervised learning. Furthermore, Lin *et al.* [19] suggested that when the data labels are available and a powerful learning model is used, the binary codes can be learning by adding some hidden layers for representing. Difference from the method presented by Lin, our method apply a sparse convolutional neural network architecture called Inception. The pre-training process of the CNN model mainly trains the image representations, and the fine-tuning learn the feature hash code. The proposed method learns image representation and hash codes in a manner of point of point by using the incremental learning nature of deep CNN (Gradient Descent Algorithm). Meanwhile, the deep learning based architecture also allows for efficient image retrieval of representation learning.

- The proposed method apply a simple but effective supervised deep learning based framework for wool fabric image retrieval.
- The retrieval strategy of the proposed method is hierarchical search, that is, coarse-level retrieval is first performed using binary codes, and then fine-level retrieval is performed using high-dimensional features.
- A dataset named WFID is created for fabric image retrieval. the dataset contains 19,564 wool fabric images of which 4,062 images have been labeled.
- The superiority and rationality of the proposed method have been demonstrated by experiments. We also rigorously discussed and evaluated key modules of the model to improve model performance.

The rest of this paper is structured as follows. Section 2 review the related work. Section 3 introduce the framework and specific steps of the proposed method. Experiments and comparative discussions are described in Section 4. Section 5 concludes the paper.

II. RELATED WORKS

According to different technical components, the two topics most relevant to this research are deep representation learning and traditional descriptors. The two topic will be reviewed in this section.

A. TRADITIONAL DESCRIPTOR

The focus of traditional image retrieval research is to manually design an algorithm for image representation which is used for discriminating or matching two images. Because Scale-Invariant Feature Transform referred to as SIFT [2] algorithm is robust to image rotation, translation and scale transformation, it is widely applied in image feature extraction and retrieval. Many studies combined Bag of Features referred to as BOF and SIFT for large-scale image retrieval, as in [20]. The idea of these methods is to first extract the local features of the image, then quantize the extracted local features into visual feature words, and finally use the histogram of the frequency of each visual feature word to represent the image and then retrieval the image. On the basis of this idea, Nister and Stewenius [21] quantified the local feature visual features into a hierarchical vocabulary, which reduces the computational complexity of feature matching, thereby improving retrieval speed and quality. To improve the speed of large-scale image retrieval, Jégou *et al.* [22] proposes a context-based measure of dissimilarity. Moreover, Jégou *et al.* [23], [24] also presented a more accurate and robust image representation method by integrating weak geometric and Hamming embedding consistency in inverted files.

There are also some works that combine several local features into global features for image retrieval. Jing [25] present a method of printed fabric image retrieval based on multi-feature fusion, in which color moments represents color features, and GIST represents spatial shape features. Based on rotation invariant and multi-scale LBP, Zhang [26] proposed a method for lace fabric image retrieval. Perronnin and Larlus [27] proposed a method to first represent the image as Fisher Vector and then use the binarization compression technique to compress the vector to speed up the retrieval.

Despite achieving significant advances in image retrieval, these studies are not always optimized for specific tasks because they are heavily dependent on manual feature extraction.

B. DEEP REPRESENTATION LEARNING

Recently, significant progress has been made on image analysis [28] by moving from the early low-level feature based algorithms to deep learning based end-to-end frameworks. Deep representation learning is typical of end-to-end frameworks and has been widely employed in a variety of visual tasks, the most common of which are image classification [8], [29]–[33], object detection [34]–[36], image segmentation [37]–[40], pix-wise image labeling [38], [41] and human centric analysis [42], [43]. Next, we will review the

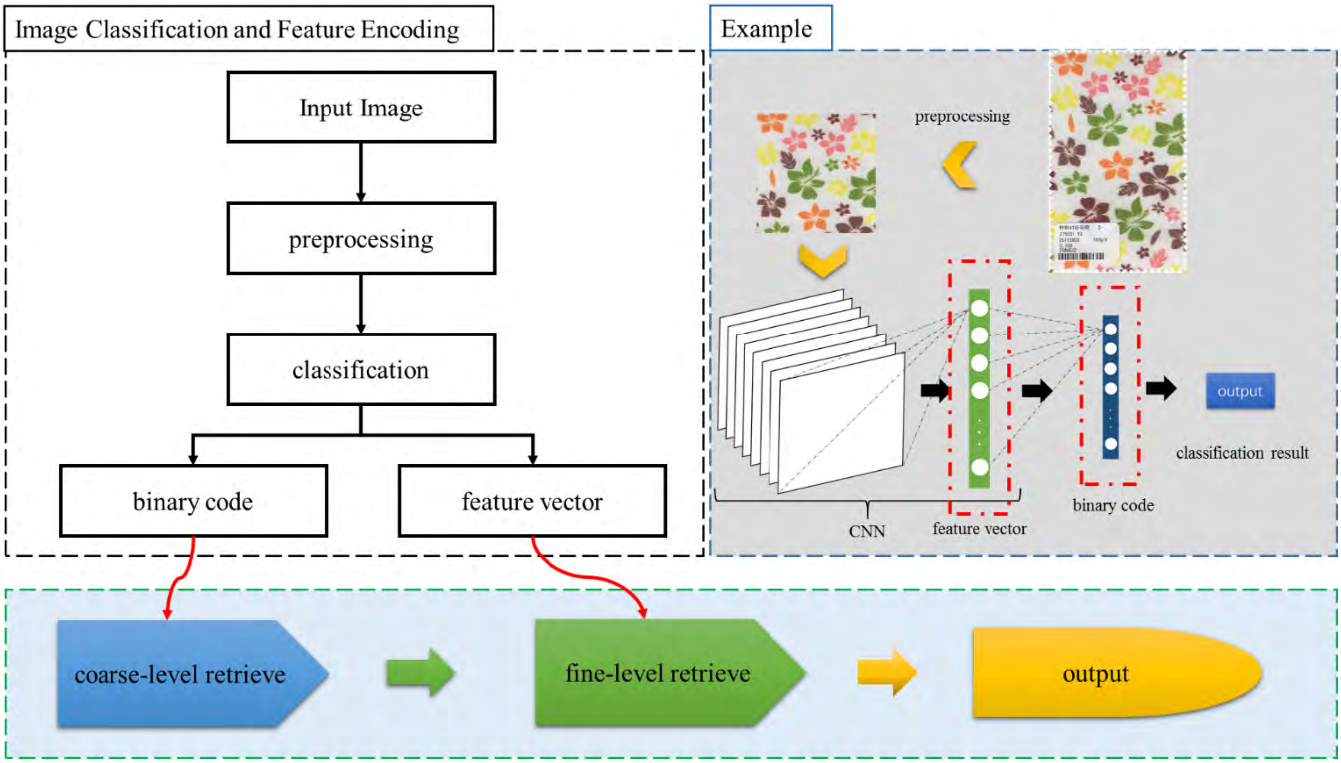


FIGURE 1. **The framework of proposed fabric image retrieval system.**

research work of deep representation learning in the field of image retrieval. Babenko *et al.* [11] proposed a global feature extraction method for image retrieval. In this method, the CNN is first pre-trained on a public large-scale dataset such as ImageNet, then the pre-trained model is fine-tuned on the target data, and finally the output of the fully connected layer is extracted as an image representation. In their follow-up study, the average and maximum pooling of the last convolutional layer of CNN was taken as an image representation and better performance. Deng *et al.* [44] present a method for fabric image retrieval based on learning deep similarity model with focus ranking. Perronnin and Larlus [27] proposed an image retrieval framework based on multi-feature fusion. The two features of this framework fusion are supervised depth representation learning and unsupervised Fisher Vector [6], [7], [27] intensity representation. However, these methods extract deep global features based on CNN's classification objective function, which merely considered feature learning and extraction and ignored the influence of feature dimension on retrieval.

Deep learning architecture has been applied for hash learning, but most of them are based unsupervised learning, such as using a deep self-encoded network to represent the image [16], [45]. Xia *et al.* [17] presented a method for fast image retrieval based on unsupervised learning for learning binary hashing codes which performed good on several public datasets.

In contrast, we present a method for wool fabric image retrieval based on deep representation learning by using binary hashing code learning, and it achieves good performance on target dataset. The following section will describe the method proposed in this paper.

III. METHODS

The proposed fabric image retrieval framework in this paper is shown in Figure 1. The framework includes three key components. The first component is to perform preprocessing on the image called histogram equalization. The second component is the feature extraction by using deep CNN. The third component is hierarchical search strategy. The proposed method is described in detail as follows.

A. IMAGE PREPROCESSING

To expose the surface features of the fabric image clearly, the image is preprocessed – image histogram equalization. Compared with the RGB, HSV is a color model for visual perception. Its main features are that the luminance component is independent of the color information, and the second is that the hue and saturation are closely linked to the way people perceive color. So we perform histogram equalization of the V component of the image in the HSV color space.

For an input image, the first step of preprocessing is to convert the image color space — from RGB color space to

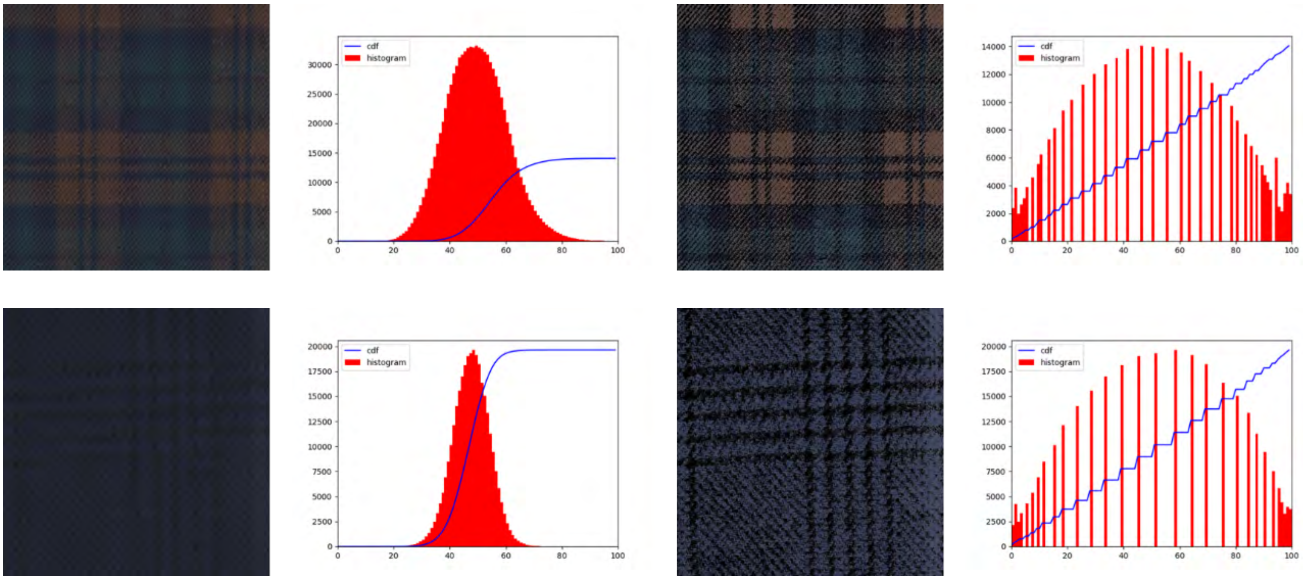


FIGURE 2. **Comparison of image equalization before and after.**

HSV color space, which can be expressed by formula 1.

$$\begin{cases} V = \frac{1}{\sqrt{3}}\left[R+G+B\right] \\ S = 1 - \frac{\sqrt{3}}{V}\min\left(R,G,B\right) \\ H = \begin{cases} \theta & \text{G} \geq \text{B} \\ 2\pi - \theta & \text{G} > \text{B} \end{cases} \\ \theta = \cos^{-1}\left[\frac{\frac{1}{2}\left[\left(R-G\right)+\left(R-B\right)\right]}{\sqrt{\left(R-G\right)^2+\left(R-B\right)\left(G-B\right)}}\right] \end{cases} \tag{1}$$

where H, S, and V represents the hue, saturation, and brightness in the HSV color space, respectively; R, G, and B represents the values of the red, green, and blue components in the original image, respectively.

Histogram equalization, also known as histogram flattening, is essentially a nonlinear stretching of an image. It can also be understood as redistributing image pixel values such that the number of pixel values within a certain grayscale range is approximately equal. By equalizing the histogram, the contrast of the peak portion in the middle of the original histogram is enhanced, while the contrast at the bottom of the valley is reduced, and the histogram of the output is a flat segmented histogram.

Let the variable r represent the pixel brightness level in the image. Normalizing the brightness level, $0 \leq r \leq 1$, where $r = 0$ for black and $r = 1$ for white. For a given image, each pixel value is random at the brightness level of [0,1]. The probability density function $p_r(r)$ is used to represent the distribution of image brightness levels. However, discrete probability density functions are generally used in image processing, in which case r^k is used to represent discrete brightness levels and probability density functions are represented by $p_r(r_k)$. Where $p_r(r_k)$ can be expressed by the following formula:

$$p_r(r_k) = \frac{n_k}{MN} \tag{2}$$

where n_k indicates the number of pixels in the image whose luminance value is k, MNrepresents the number of pixels in the image. Then the mathematical expression of the image for histogram equalization can be expressed as

$$s_k = (L-1)\sum_{j=0}^{k} p_r(r_j) = \frac{(L-1)}{MN}\sum_{j=0}^{k} n_j \tag{3}$$

Thus, a processed (output) image is gained by mapping each pixel in the input image with an intensity of r_k into a corresponding pixel with a level of s_k in the output image. The transformation (mapping) $T(r_k)$ in this equation is called a histogram equalization transformation used in proposed method.

The left column in Figure 2 shows the two original image from fabric image dataset. Because the picture was taken with a scanner, the picture appears dark. The second column shows the brightness distribution histogram of the original image. The third column shows the result of performing histogram equalization on each of the two image, and the results show significant improvement. The right column shows the brightness distribution histogram of the equalized images. Given the significant contrast differences between the original images, the example illustrates the power of histogram equalization as an adaptive contrast enhancement tool.

B. IMAGE FEATURE EXTRACTION

Recent studies have demonstrated that the feature activations of CNNs induced by the input image can serve as the image representation or visual signatures. The application of these mid-level feature or image representations indicates

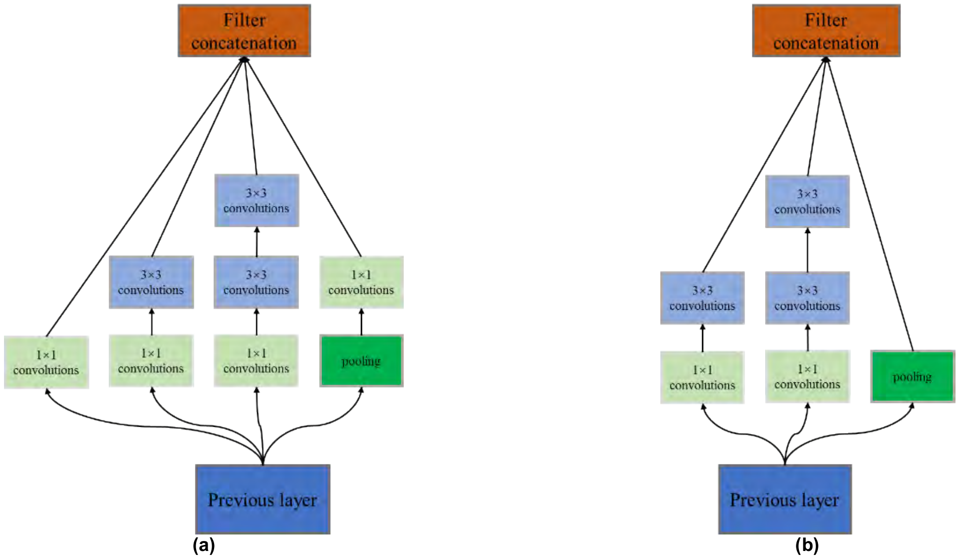


FIGURE 3. The two inception module used in the network. (a) Inception module 1. (b) Inception module 2.

impressive improvement on the task of object recognition, image classification, retrieval, and others. However, using these high-dimensional features directly as an image feature vector and then as an index of the image, the computer computation cost is expensive. To improve retrieval speed and accuracy, one of the most effective methods is to binary encode features. The similarity of the binary compact codes can be characterized by a hash or Hamming distance.

In this work, we proposed a method based deep convolutional neural network to encode the image. The encoding process is performed in two steps:

1) Train the classification model on the labeled fabric dataset to extract the mid-level image representations;

2) Train the encoder based on the mid-level image representations.

The first step is actually a classification task. The proposed method use a Network based on Inception to do this task. The main design idea of Inception's architecture is to find a locally optimal sparse structure in a convolutional visual network. This structure needs to be covered and approximated by dense components that can be obtained. As these "inception modules" are stacked on top of each other, their output correlation statistics are bound to vary. As is shown in Figure 3, in each inception model, 1 × 1 convolutions are used to reduce the input dimension before each 3 × 3 convolutions. In addition to dimensionality reduction, they are also used for data rectified linear activation, which makes it a dual mission. In addition, because the pooling operation plays an important role in the existing level of CNNs, a pooling path is added to each Inception module. Figure 3 shows two inception modules that are also used in the convolutional neural network proposed in this paper.

In the paper "Rethinking the Inception Architecture for Computer Vision", Szegedy *et al.* [33] proposed four principles that generally need to be followed when designing neural networks: (1) Avoid representational bottlenecks; (2) Higher-dimensional representations are easier to localize in the network; (3) Spatial aggregation can be done on lower-dimensional embedding without causing any or all loss in presentation capabilities; (4) Balance the width and depth of the network. Following the four principles, the network architecture in proposed method is described in Table 1.

TABLE 1. The outline of the CNN architecture.

Type	patch size/stride	output size
Input	448×448×3	
Convolution	3×3/2	224×224×28
Convolution	3×3/1	224×224×64
Pool	3×3/2	112×112×64
Convolution	3×3/1	112×112×96
Convolution	3×3/2	56×56×192
Pool	3×3/2	28×28×192
inception×2	inception module 1	28×28×320
Inception	inception module 2	14×14×576
inception×4	inception module 1	14×14×786
Inception	inception module 2	7×7×980
inception×2	inception module 1	7×7×1024
Pool	7×7	1×1×1024
Softmax	Classifier	

After training this CNN on the fabric image dataset, we can get a 1024-dimensional image mid-level feature vector. The efficiency of directly using it for image retrieval is relatively low, due to the high computational cost in the feature matching process. So before the classifier a layer of fully connected hidden layers (128 nodes in this hidden layer are set in this article, reasons will be discussed later in this paper) is added to binary code the extracted features. This added hidden layer

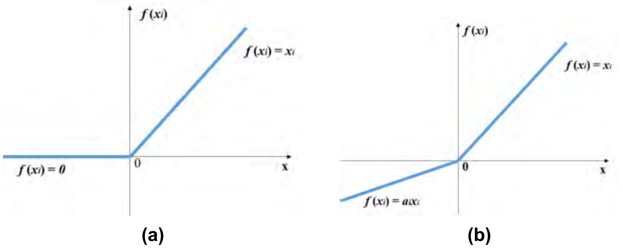


FIGURE 4. (a) ReLU activation function. (b) PReLU activation function.

uses PReLU as its activation function. The activation function defined as:

$$f(x_i) = \begin{cases} x_i & if\ x_i > 0 \\ a_i x_i & if\ x_i \leq 0 \end{cases} \tag{4}$$

where x_i denotes the input of the nonlinear activation function f on the ith channel, and a_i is a coefficient that controls the slope of the negative portion. The subscript i in a_i indicates that nonlinear activation is allowed to vary across different channels. If $a_i = 0$, it becomes ReLU activation function; When a_i is a learnable variable, the function f is called as Parametric ReLU (PReLU). Figure 4 shows the shapes of ReLU and PReLU. Eq. (4) is equivalent to $f(x_i) = \max(0, x_i) + a_i \min(0, x_i)$.

To achieve feature encoding learning, it is necessary to fine-tune the encoding network on fabric image dataset based on back propagation. The weight of the deep convolutional neural network is initialized by the pre-training model, and then these parameters are controlled by a relatively small learning rate. The initialization parameters of the coding layer and the final classification layer are randomly initialized.

To avoid over-fitting, we use the L2-constrained softmax loss as the loss function which added an additional L2-constrained to the regular softmax loss. The loss is given by Equation 5,

$$\begin{aligned} & \arg\min -\frac{1}{M}\sum_{i=1}^{M} \log \frac{e^{W_{y_i}^T f(X_i)+b_{y_i}}}{\sum_{j=1}^{C} e^{W_j^T f(X_i)+b_j}} \\ & \text{subject to } \|f(X_i)\|_2 = \alpha, \quad \forall i = 1, 2, \ldots M \end{aligned} \tag{5}$$

where X_i is the input image in a mini-batch of size M, y_i is the corresponding class label, C denote the number of subject classes, $f(X_i)$ denote the features descriptor obtained from the penultimate layer of the proposed network, and b and W are the bias and weights for the last layer of the network which acts as a classifier. So, the framework of proposed method can be described in Figure 5.

C. IMAGE RETRIEVAL VIA HIERARCHICAL SEARCH

Through in-depth analysis and visualization of deep CNN, Zeiler and Fergus [46] suggested that the shallow layers in the neural network can learn the local low-level visual descriptors such image outline, while the deeper layers in the neural network can capture the high-semantic information suitable for recognition. We adopt a hierarchical search method whose search strategy is to use Fc layer coding for coarse-level search first, and then use Fn layer for fine-level retrieval. Firstly, use the code as an index of image search database to determine some image candidate pools with similar high-level semantic features. Then, to further select images with similar appearance, the deepest mid-level image representations is used to rank similarity.

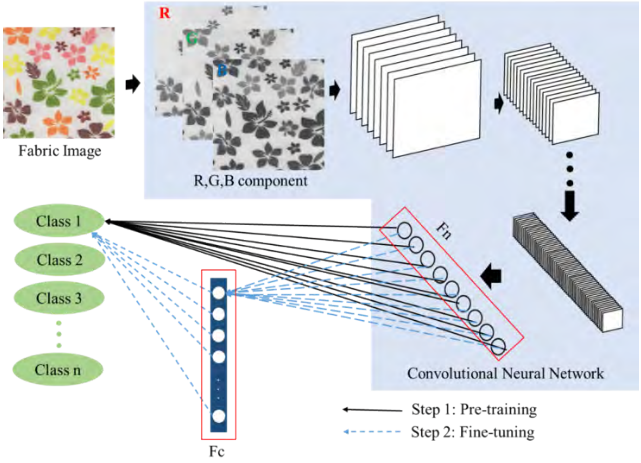


FIGURE 5. The framework of the proposed method.

Given a query image I, the process of extracting its encoding is as follows:

a) Extract the image signature which is the output of the encoding layer denoted by $Out(c)$;

b) Binarize the activations by a threshold.

We can output the binary codes of B by

$$B^j = \begin{cases} 1 & Out^j(c) \geq 0 \\ 0 & Out^j(c) < 0 \end{cases} \tag{6}$$

where $j = 1, 2, 3 \cdots b$ (b is the number of nodes in the encoding layer – Fc). Let $\boldsymbol{S} = [47]$ indicates the dataset consisting of n images for retrieval. The corresponding binary codes of each image extracted by the deep convolutional neural network are denoted by $\boldsymbol{Sc} = [31]$ with $B_i \in \{0, 1\}^b$. For a query image I_q with binary code of $\mathrm{B_q}$. For this kind of coding, in this paper, using the number of different bits in the code (Hamming distance) to calculate the similarity of two vectors:

$$SIM_H = \frac{N - (\#(B_{Q,i} \neq B_{D,i}))}{N} \tag{7}$$

where N is the number of elements in the HASH code, and $\#(B_{Q,i} \neq B_{D,i})$ represents the number of different digits in the code, and the similarity of the code is denoted by SIM_H. We identify a pool if n candidates, $\boldsymbol{P} = \{I_1^c, I_2^c \ldots I_m^c\}$, if the SIM between B_q and $B_d \in \boldsymbol{S}$ is lower than a threshold.

For a query image q, a set of candidate pools $\boldsymbol{P}$ is obtained by coarse-level search. The features extracted from the layer Fn is used to select the top k ranked image. Let V_q represent the feature vectors of the query image q, and V_i^P represent the

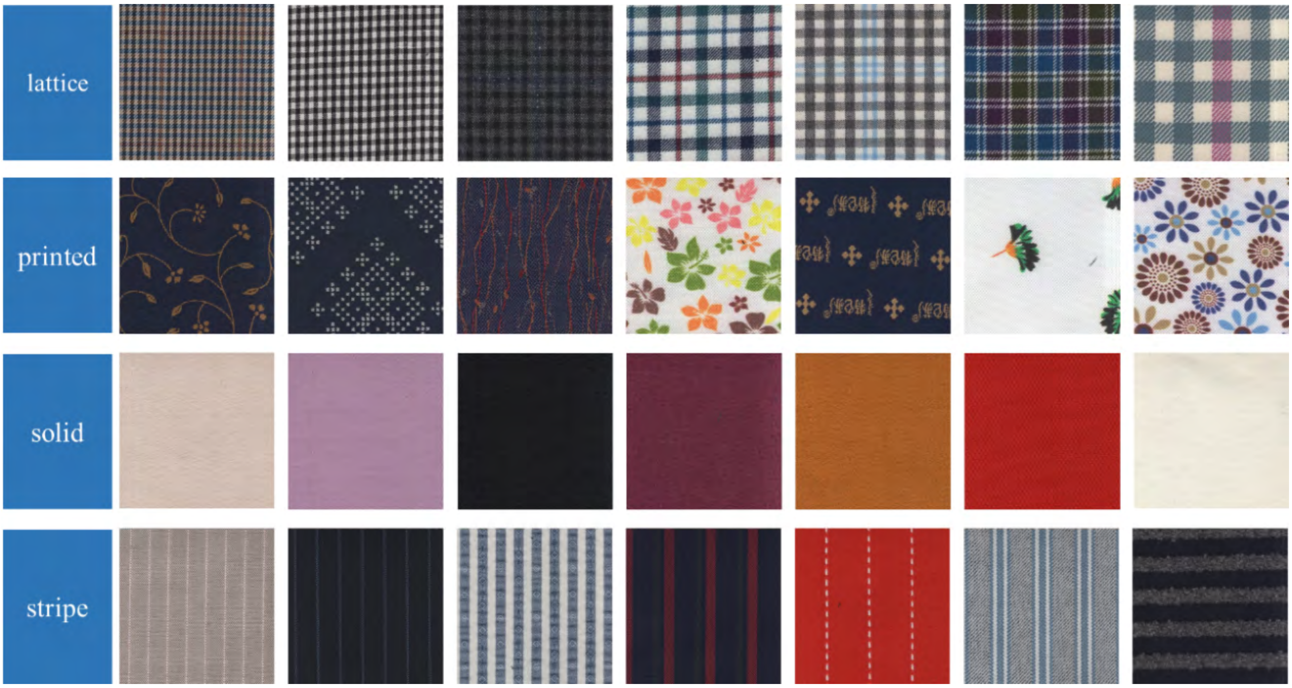


FIGURE 6. Sample wool fabric images in the database.

feature vectors of the image I_i^c from the pool $\boldsymbol{P}$. Euclidean distance is used to measure the distance of the corresponding two feature vectors.

$$Dis_F = \left\| V_q - V_i^P \right\| \tag{8}$$

The smaller the Euclidean distance denoted by Dis_F between vectors, the higher the similarity between the two image. Each candidateI_i^c is ranked in ascending order by the similarity; hence, top k ranked image are identified.

IV. EXPERIMENTS

A. DATASET AND IMPLEMENTATION

To the best of our knowledge, there are not public datasets specifically for the fabric image retrieval problem. Due to the potential value of fabric image retrieval in many applications, e.g., large-scale fabric searching and products design, we create a large-scale fabric image retrieval dataset in this study. Since the fabric are from the woolen factory, we abbreviated this image database as WFID. 19,564 images are collected using a scanner with a resolution of 200 dpi and a size of 448 × 448. According to the appearance of the fabrics, we classify the fabrics into four categories: lattice fabrics, printed fabrics, solid fabrics, and striped fabrics and label 4,062 images for training the model.

As shown in Figure 7, there are three logical main lines of the woolen fabric image retrieval system, which are the training of the classification model, the establishment of a feature database, and the image retrieval process. The system's hardware environment is an HP workstation (Z840 TOWER: CPU– E5-2623 v4 @ 2.60GHz, Memory 32G) with a NVIDIA TITAN XP GPU (11G graphics memory).

1) TRAINING OF THE MODEL

We train the CNN model as shown in Section 3.2 to utilize the deep learning framework –TensorFlow [37]. During training, bactch_size is 64, the number of training steps is 50,000, and the optimizer uses ADAM [48]. As shown in Figure 8(a), the specified learning rate is decayed by exponential decay. The total loss changes during training as shown in Figure 8(b). The model fine-tuning process actually trains only the parameters in the coding layer and the classifier, so the training has taken 5,000 steps and the model has converged.

2) THE ESTABLISHMENT OF DATABASE

Extracting the features of all images in the image library using the trained model, including binary code and feature vector. In this paper, use Sqlite3 database [49]{Pond, 2005 #40} to store image feature vector.

3) THE RETRIEVAL PROCESS

Given a query image, first we need to extract the output of the encoding layer using the trained model, and then use the output to do coarse-level retrieval to get a candidate pool. Then use the feature vector to select the image with high similarity from the candidate pool as the retrieval output.

B. EVALUATION METRICS

In this study, ranking-based criteria is used to evaluate the algorithm. Given a query image, retrieval results, and'

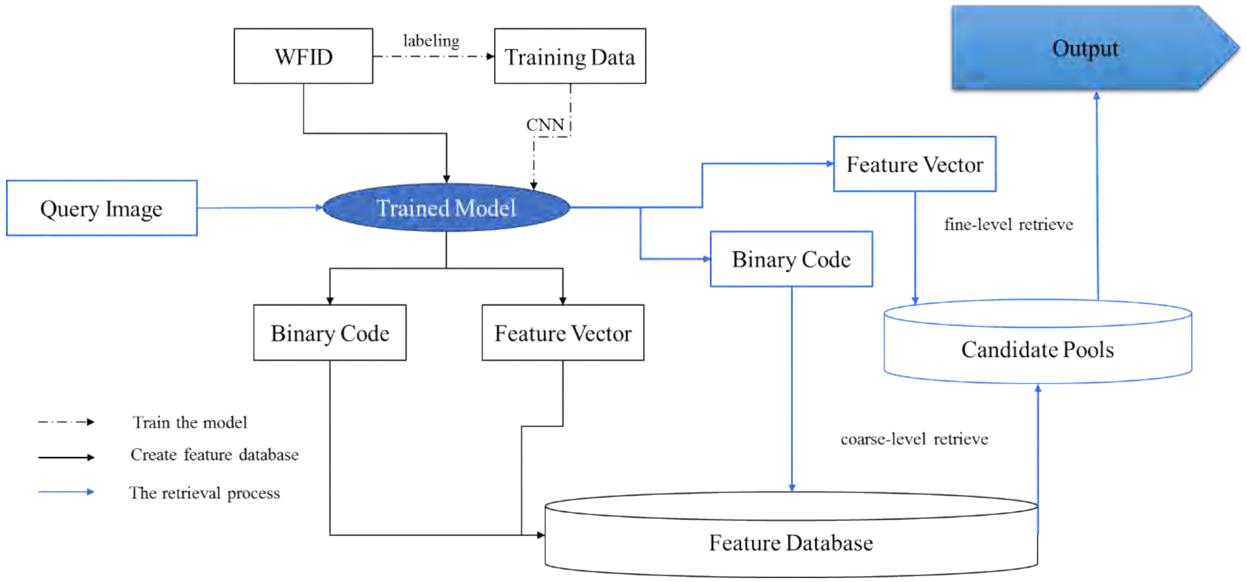
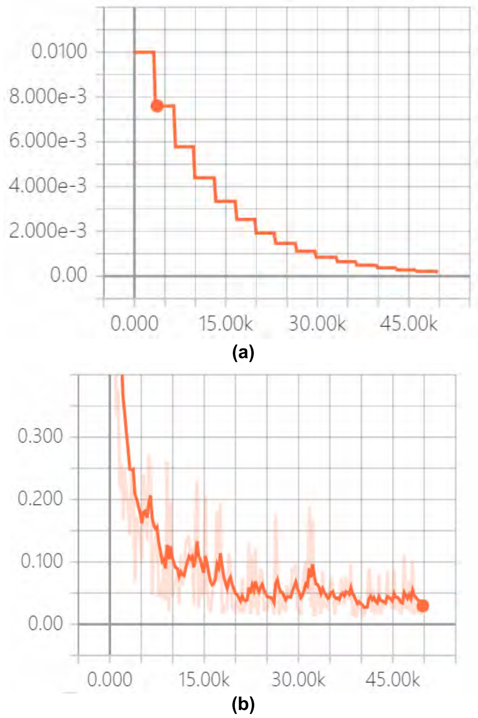


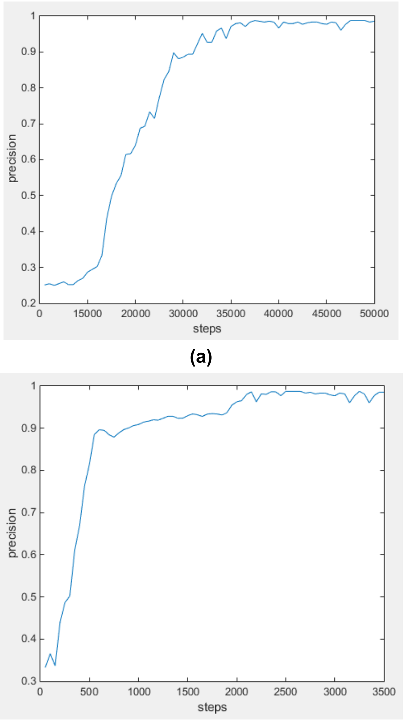
FIGURE 7. The architecture of wool fabric image retrieval system.



(a)

(b)

FIGURE 8. (a) Learning rate and (b) total loss curve during training.



(a)

(b)

FIGURE 9. Precision curve during training. (a) Precision change curve during pre-training. (b) Precision change curve during fine-tuning.

evaluation methods between similar images, a rank can be assigned for each image in the dataset. In this paper, we evaluate the ranking of top k images with respect to a query image by an error rate:

$$E@k = \frac{\sum_{i=1}^{k} R(i)}{k} \tag{9}$$

where k denotes the number of outputs for analyzing. The ground truth relevance between a query image and the ith ranked image is denoted by $R(i)$. In this study, the appearance of image is only taken into consideration, and $R(i) \in \{0, 1\}$ with 1 for the query image and the ith image with the dissimilar appearance and 0 otherwise.

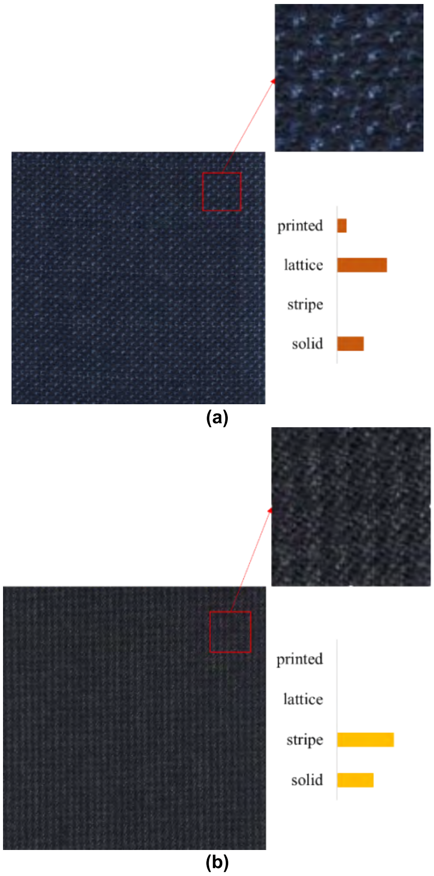


FIGURE 10. Error classification example analysis. (a) Solid→lattice. (b) Solid→stripe.

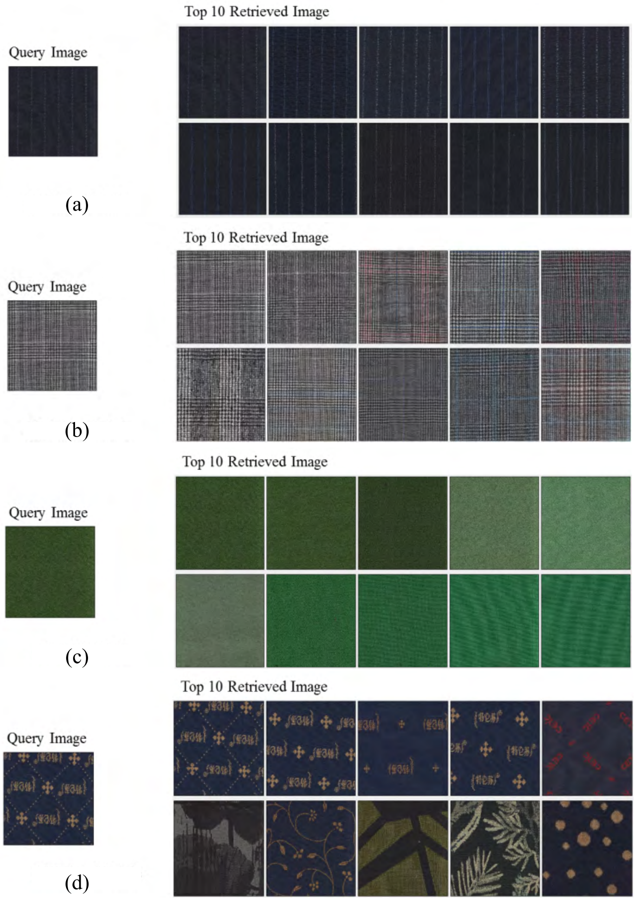


FIGURE 11. The retrieval error rate in each category. (a) Stripe fabric. (b) Lattice fabric. (c) Solid fabric. (d) Printed fabric.

C. RESULTS

1) PERFORMANCE OF IMAGE CLASSIFICATION

All parameters before the Fc layer are trained in pre-training (the amount of parameters is large), while only the parameters of the Fc layer and the classifier are trained in fine-tuning (the amount of parameters is small). So in the pre-training the accuracy converges slowly (it starts to converge at about 35000 steps), while in the fine-tuning the accuracy quickly reaches the peak and converges, as shown in Figure 9. As can be seen from the figure, in the pre-training process, the classification accuracy of the model can reach 98.56%. Also in the fine-tuning process, the accuracy of the classification can be achieved by fitting the network parameters to 98.56%. Through the analysis of misclassified images, it has been found that misidentification mostly results in those images with multiple categories of features. As shown in Figure 10(a), the label of the image is a solid fabric, but the model recognizes it as a lattice fabric because the fabric in this image has the characteristics of a lattice fabric. Similarly, the image in Figure 10(b) has very fine stripes, so the model identifies the fabric in the image as a striped fabric. In general, the trained classification model learned some ‘‘rules’’ for human classification.

2) PERFORMANCE OF IMAGE RETRIEVAL

In this experiment, about 100 images (including one quarter each of lattice, stripe, solid and printed fabrics) were selected to test the method proposed in this paper. We evaluate the ranking of top 10 images with respect to a query image by an error rate. The retrieved results of some images are shown in Figure 11. For stripe fabrics, the most obvious feature is the color, the spacing between the stripes, and the type of stripes (horizontal or vertical stripes). Figure 11(a) shows a fine-vertical-stripe fabric and its top 10 retrieved results. It can be found through analysis that they are highly relevant in appearance. The lattice fabrics in the query image and search results in Figure 11(b) are all hidden lattice fabrics, with only slight differences in appearance. For solid fabrics with color as the main feature, both the search results and the original images belong to the green system, as shown in Figure 11(c), and the higher the ranking is, the higher the similarity is. In general, the pattern cycle of the printed fabric is relatively large, the normal size of the picture is difficult to include a cycle, and the pattern of the printed fabric is varied, so the correlation between the printed fabric images is not well evaluated, and this evaluation is inevitably subjective. Compared with other fabrics, the error rate of image retrieval

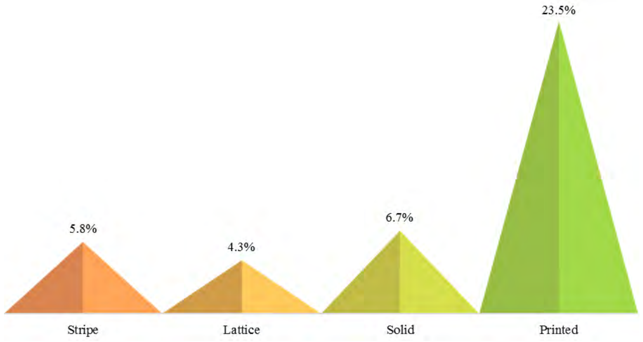


FIGURE 12. The retrieval error rate in each category.

of printed fabrics is high, and the error rate of the images shown in Figure 11(d) is 20%. Figure 12 show the retrieval error rate in each category.

D. DISCUSSION

In image retrieval, the most consumed time is the calculation of similarity between features. For a query image, feature extraction is performed only once, and the extracted features are to be calculated for similarity with all image features in the feature database. In this paper, the strategy to improve the retrieval speed is to use the extracted binary code to narrow the search range first, to obtain a candidate pool, and then to select the pictures with higher similarity from the candidate pool. The shorter the binary code length (number of nodes), the faster the search speed will be.

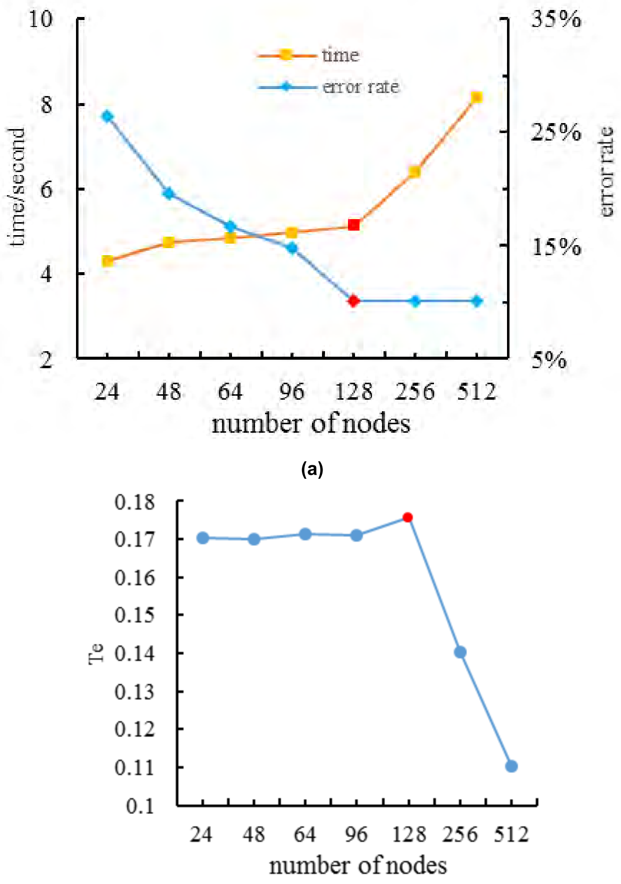
If the output of Fn (1024 dimensional) is directly used as the index of image in WFID for retrieval, the average error rate of the retrieval result is 10.08%, and the average retrieval time is 15.38s. Figure 13(a) shows the average time and average error rate of hierarchical retrieval using different nodal numbers at the Fc encoding layer. In comparison, the strategy of using hierarchical retrieval can greatly reduce the average retrieval time. On the one hand, it is because the length of binary code is small; on the other hand, it is because the calculation amount of hamming distance between vectors is much smaller than that of Euclidean distance.

We use *Te* to indicate the time efficiency of retrieval:

$$Te = \frac{1 - Er}{T} \tag{10}$$

where *Er* denotes the average error rate of the retrieval result, *T* denotes the average retrieval time. The time efficiency curve of different nodes is shown in figure 13 (b). The peak value of the curve is obtained when the number of nodes is 128, so the number of nodes of the encoding layer in this paper is 128.

Table 2 shows the comparison of results with image histogram equalization and without the preprocessing. It can be seen from the table that the equalization improves the retrieval accuracy of striped fabrics, lattice fabrics and solid fabrics, but the retrieval accuracy of printed fabrics is not improved much. Because there are many lattice and striped fabrics with weak textures in the WFID, these fabrics are likely to be recognized as solid fabrics without image enhancement. The average time taken for image equalization is 0.27s, but the equalization reduces the average error rate by 8.75%, so image equalization improves the time efficiency of retrieval.



(a)

(b)

FIGURE 13. Test indicator results for the proposed method. (a) Time and error rate curves. (b) Time efficiency curve.

TABLE 2. Comparison of results with equalization and without equalization.

Evaluation		without equalization	with equalization
Error Rate	stripe	12.5%	5.8%
	lattice	19.6%	4.3%
	solid	18.2%	6.7%
	printed	25.0%	23.5%
Average Error Rate		18.83%	10.08%
Time/second		4.85	5.12
Time Efficiency		0.167	0.176

E. COMPARISON

Since image representation is crucial to image retrieval results, this study compare the retrieval results and evaluation metrics obtained by features from different network

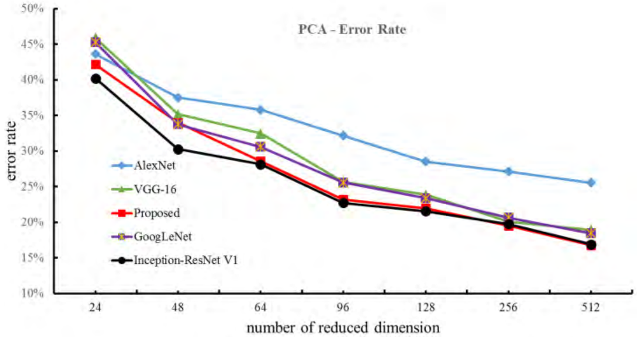


FIGURE 14. Three model evaluation indicators using PCA.

framework: (1) AlexNet: feature from the full connection F7 [8]; (2) VGG-16 [30]: feature from the Fc-4096; (3) GoogLeNet [32]: feature from the dropout layer; (4) Inception-Resnet V1 [47]: feature from the average pooling layer; (5) Proposed: feature from the Fc. Note that the dimensions of the original feature outputted by the first two models are 4,096, and such a high feature dimension cannot be directly used for efficient large-scale image retrieval systems. In addition, the feature output dimensions of the GoogLeNet and Inception-Resnet V1 framework are 1024 and 1792, respectively, and they are both deep learning frameworks based on the Inception structure. The Principal Component Analysis (PCA) (followed by Bebenko *et al.* [11]) and Feature Encoding (FE) are applied for reducing the dimension to smaller ones (24, 48, 64, 96, 128, 256, 512) and perform comparison on this feature dimensions.

Figure 14 shows the comparison results measured in average error rate with respect to feature dimension reduced by PCA. It reveals the other tested models all perform better than the AlexNet baseline one. The evaluation results in error rate demonstrate the superiority of the proposed model. It shows that the proposed model has an advantage in the representation of fabric images. It also suggests that the proposed model can perform much better than the existing ones in fabric image retrieval systems.

The comparison results measured in error rate, average time and time efficiency with respect to the number of nodes learned by deep CNN is shown in Figure 15. Since the proposed model has an extracted feature dimension of 1024 and the other two models have a feature dimension of 4096, as shown in Figure 15(a), the average time result using the proposed model is better than the other two models. VGG-16, Inception-ResNet V1 and the proposed model have the same excellent performance in the evaluation metrics of retrieval results – average error rate. However, the comprehensive evaluation index of the proposed method – time efficiency is significantly better than the others, as we can see from Figure 15(b) and Figure 15(c). It suggests again that the proposed method can perform much better than the other methods.

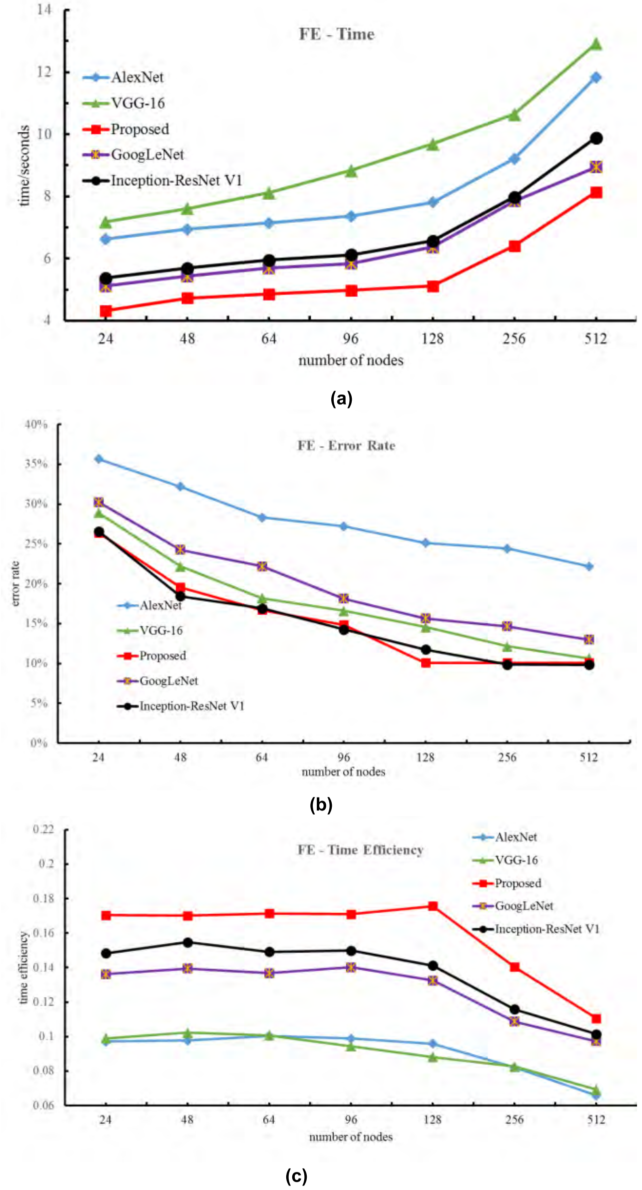


FIGURE 15. Three model evaluation indicators using proposed framework. (a) FE-Time curves. (b) FE-Error rate curves. (c) FE-Time efficiency curves.

To evaluate the retrieval performance, we compare the proposed method with several methods including CMG [25], MS+LBP [26] (based on traditional descriptor), Fast SH [50], SDH [51] (based on traditional hashing methods), CNNH+ [17], DSH [52] (based on deep hashing methods). The retrieval is performed by randomly selecting 400 query images from testing set for the system to retrieval relevant ones from the WFID. Table 3 shows the retrieval error rates of several methods listed and the proposed method. As can be seen, the performance of deep learning based methods is significantly better than those methods which are based on traditional descriptor and traditional hashing. In addition, the hash algorithm-based method is much faster than the manual feature extraction based method, which was found in the experiment. Furthermore, the proposed method achieves

TABLE 3. Retrieval error rate of several methods.

Methods	Error Rate @k=10
CMG[25]	35.8%
MS+LBP[26]	42.5%
Fast SH[50]	37.5%
SDH[51]	34.2%
CNNH+[17]	17.5%
DSH[52]	16.8%
Proposed Method	10.1%

better performance than other un-supervised and supervised methods, and it attains an error rate of 10.1% on WFID.

V. CONCLUSION

In this paper, we present a simple yet effective deep learning framework, which aims to retrieve the similar image from the database. We apply a convolutional neural network based on sparse network structure – Inception to feature extraction and encoding of images. The retrieval strategy of the proposed method is to use a hierarchical search, that is, to perform coarse-level retrieval on the image using a binary encoding similar to hash encoding, and then use the extracted high-dimensional features to perform fine-level retrieval. Experimental results show that, the proposed method has good performance on WFID with an average error rate of 10.08%. Comparison experiments with VGG-16, AlexNet, GoogLeNet and Inception-ResNet V1 illustrate that the proposed method outperforms them in comprehensive retrieval performance.

REFERENCES

[1] K. K. Seo, "An application of one-class support vector machines in content-based image retrieval," *Expert Syst. Appl.*, vol. 33, no. 2, pp. 491–498, 2007.

[2] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[3] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.

[4] D. Galvez-López and J. D. Tardos, "Bags of binary words for fast place recognition in image sequences," *IEEE Trans. Robot.*, vol. 28, no. 5, pp. 1188–1197, Oct. 2012.

[5] H. Jegou, M. Douze, C. Schmid, and P. Perez, "Aggregating local descriptors into a compact image representation," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3304–3311.

[6] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[7] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the Fisher kernel for large-scale image classification," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 143–156.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[9] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[10] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and F. F. Li, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 248–255.

[11] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, "Neural codes for image retrieval," in *Proc. Eur. Conf. Comput. Vis.*, vol. 8689, Zürich, Switzerland, Sep. 2014, pp. 584–599.

[12] B. Kulis and T. Darrell, "Learning to hash with binary reconstructive embeddings," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2009, pp. 1042–1050.

[13] M. Norouzi and D. J. Fleet, "Minimal loss hashing for compact binary codes," in *Proc. Int. Conf. Int. Conf. Mach. Learn.*, 2011, pp. 353–360.

[14] A. Gionis, P. Indyk, and R. Motwani, "Similarity search in high dimensions via Hashing," in *Proc. Vldb*, vol. 8, 1999, pp. 518–529.

[15] J. Shao, F. Wu, C. Ouyang, and X. Zhang, "Sparse spectral hashing," *Pattern Recognit. Lett.*, vol. 33, no. 3, pp. 271–277, Feb. 2012.

[16] Y. Weiss, A. Torralba, and R. Fergus, "Spectral hashing," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 282, 2008, pp. 1753–1760.

[17] R. Xia, Y. Pan, H. Lai, C. Liu, and S. Yan, "Supervised hashing for image retrieval via image representation learning," in *Proc. AAAI Conf. Artif. Intell.*, 2014, pp. 1–2.

[18] W. Liu, J. Wang, R. Ji, Y.-G. Jiang, and S. F. Chang, "Supervised hashing with kernels," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2074–2081.

[19] K. Lin, H.-F. Yang, J.-H. Hsiao, and C.-S. Chen, "Deep learning of binary hash codes for fast image retrieval," in *Proc. Comput. Vis. Pattern Recognit. Workshops*, Dec. 2015, pp. 27–35.

[20] J. Sivic and A. Zisserman, "A text retrieval approach to object matching in videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2003, pp. 1470–1477.

[21] D. Nistér and H. Stewénius, "Scalable recognition with a vocabulary tree," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2006, pp. 2161–2168.

[22] H. Jegou, H. Harzallah, and C. Schmid, "A contextual dissimilarity measure for accurate and efficient image search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 1–8.

[23] H. Jegou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 304–317.

[24] H. Jégou, M. Douze, and C. Schmid, "Improving bag-of-features for large scale image search," *Int. J. Comput. Vis.*, vol. 87, no. 3, pp. 316–336, May 2010.

[25] J. Jing, Q. Li, P. Li, and L. Zhang, "A new method of printed fabric image retrieval based on color moments and gist feature description," *Textile Res. J.*, vol. 86, no. 11, pp. 1137–1150, 2015.

[26] L. Zhang, X. Liu, Z. Lu, F. Liu, and R. Hong, "Lace fabric image retrieval based on multi-scale and rotation invariant LBP," in *Proc. Int. Conf. Internet Multimedia Comput. Service*, 2015, pp. 1–5.

[27] F. Perronnin and D. Larlus, "Fisher vectors meet neural networks: A hybrid classification architecture," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3743–3752.

[28] A. Khatami, A. Khosravi, T. Nguyen, C. P. Lim, and S. Nahavandi, "Medical image analysis using wavelet transform and deep belief networks," *Expert Syst. Appl.*, vol. 86, pp. 190–198, Nov. 2017.

[29] S. Ioffe and C. Szegedy. (2015). "Batch normalization: Accelerating deep network training by reducing internal covariate shift." [Online]. Available: https://arxiv.org/abs/1502.03167

[30] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[31] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *Proc. AAAI*, 2016, pp. 1–12.

[32] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1–9.

[33] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Dec. 2015, pp. 2818–2826.

[34] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[35] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2015, pp. 21–37.

[36] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 779–788.

[37] M. Abadi *et al.* (2016). "TensorFlow: Large-scale machine learning on heterogeneous distributed systems." [Online]. Available: https://arxiv.org/abs/1603.04467

[38] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

[39] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1520–1528.

[40] B. Vijay, K. Alex, and C. Roberto, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[41] T. Chen, L. Lin, L. Liu, X. Luo, and X. Li, "DISC: Deep image saliency computing via progressive representation learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1135–1149, Jun. 2016.

[42] L. Lin, K. Wang, W. Zuo, M. Wang, J. Luo, and L. Zhang, "A deep structured model with radius-margin bound for 3D human activity recognition," *Int. J. Comput. Vis.*, vol. 118, no. 2, pp. 256–273, 2016.

[43] S.-Z. Chen, C.-C. Guo, and J.-H. Lai, "Deep ranking for person re-identification via joint representation learning," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2353–2367, May 2016.

[44] D. Deng, R. Wang, H. Wu, H. He, Q. Li, and X. Luo, "Learning deep similarity models with focus ranking for fabric image retrieval," *Image Vis. Comput.*, vol. 70, pp. 11–20, Feb. 2018.

[45] A. Krizhevsky and G. E. Hinton, "Using very deep autoencoders for content-based image retrieval," in *Proc. Eur. Symp. Artif. Neural Netw. (Esann)*, Bruges, Belgium, 2011, pp. 1–5.

[46] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 8689, 2014, pp. 818–833.

[47] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. AAAI*, 2017, p. 12.

[48] D. P. Kingma and J. Ba. (2016). "Adam: A method for stochastic optimization." [Online]. Available: https://arxiv.org/abs/1412.6980

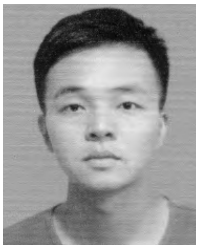
[49] S. L. K. Pond, S. D. W. Frost, and S. V. Muse, "HyPhy: Hypothesis testing using phylogenies," *Bioinformatics*, vol. 21, pp. 676–679, Mar. 2005.

[50] G. Lin, C. Shen, Q. Shi, A. van den Hengel, and D. Suter, "Fast supervised hashing with decision trees for high-dimensional data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1963–1970.

[51] F. Shen, C. Shen, L. Wei, and H. T. Shen, "Supervised discrete hashing," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 37–45.

[52] H. Liu, R. Wang, S. Shan, and X. Chen, "Deep supervised hashing for fast image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2064–2072.

JUN XIANG received the master's degree in textile engineering from Jiangnan University, Wuxi, China, where he is currently pursuing the Ph.D. degree in textile science and engineering with the School of Textile and Clothing. His research interests include image analysis, textile measurement, machine learning, and intelligent manufacturing.



NING ZHANG received the B.S. degree in textile engineering from Jiangnan University, Wuxi, China, in 2016, where he is currently pursuing the Ph.D. degree with the School of Textile and Clothing. His current research interests include interactive genetic algorithm and its application in textile and garment industry.

RURU PAN received the B.S. and Ph.D. degrees in textile engineering from Jiangnan University, Wuxi, China, in 2005 and 2010, respectively, where he is currently an Associate Professor with the School of Textile and Clothing. His current research interests include digital textile technology and digital image processing of textile.

WEIDONG GAO received the B.S. and M.S. degrees in textile engineering from the Wuxi Institute of Light Industry, Wuxi, China, in 1982 and 1985, respectively, and the Ph.D. degree in textile engineering from Donghua University, Shanghai, China, in 2011. He is currently a Full Professor with the School of Textile and Clothing, Jiangnan University, Wuxi. His current research interests include intelligent textile technology, intelligent weaving, and digital image processing of textile.