IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 May 2023, accepted 15 June 2023, date of publication 19 June 2023, date of current version 17 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3287630

RESEARCH ARTICLE

Fine-Grained Retrieval Method of Textile Image

SHUTAO TAN[1], LIANG DONG[1], MIN ZHANG[2], AND YE ZHANG[1]
[1]Department of Information,The Affiliated Hospital of Jiaxing University, Jiaxing, Zhejiang 314001, China
[2]Department of Intelligent Engineering Technology, Jiangsu Vocational College of Finance and Economics, Huai'an, Jiangsu 223001, China

Corresponding author: Liang Dong (tst960217@126.com)

This work was supported in part by the Key Discipline of Jiaxing General Practice Medicine Construction Project 2023-fc-002, and in part by the Fund of the Huaian Science and Technology Project HAB202237.

ABSTRACT There are a large category of textile images that have the characteristics of high local feature repetition rate and complex background information. These types of images have significant intra-class differences and small inter-class differences, making it impossible to perform classification training. Making it difficult for existing methods to accurately retrieve textile images. To improve the retrieval accuracy of textile images, this paper defines multiple repeated local fine-grained features in textile images as textile image ''feature components'', extracts multiple ''feature components'' from a textile image, and fuses the ''feature components'' to generate the definition of textile ''fingerprints''. We propose an image retrieval method based on a pre-trained Mask R-CNN model to extract multiple ''feature components'' in the textile image, then extract the depth feature of the textile again through a convolution neural network, and fuse the extracted depth feature to obtain the ''fingerprint'' of the textile. The obtained ''fingerprint'' can effectively eliminate the interference of large area background and a large number of local repeated features in the textile image, and improve the efficiency of textile retrieval. A series of comparative experiments are carried out on textile image data sets with repeated features. The experimental results show that the proposed method is generally effective.

INDEX TERMS Mask R-CNN, image retrieval, deep feature fusion, fine-grained image retrieval.

I. INTRODUCTION

In recent years, with the development of the textile image design industry, the number of textile images has grown rapidly, showing an exponential growth trend. How to better apply textile image data to realize convenient, fast, and accurate query and retrieval of image information required by users, so that managers can be freed from a large number of monotonous manual management work, has become an urgent problem to be solved. Image retrieval technology has a wide range of applications in various industrial fields. There are many problems such as time-consuming and laborious classification search of textile images, and low image accuracy required by relevant practitioners for retrieval. For example, the 2020 China Shaoxing Keqiao International Textile Surface Accessories Expo showed 450000 fabrics. How to effectively screen similar images that customers need from local fabric patterns is a typical problem. The textile fabric images accumulated over a long time are complex and diverse, with a series of characteristics such as a high repetition rate of internal features of images, complex background information of images dominated by a single pattern, and large difference in image size.

The mainstream solution is to use content-based image retrieval, namely CBIR (Content-based Image Retrieval) technology. This image retrieval technology enables users to input an image and find other images with the same or similar content. Mao [1] put forward the concept of content-based image retrieval technology in his paper in 1992. In his paper, he built an image database based on color and shape and provided certain retrieval functions for experiments. Since then, the CBIR has been widely used in various research fields. Learning effective feature representation and similarity measurement is crucial to the retrieval performance of CBIR systems. Despite decades of extensive research, it is still one of the most challenging and open issues.

With the development of deep learning technology, the convolutional neural network (CNN) model has achieved

The associate editor coordinating the review of this manuscript and approving it for publication was Zhongyi Guo.

success in image retrieval, classification, and other related fields, making the representation of image depth features a research hotspot. Compared with traditional features, depth features can extract high-level semantic information from images through multi-layer convolution calculation. There are two kinds of depth learning methods in image retrieval. One is based on the CNN model pre-trained on ImageNet [2] (such as VGG [3], ResNet [4], and [5], [6], [7], [8]), which uses the output of the full connection layer as image features to improve image retrieval accuracy. Babenko et al. [6] first expressed the whole image of the image retrieval task as a global neural code as the feature of image retrieval; After that, Sharif et al. [7] and Yue et al. [8] tested the effect of retrieving several commonly used image databases on different networks. Although the expression ability of the neural network has been improved with the deepening of the neural network, the effect of directly applying the pre-trained CNN model as a feature extractor is always unsatisfactory, especially in some fine-grained image feature extraction, and the detection accuracy has not been significantly improved compared with traditional features. The other method is based on the fine-tuned pre-trained model to obtain a network model more suitable for the image classification database. E.g. Radenović [9], [10] fine-tune the CNN by mining the positive and negative samples in the database. The fine-tuning model is very important for learning the fine-grained image classification in image retrieval, but the fine-tuning model will lead to the problem of ''catastrophic forgetting'', almost losing the recognition ability of the original data set, and the fine-tuning method cannot migrate the model to different data sets.

There are many unique features in textile images, and the features extracted by the pre-trained model of ImageNet are not completely applicable. One kind of textile image has the characteristic of a high local feature repetition rate, which makes the depth network extract the global feature of image repeated distribution, and the extracted global feature will interfere with fine-grained image feature retrieval; Another kind of textile image is composed of a single subject and complex background. The image background features extracted by the depth network are irrelevant to image retrieval and may also cause interference.

This paper proposes a method of textile image fine-grained retrieval. Mask R-CNN model [24] is used to extract the fine-grained features in the textile image as the ''feature component'' of the textile image, and the depth features of the textile fingerprint are extracted through the trunk network of Mask R-CNN, and the extracted depth features are weighted and fused as the ''fingerprint'' of the textile, improving the accuracy of the textile image fine-grained retrieval. The main contributions include three aspects: 1) A method based on an improved Mask R-CNN model to extract ''feature components'' with fine-grained features in textile images is proposed. To solve the problem of the high repetition rate of internal features in textile images, search and locate the '' feature components'' with high repetition in textile images, and use one of the image features in image retrieval, so that image retrieval is not affected by high repetition features; Because of the complex image background in the textile image, the key part of the image is searched, and the main part is used to replace the whole image for image retrieval, to eliminate the interference of the image background on the main target. The resolution of textile images is between $1024 * 1980$ and $4096 * 4096$. The above extracted ''feature component'' is used as the input image in the retrieval, which reduces the input size and improves the efficiency of image retrieval index construction. 2) A weighted fusion method of textile image ''feature components'' is proposed, which uses multiple extracted ''feature components'' to construct a textile image ''fingerprint'' with a 1:1 weighting ratio. Use the fused textile features to calculate the similarity and sort the results accordingly, which improves the impact of the features of a single '' feature component'' on the image features and greatly improves the image retrieval performance. 3) The experiment shows that the feature fusion method of the textile image can effectively fuse the fine-grained features in a textile image. Compared with traditional methods and depth learning methods, this fusion method has shown good performance in many kinds of CNN depth feature extraction.

II. RELATED WORK

Textile image retrieval can be classified as content-based image retrieval, which has been an important research topic in the field of computer vision for the past two decades. Early traditional image features based on manual design can be divided into global features and local features. Global features are usually based on image statistics, such as color histogram [11]; Local features are descriptions of detail, such as the Scale Invariant Feature Transform (SIFT) [13] descriptor. The advantage of traditional features is that the extraction process is simple and direct, does not require learning and training steps, and meets the unsupervised requirements of content-based image retrieval tasks. However, the ability to express visual features based on pixel values is limited, especially the difficulty in describing the semantic level of image content.

In recent years, deep learning has been widely used in the field of computer vision, such as image classification and image recognition. The convolution neural network method has gradually replaced the traditional method. Especially for large-scale image data classification, the establishment of ImageNet [2] provides sufficient data samples and classification criteria for deep learning model training. Some algorithms fine-tune the network through additional collected databases (Fine-tune) [14], [17], which can have good retrieval performance on specific categories of images. Fine-tune models have been fitted and the transfer ability of fine-tuning results is weak, resulting in textile image data being difficult to classify and model fit on some images irrelevant to the collected image content to improve retrieval accuracy.

Xie [16] first proposed the concept of ''fine-grained image retrieval'' in 2015, and unsupervised retrieval of fine-grained images is a very challenging cutting-edge issue. This problem requires the algorithm to achieve object localization under unsupervised conditions and to select convolutional feature descriptors based on the localization results. Perform averaging and maximum pooling operations on the retained depth features, and then cascade them to form the final image representation to improve retrieval performance. Textile image retrieval can be considered as a separate classification for each image, making it difficult to classify it for traditional classification models for retrieval. Fine-grained image retrieval provides a new direction for textile image retrieval. Textile image retrieval uses local images to retrieve images containing similar textures and styles. Local details of images play a more important role than global structures. How to maximize the preservation of local image features is a feasible direction to improve retrieval accuracy.

SCDA [17] is a fine-grained image retrieval method in an unsupervised setting. SCDA found that even if the data set used by the pre-trained model is very different from the retrieved data set, using the features extracted by the pre-trained model can still segment the foreground target well and remove the influence of background noise. This provides a theoretical basis for the segmentation of textile images in this article. SCDA extracted the image representation vectors of relu5-2 and pool5 respectively, and weighted cascaded them as the final representation vectors. The weight represented by pool5 is 1, while the weight of releu5-2 is 0.5. SCDA found that the Semantic information of the relu5-2 feature is not as rich as that of the pool5 feature, but the target detection of the relu5-2 feature is more accurate than that of pool5. Integrating lower layers, such as pool4, may result in some performance loss. In addition, SCDA uses both the input of the original image and the input of the horizontally flipped original image to extract depth features, and cascades them into the final image representation.

In the past few years, many fine-grained recognition methods [18], [19], [20], [21], [22], [23] have been proposed, which can be roughly divided into three categories. The first type of recognition method attempts to obtain a learning model suitable for the new data set by fine-tuning the pre-trained model, learning more discriminative feature representation, and achieving fine-grained image classification, such as [19] and [21]; The second method aims at objects in fine-grained images to eliminate the influence of pose changes and camera positions, such as [20]; The third category focuses on component-based representation because it is generally believed that the subtle differences between fine-grained images mainly lie in the unique attributes of object components. Chen et al [26] proposed the destruction-construction learning framework (DCL), which inputs images without object structure into network learning, forcing the classification network to focus on more meaningful local sub-regions to classify the features of fine-grained images. Wei et al. [27] proposed locating multiple fine-grained features to discard the convolution descriptors of background and noise areas to improve the recognition accuracy of fine-grained levels. Wei et al [28] proposed the Two Level Progressive Attention Convolutional Network to use a two-layer attention mechanism to accurately locate fine-grained features and improve retrieval accuracy. Gao [29] propose a Chinese character part segmentation method based on Faster R-CNN which proves the effectiveness of the proposed method. This method greatly improves the recognition accuracy of Chinese characters.

The difficulty of the textile image retrieval task in this paper compared to other datasets is that it belongs to content-based image retrieval, where each textile image can be considered as a category. Traditional training of a new CNN model through classification is not suitable for this type of dataset. The local feature is essential for distinguishing different textile images. Because of the high repetition rate of local features and the complexity of background information in textile images, using local feature maps as feature matching methods in fine-grained image retrieval, the Mask R-CNN is used to extract important fine-grained local image features in textile images.

The extraction of local features from textile images helps to eliminate the interference of noise background and improve the retrieval accuracy by discarding the convolution descriptors of background and noise area as well as fine-grained image retrieval tasks [26], [27]. For the dataset in this paper, the effect of the same ''feature components'' on the depth feature extraction is excluded to improve the overall expression of local textile features.

III. CONSTRUCTION METHOD OF TEXTILE IMAGE FINE-GRAINED RETRIEVAL

A. OVERVIEW OF CONSTRUCTION METHODS OF TEXTILE IMAGE FINE-GRAINED RETRIEVAL

Aiming at the problems of the high repetition rate of core components and a large area of background information interference in textile images, this paper adopts the textile image fine-grained retrieval method, the key of which is the ''feature component'' extraction method. The ''feature component'' of the textile image refers to the image part containing the core image features of the textile image. The most difference between the proposed textile image fine-grained retrieval method and the deep learning method used in general image retrieval is that the features extracted by the pre-trained model are used to locate the fine-grained image ''feature components'' of the textile in the model selection of the main image feature extraction. In addition, the proposed textile image fine-grained retrieval method uses the pre-trained model of ImageNet to extract features without any fine-tuning on the target image data set. Fine-tuning is prone to over-fitting and the portability of fine-tuning results is weak. It can be seen that the proposed method has strong portability and applies to other textile image data sets or image data sets with the same characteristics.

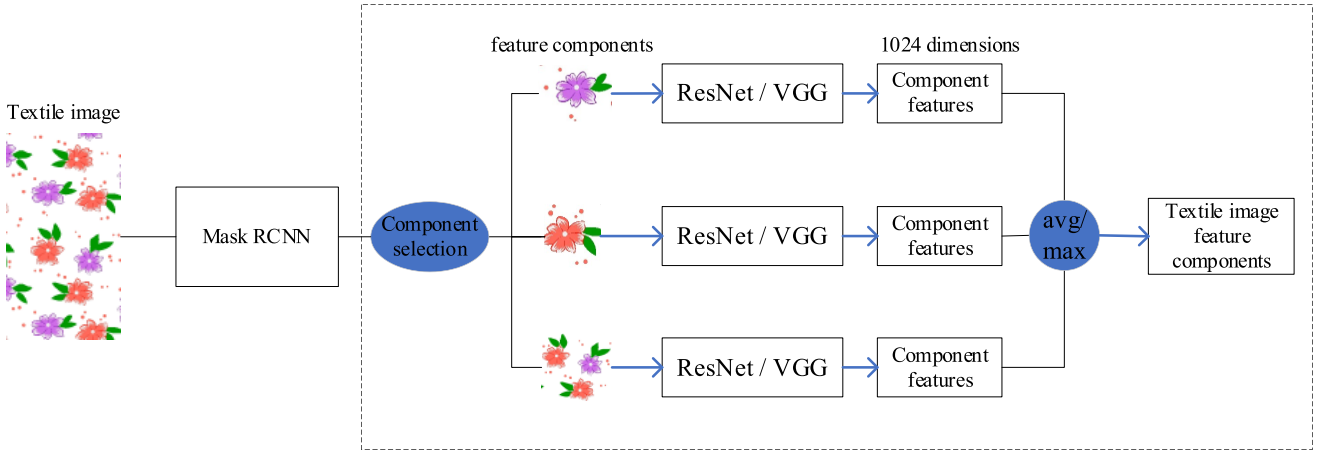


FIGURE 1. Network architecture of textile image fine-grained retrieval.

The overall process of the textile image fine-grained retrieval method is shown in Figure 1. The image is fed into the improved pre-trained Mask R-CNN model to extract multiple ''feature components'' in the textile image, and then the depth feature of the textile ''feature components'' is extracted through ResNet or VGG model, and the extracted depth feature is fused as the ''fingerprint'' of the textile. Through the image feature deletion mechanism of Mask R-CNN, the ''fingerprint'' can effectively eliminate the interference of the inherent large area background and a large number of internal repetitive features of the textile image, and improve the retrieval efficiency of the textile ''feature component''.

B. MASK R-CNN OVERVIEW

Mask R-CNN model is a target detection network framework developed by He Kaiming [24] and others based on the general framework of Faster R-CNN [25] in 2017. It is one of the excellent algorithms in object recognition and segmentation at present. Mask R-CNN is mainly divided into the following five parts: ResNet backbone network, feature extraction network feature pyramid networks (FPN) [33], regional proposal network (RPN) [25], ROIAlign, and image output network. The Mask R-CNN model uses the ResNet backbone network, which well solves the training difficulties caused by network depth, and the network performance is good. The Mask R-CNN algorithm first feeds the image to be detected into the pre-trained ResNet backbone network to obtain depth features and then performs feature fusion on each convolution layer. Five feature maps P2, P3, P4, P5, and P6 corresponding to the anchor side length of 32, 64,128,256,512 are sequentially input into the RPN. After the RPN generates several candidate regions, the NMS is used to filter out accurate candidate regions, Then the fixed dimension feature vector is obtained through the ROI Align layer, and the target object is located, classified, and segmented simultaneously through the full connection layer. The main network is shown in Figure 2.

The region recommendation network RPN is a full convolution neural network, which can accept image input of different sizes. After a full convolution operation, the features of multiple regions can be obtained. Unlike the traditional convolutional neural network, the full convolutional neural network has no full connection layer. Different from the traditional selective search method of candidate region generation, RPN takes the feature map as the input, uses the sliding window method to generate multi-scale anchor points in the feature map, judges the target region category through the classification layer, and obtains the target location through the boundary regression layer.

In this paper, a 3×3 convolution kernel is used to extract 256-dimensional feature vectors for each anchor point. The loss function of RPN is composed of the classification loss function and boundary regression loss function. The loss function is defined as:

$$L\left(\{p_i\}, \{t_i\}\right) = \frac{1}{N_{cls}} \sum_i L_{cls}\left(p_i, p_i^*\right) + \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}\left(t_i, t_i^*\right) \quad (1)$$

where i is the index representative of the anchor point in each batch of training and p_i is the classification prediction probability of the anchor point i, p_i^* is the true value. If the candidate box is a positive label, set its true value p_i^* to 1. If it is a negative sample, p_i^* is 0. t_i represents the four parameterized coordinate vectors of the positive sample anchor point, t_i^* is the coordinate vector corresponding to the real area, N_{cls} is the binary loss function (whether it is the target), and L_{reg} is the boundary regression loss function. N_{cls} and N_{reg} are the normalization coefficients of the classification loss function and the boundary regression loss function, respectively. λ is the weight balance parameter between the two loss functions in order to make the proportion of the two loss functions tend to be the same. The three parameters are generally 256, 2400, and 10.

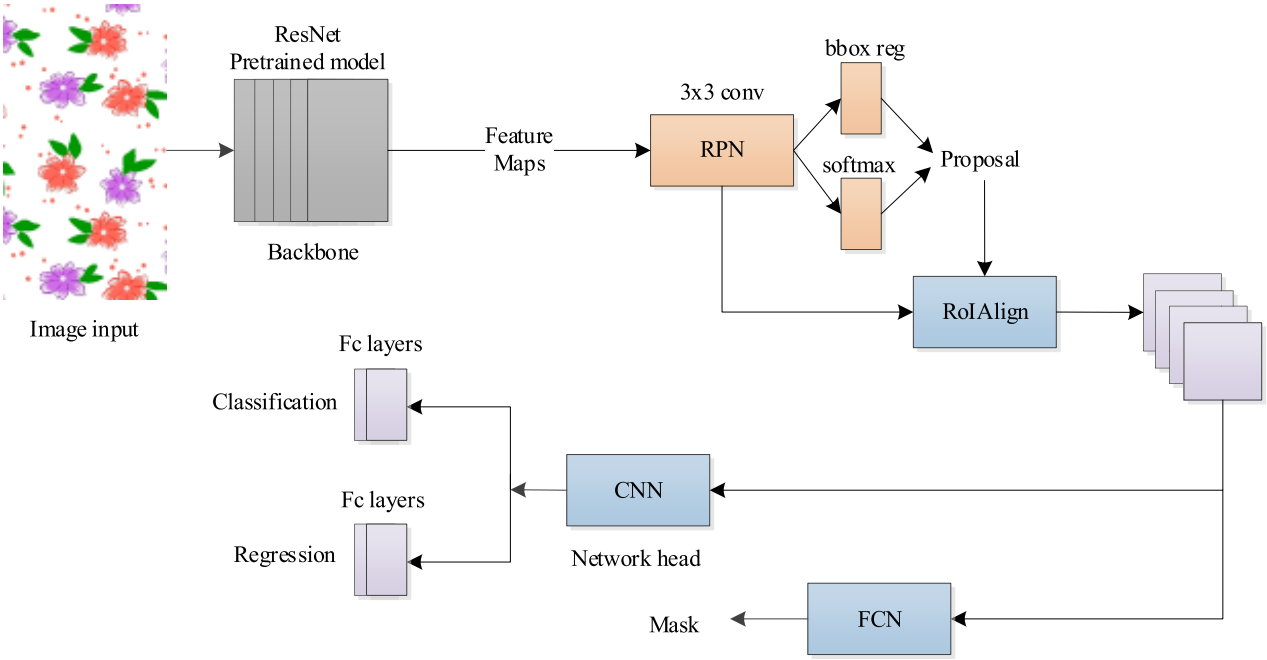


FIGURE 2. Mask R-CNN model structure.

C. ACQUISITION OF IMAGE "FEATURE COMPONENT"

The method proposed in this paper extracts multiple fine-grained textile "feature components" with the most features in the textile image. Figure 1 shows the algorithm flow of partial fine-grained features extracted from textile fabric images as "feature components". The resolution of textile images is between 1024 * 1980 and 4096 * 4096. As mentioned above, the feature pyramid network in the Mask R-CNN model can obtain input images of any resolution and generate an output of the corresponding size, which can effectively compress the size of textile images needed for retrieval. The proposed method uses the RPN regional recommendation network to locate the fine-grained feature " feature component" of textiles in the fine-grained image for use in the subsequent descriptor selection process. Each textile image has 1-4 "feature components".

In the feature pyramid network, the key of the "feature component" extraction method is to use the "feature component" that repeatedly appears in the textile image to enhance the probability value of the output anchor point, and prevent the target from missing detection due to the NMS algorithm mistakenly deleting the "feature component". Due to the low matching between the textile image and the pre-trained mask R-CNN model, the missed detection rate is high, and lower the ROI detection threshold DETECTION_MIN_CONFIDENCE is 0.2, making more "feature components" available for selection. The proposed method will return about 10 candidates " feature components" for each image. According to the size of the generated label and frame, objects similar to the label and frame are regarded as the same "feature components", and the duplicate "feature components" are removed.

The image retrieval feature fusion method designed in this paper, that is, the method of using "feature components" to generate textile "fingerprint", is as follows: For multiple "feature components" extracted from a textile image using Mask R-CNN model, record each "feature component" to extract convolution features again using convolution neural networks, such as ResNet and VGG model, and fuse the depth features extracted from each component by weighted average, The "fingerprint" obtained by weighted fusion of corresponding textile images can be expressed as $\bar{x} = \frac{x_1+x_2+x_3+\ldots+x_k}{k}$. Among them, the convolution feature calculation method is to use the maximum pooling layer, the average pooling layer, or the layer with the best effect in the ResNet or VGG model to calculate the depth feature. The quality of the feature calculation method is determined by experimental comparison.

The method presented in this paper is called Mask R-CNN-FF(feature fusion). When retrieving, the retrieved images are sorted by cosine distance with the textile image "fingerprint" in the database, the whole retrieval process is completed, and the corresponding Topk and mAP accuracy rates are calculated.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

In this paper, Pytorch is used to implement CNN. ImageNet pre-trained model is used as the unsupervised retrieval model for the whole network, and the deep convolution features required for image retrieval are extracted. The final feature vector used is 1024 dimensions. All experiments were carried out on Dell T630 Tower Server, which is equipped with GPU: Tesla K40; Memory capacity: 64G DRAM; CPU model: E5 processor*2.

Algorithm 1 Extraction of Textile Image "Feature Components"

Input: Textile Image M=w*h;
Output: Textile "fingerprint"

1. Adjust R OI detection threshold DETECTION_in Mask R-CNN model MIN_ CONFIDENCE is 0.2;
2. **For** each textile image:
3. Feed the image to be detected into the pre-trained convolution neural network to get the depth features;
4. Feature fusion is performed on each convolution layer to get a feature map.
5. Input the extracted feature map into the RPN to generate several candidate regions;
6. Use non-maximum suppression (NMS) to filter out exact candidate regions;
7. Feed the obtained "feature components" into the ROIAlign layer and map them to fixed dimension vectors;
8. Mark all the candidate areas and find all the "feature components" in the image M, and mark them as T= T1, . . . , T n| n < 10;
9. Calculate the area of all the feature components in T and select those with the same area.
10. Identify "feature components" of the same area that are duplicated, eliminate duplicates and place the resulting duplicate "feature components" in List P first;
11. **If** 10 "feature components", the top three tag targets are not in P:
12. Put the target in list P;// Place high recognition "feature components " in the list P.
13. Output list P is used as the textile "feature components";

The purpose of this paper's textile image retrieval task is to compare local textile photographs to images in the database in order to return related images that users may need for selection and redesign.

A. IMAGE RETRIEVAL TASKS

The textile image fine-grained retrieval method proposed is suitable for image retrieval tasks that contain fine-grained features that occur repeatedly and a background that needs to be deleted and enlarged. The textile image database images of the validation experiments were selected from the textile images displayed at the International Textile Fabric Expo in China Textile City. The images are rich and varied. Most of the textile images have the fine-grained feature of repetition. There is a high rate of missing "feature components" because the pre-trained Mask R-CNN model and the textile image do not match well. The 1100 textile images used generated 500 effective textile "feature components". Most textile images extracted 1-4 different numbers of textile image "feature components" to detect the matching degree of "feature components". An example of the data in the textile image database is shown in Figure 3.

B. EVALUATING INDICATOR

We choose the classic evaluation index mean average precision (mAP) in image retrieval as the evaluation criterion for the fine-grained image retrieval task of the image library. Take the area enclosed by each query precision-recall rate curve and coordinate axis as the accuracy of the query, and calculate the average of the accuracy of all queries to get mAP, with the maximum value of 100%. Intuitively, mAP represents the average value of multiple image retrieval accuracy. For the first k returned results of each image retrieval result, record the returned results as top k; The sum of the top j images retrieved for the time is recorded as t_{ij}.

The corresponding query precision AP_i for the ith time can be expressed as:

The precision AP_i expressed as:

$$AP_i = \left(\frac{t_{i,1}}{1} + \frac{t_{i,2}}{2} + \frac{t_{i,3}}{3} + \cdots + \frac{t_{i,k_i}}{k_i}\right)/k_i \quad (2)$$

where k_i is no more than k. The average query precision is expressed as:

$$mAP = (AP_1 + AP_2 + AP_3 + \cdots + AP_n)/n \quad (3)$$

C. SIMILARITY MEASURE

Experiment on real textile image data, use different similarity measures to obtain the nearest search similarity score between images, get a ranking of database images from the most similar to the least similar, and then calculate the retrieval accuracy of the results.

The first comparative experiment is to use traditional manual features as the baseline model to explore the differentiation of color and texture to image features, and to explore the differentiation of color histogram, direction gradient histogram, and edge histogram to textile images. Table 1 uses manual features to observe the feature distribution of the dataset. Different feature distances are measured by the common cosine distance, and the results are shown in Table 1.

The findings demonstrate that while texture-based algorithms perform poorly in terms of retrieval, color distribution techniques work well. The majority of textile images are different varieties of flowers, the texture similarity is quite high, and a significant portion of image retrieval exhibits the characteristics of a small gap between classes. These factors may be the primary causes of the textile photos' excellent color distribution distinction.

The effectiveness of the supervised method depends largely on the data set used in the training model, while the proposed unsupervised textile image fine-grained retrieval method can better use the convolution features extracted from the pre-trained and optimized CNN model to represent the local image, and fuse the local image features, without further supervised training. Considering that most textile image designs are obsolete and data sets with labels are difficult to collect, it is unrealistic to adjust the models separately for different tasks.

The proposed unsupervised textile image fine-grained retrieval method is very suitable for this situation. The "feature component" preserves the important features of the retrieval object, and can significantly suppress the impact of background noise and the repeatability of the "feature

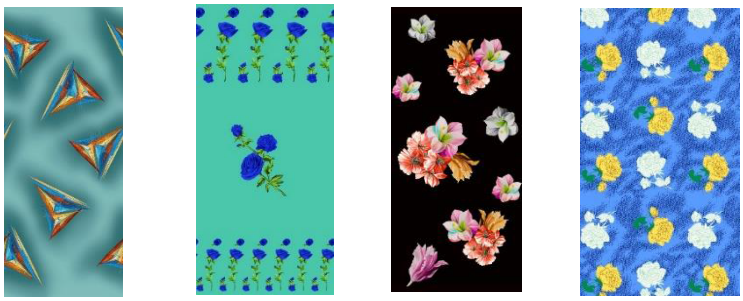


FIGURE 3. Textile image.

TABLE 1. The effectiveness of hand-designed image features.

models	top1	top5	mAP
ColorHistogram	0.719	0.736	0.682
HOG	0.085	0.125	0.104
Edge	0.012	0.019	0.039

component'', making better use of the convolution features extracted from the pre-trained CNN model.

The backbone network used by the latest model is ResNet [25], [26], [27], [28], [29], [30], [31], [32]. In order to verify the effectiveness of the proposed textile fine-grained retrieval method, a variety of control experiments were designed for comparison. ResNet and VGG model can represent the performance of current CNN models. The latest model also uses ResNet pre-trained model as the backbone network. This paper selects some CNN pre-trained models with superior performance as the baseline for this method. ResNet-101 and VGG-16 convolutional neural networks based on ImageNet pre-trained were used as the baseline.

The comparison results of retrieval accuracy of textile image retrieval using the maximum pooling feature under different measurement methods are shown in Table 2. The results show that the method of cosine measure is the most effective for similarity measurement, so the retrieval effect of cosine measure distance is the criterion in the following experimental comparison.

The comparison results of retrieval effects of different pooling methods in textile image data sets are shown in Table 3. The results show that the feature retrieval accuracy of maxpool extraction is the highest, followed by avgpool, which is consistent with the feature that the classification features in the textile image data set are mainly concentrated in multiple parts. In addition, the experimental results show that the accuracy of the depth feature of the FC layer in CNN is lower.

In the convolutional neural network, RGB three-channel image features are more distinguishable than grayscale image features, so subtracting the RGB three-channel mean value of ImageNet from the ImageNet pre-trained model is conducive to the extraction of color features. The Table 2 experiments is

TABLE 2. Comparison of retrieval accuracy using maximum pooling feature under different measures.

models	top1	top5	mAP
ResNet-101(max-cosine)	0.837	0.861	0.758
ResNet-101(max-square)	0.829	0.847	0.728
ResNet-101(max-d1)	0.833	0.847	0.731
ResNet-101(max-d2-norm)	0.455	0.547	0.500

TABLE 3. The accuracy comparison among different pooling methods.

models	top1	top5	mAP
ResNet-101(max-cosine)	0.837	0.861	0.758
ResNet-101(avg-cosine)	0.817	0.837	0.730
ResNet-101(fc-cosine)	0.605	0.621	0.547

to use the depth features of ImageNet pre-trained ResNet-101 and ResNet-50 for image retrieval of textile images.

D. ANALYSIS OF EXPERIMENTAL RESULTS

The maxpool pooling layer feature and the cosine distance is used to measure the distance of different features. The retrieval results are compared with the textile fingerprint fusion method proposed in this paper, and the results are shown in Table 4. It can be seen from the results that the fingerprint fusion method has greatly improved the performance of the pre-trained model, by about 15% on the top 1 index, and by about 23∼28% on the mAP index.

The Table2 and Table3 experiments used the depth feature of ImageNet pre-trained VGG16 to retrieve the image of textiles, as set above. The retrieval results are compared with the textile fingerprint fusion method proposed in this paper. The fusion method of depth features of textile ''feature components'' has obvious performance improvement, with the retrieval accuracy of top1 and top5 improved by 7-8%, and the retrieval performance in mAP improved by 18%. The experiment shows that the method in this paper is universally applicable to the normal CNN model.

The main solution to this issue is to more accurately find and extract the ''feature components'' that textile image identification relies on. It's important to note that the fusion and

TABLE 4. The effectiveness of textile image fine-grained retrieval with Resnet-101.

models	top1	top5	mAP
ResNet-101	0.817	0.837	0.729
Mask R-CNN-FF(ResNet-101)	0.953	0.965	0.958
ResNet-50	0.764	0.786	0.671
Mask R-CNN-FF(ResNet-50)	0.923	0.951	0.952

TABLE 5. The effectiveness of textile image fine-grained retrieval with Vgg-16.

models	top1	top5	mAP
VGG-16	0.862	0.872	0.733
Mask R-CNN-FF(VGG-16)	0.938	0. 952	0.954

use of ''feature components'' reduce the need for extraneous positioning information in textile images and increases the retrieval precision of textile images.

V. CONCLUSION

This paper proposes a textile image fine-grained retrieval method, which can effectively eliminate the interference of the inherent large-area background and a large number of internal repetitive features of the textile image to improve the retrieval efficiency of the textile ''feature component'' by extracting the fineness of the textile image. The key point is to use Mask R-CNN to extract the ''feature component'' in the textile image and generate the weighted and aggregated depth convolution feature according to the ''feature component'', which makes it possible to pay more attention to the important and repeated identification features of the retrieval object, and effectively improve the efficiency of image retrieval. This method can greatly improve the precision of fine-grained image retrieval tasks with repetitive features in textile images without relying on a large number of external data set training. The proposed unsupervised weighted aggregate textile feature is very suitable for the situation where it is difficult to collect the detailed annotated training data set in this field. In this paper, we found that the depth features extracted by CNN still have good retrieval efficiency after being weighted and averaged equally.

The method proposed in this article has the following shortcomings: the fine-grained feature extraction of textile images relies on the target detection method of Mask R-CNN, and the target detection performance of the latter still needs to be improved.

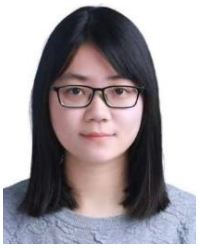
REFERENCES

[1] J. Mao and A. K. Jain, ''Texture classification and segmentation using multiresolution simultaneous autoregressive models,'' *Pattern Recognit.*, vol. 25, no. 2, pp. 173–188, Feb. 1992.
[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, ''ImageNet: A large-scale hierarchical image database,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[3] K. Simonyan and A. Zisserman, ''Very deep convolutional networks for large-scale image recognition,'' in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–4. [Online]. Available: https://arxiv.org/abs/1409.1556
[4] K. He, X. Zhang, S. Ren, and J. Sun, ''Deep residual learning for image recognition,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[5] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, ''CNN features off-the-shelf: An astounding baseline for recognition,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2014, pp. 806–813.
[6] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, ''Neural codes for image retrieval,'' in *Proc. Eur. Conf. Comput. Vis.*, Zürich, Switzerland, 2014, pp. 584–599.
[7] A. S. Razavian, J. Sullivan, S. Carlsson, and A. Maki, ''Visual instance retrieval with deep convolutional networks,'' *ITE Trans. Media Technol. Appl.*, vol. 4, no. 3, pp. 251–258, 2014.
[8] J. Y.-H. Ng, F. Yang, and L. S. Davis, ''Exploiting local features from deep networks for image retrieval,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 53–61.
[9] F. Radenovi, G. Tolias, and O. Chum, ''CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples,'' in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 3–20.
[10] F. Radenović, G. Tolias, and O. Chum, ''Fine-tuning CNN image retrieval with no human annotation,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1655–1668, Jul. 2019.
[11] Z. Hao and W. Jianxin, ''A survey on unsupervised image retrieval using deep features,'' *J. Comput. Res. Develop.*, vol. 55, no. 9, pp. 1827–1842, 2018, doi: 10.7544/issn1000-1239.2018.20180623.
[12] M. J. Swain and D. H. Ballard, ''Color indexing,'' *Int. J. Comput. Vis.*, vol. 7, pp. 11–32, Nov. 1991.
[13] D. G. Lowe, ''Distinctive image features from scale-invariant keypoints,'' *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
[14] H.-F. Yang, K. Lin, and C.-S. Chen, ''Cross-batch reference learning for deep classification and retrieval,'' in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 1237–1246.
[15] Y. Liu, Y. Guo, S. Wu, and M. S. Lew, ''DeepIndex for accurate and efficient image retrieval,'' in *Proc. 5th ACM Int. Conf. Multimedia Retr.*, Jun. 2015, pp. 43–50.
[16] L. Xie, J. Wang, B. Zhang, and Q. Tian, ''Fine-grained image search,'' *IEEE Trans. Multimedia*, vol. 17, no. 5, pp. 636–647, May 2015.
[17] X. Wei, J. Luo, J. Wu, and Z. Zhou, ''Selective convolutional descriptor aggregation for fine-grained image retrieval,'' *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2868–2881, Jun. 2017.
[18] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, ''The application of two-level attention models in deep convolutional neural network for fine-grained image classification,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 842–850.
[19] M. Simon and E. Rodner, ''Neural activation constellations: Unsupervised part model discovery with convolutional networks,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1143–1151.
[20] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, ''The Caltech-UCSD birds-200-2011 dataset,'' California Inst. Technol., Pasadena, CA, USA, Tech. Rep. CNS-TR-2011-001, 2011.
[21] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, ''3D object representations for fine-grained categorization,'' in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Dec. 2013, pp. 554–561.
[22] Y. Gong, L. Wang, R. Guo, and S. Lazebnik, ''Multi-scale orderless pooling of deep convolutional activation features,'' in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 392–407.
[23] M. Paulin, M. Douze, Z. Harchaoui, J. Mairal, F. Perronin, and C. Schmid, ''Local convolutional features with unsupervised training for image retrieval,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 91–99.
[24] K. He, G. Gkioxari, P. Dollár, and R. Girshick, ''Mask R-CNN,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 386–397, Feb. 2020.
[25] S. Ren, K. He, R. Girshick, and J. Sun, ''Faster R-CNN: Towards real-time object detection with region proposal networks,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[26] Y. Chen, Y. Bai, W. Zhang, and T. Mei, ''Destruction and construction learning for fine-grained image recognition,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5157–5166.
[27] X.-S. Wei, C.-W. Xie, J. Wu, and C. Shen, ''Mask-CNN: Localizing parts and selecting descriptors for fine-grained bird species categorization,'' *Pattern Recognit.*, vol. 76, pp. 704–714, Apr. 2018.
[28] H. Wei, M. Zhu, B. Wang, J. Wang, and D. Sun, ''Two-level progressive attention convolutional network for fine-grained image recognition,'' *IEEE Access*, vol. 8, pp. 104985–104995, 2020.
[29] X. Gao, F. Yang, T. Chen, and J. Si, ''Chinese character components segmentation method based on faster RCNN,'' *IEEE Access*, vol. 10, pp. 98095–98103, 2022, doi: 10.1109/ACCESS.2022.3206832.

[30] H. Zhang, J. Xue, and K. Dana, "Deep TEN: Texture encoding network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 708–717.
[31] Z. Chen, F. Li, Y. Quan, Y. Xu, and H. Ji, "Deep texture recognition via exploiting cross-layer statistical self-similarity," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5231–5240.
[32] B. Peng, M. Chi, and C. Liu, "Non-IID federated learning via random exchange of local feature maps for textile IIoT secure computing," *Sci. China Inf. Sci.*, vol. 65, Jun. 2022, Art. no. 170302.
[33] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," 2016, *arXiv:1612.03144*.

SHUTAO TAN was born in Jiaxing, Zhejiang, China, in 1996. He received the bachelor's degree in software engineering from Hangzhou Normal University, in 2018, and the master's degree in software engineering from the Zhejiang University of Technology, in 2021. He is currently with the Information Technology Department, The First Hospital of Jiaxing Affiliated Hospital. His research interests include image processing and computer vision.

LIANG DONG was born in Jiaxing, Zhejiang, China, in 1984. He received the master's degree from the Hangzhou University of Electronic Science and Technology, in 2009. He is currently a Senior Engineer in the major of electronic information engineering. He is also with the Information Technology Department, The First Hospital of Jiaxing Affiliated Hospital.



MIN ZHANG received the M.E. degree in computer science and technology from the Zhejiang University of Technology. Since 2021, she has been with the Jiangsu Vocational College of Finance and Economics, Huai'an, China. She is currently a lecturer. Her current research interests include the next generation wireless LAN and heterogeneous networks.

YE ZHANG was born in Jiaxing, Zhejiang, China, in 1986. She received the M.S. degree in clinical medicine from the Medical School of Zhejiang University, Hangzhou, in 2012. She is currently a General Practitioner with The First Hospital of Jiaxing Affiliated Hospital, Jiaxing University. Her research interests include internet + medical treatment and undifferentiated diseases in general practice.