Hindawi
Journal of Sensors
Volume 2021, Article ID 6631029, 9 pages
https://doi.org/10.1155/2021/6631029

Research Article

An Ensemble Deep Neural Network for Footprint Image Retrieval Based on Transfer Learning

Dechao Chen,[1] **Yang Chen**,[2] **Jieming Ma**,[3] **Cheng Cheng**,[2] **Xuefeng Xi**,[2,4] **Run Zhu**,[5] **and Zhiming Cui**[2,4]

[1]*School of Geography Science and Geomatics Engineering, Suzhou University of Science and Technology, Suzhou 215009, China*
[2]*School of Electronic and Information Engineering, Suzhou University of Science and Technology, Suzhou 215009, China*
[3]*School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou 215000, China*
[4]*Virtual Reality Key Laboratory for Intelligent Interaction and Application Technology of Suzhou, Suzhou 215000, China*
[5]*Public Security Bureau of Kunshan City, Kunshan 215300, China*

Correspondence should be addressed to Dechao Chen; dcchen1205@tom.com and Xuefeng Xi; xfxi2009@qq.com

Received 29 October 2020; Revised 6 November 2020; Accepted 26 February 2021; Published 18 March 2021

Academic Editor: Marimuthu Palaniswami

Copyright © 2021 Dechao Chen et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

As one of the essential pieces of evidence of crime scenes, footprint images cannot be ignored in the cracking of serial cases. Traditional footprint comparison and retrieval require much time and human resources, significantly affecting the progress of the case. With the rapid development of deep learning, the convolutional neural network has shown excellent performance in image recognition and retrieval. To meet the actual needs of public security footprint image retrieval, we explore the effect of convolution neural networks on footprint image retrieval and propose an ensemble deep neural network for image retrieval based on transfer learning. At the same time, based on edge computing technology, we developed a footprint acquisition system to collect footprint data. Experimental results on the footprint dataset we built show that our approach is useful and practical.

1. Introduction

The research of image retrieval began in the 1970s. Initially, it was based on text-based image retrieval; the characteristics of the image are described by using text [1, 2]. In the 1990s, content-based image retrieval appeared, i.e., the image color and texture were analyzed, and some shallow classifiers such as SVM and other technologies for image retrieval were used to improve the accuracy of the search [3, 4]. But these methods still cannot solve the semantic gap [5, 6]. With the research and development of deep learning, the convolutional neural networks (CNN) [7–9] have performed well in image retrieval and recognition in recent years. With various types of image recognition competitions held, such as ImageNet [10] and Kaggle [11], multiple models changed by CNN have performed well, such as AlexNet [12], VGG, GoogLeNet [13], ResNet, and DenseNet. These models have dominated the field of computer vision by their superior recognition accuracy. Applying CNN to perform image retrieval and identification of content has high reliability.

At present, face recognition [14], fingerprint, palm print automatic identification, and retrieval technologies have been already well-developed in the investigation of criminal cases in public security. Footprints, as another significant trace of crime scenes, also have an essential role. However, the traditional footprint search and identification work often require a lot of human resources, time, and experience. Manual retrieval under large data volume is also prone to mistakes. Therefore, the automatic footprint retrieval system has substantial application requirements. However, based on the traditional shallow machine learning method, the footprint image retrieval system is not only time-consuming, but also the accuracy needs to be improved.

To address the above issues, we explore the effect of convolution neural networks on footprint image retrieval and propose a novel approach. Our work has three significant

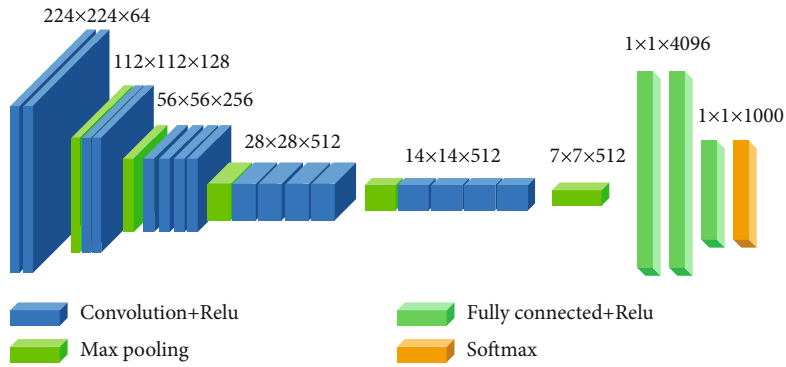

Figure 1: Structure of the VGG19 model.

contributions. Firstly, an ensemble deep neural network for feature extraction of footprint images was proposed. To the best of our knowledge, this was the first attempt at addressing the feature extraction of footprint images. By connecting the footprint image features extracted from three different models, we can get richer features for retrieval. This method can obtain excellent and stable search results when using the cosine distance. Secondly, transfer learning is used to pretrain the model, which makes our way have an outstanding performance on the footprint dataset with smaller data size. The experiments show that the approach is useful and practical. Finally, a footprint image dataset is initially constructed, which plays an essential role in the footprint image retrieval task for public social security.

2. Related Works

CNN is an improved feed-forward neural network based on a fully connected neural network [15], which dramatically reduces the parameters that need to be calculated. Its network structure is mainly composed of a convolution layer, pool layer, full connection layer, and some activation functions. In the convolution layer, the features of the input image are extracted using a convolution operation to use different convolution kernels on the input image [16–18]. The results of the convolution are changed nonlinearly to obtain the output as the input of the next layer. If the feature map is relatively large after convolution, the dimension can be reduced by the pooling operation. The pooling operation has two types, the maximum pooling operation [19] and the mean pooling operation. After the pooled layer, the depth of the output image is unchanged, and it is still the number of feature maps. The pooling layer can prevent overfitting of the model to a certain extent, and it is more convenient to connect all the neurons with weights and obtain the same output connection as the traditional neural network.

Based on the development of CNN, there have been many typical CNN networks, such as the LeNet-5 model, the AlexNet model, the VGGNet model, the GoogleNet model, ResNet model, and DenseNet. Here, we choose three typical models for experiments.

2.1. VGGNets. The VGGNet [20] model was invented by the University of Oxford's Visual Geometry Group and achieved a high ranking in the international image recognition competition. From the overall design perspective, VGGNet has a similar design style to AlexNet. It is a more in-depth network architecture built on AlexNet. Its network structure is mainly composed of 5 groups of convolution layers, five layers of pooling layers, and three layers of fully connected layers. There are 2 or 3 convolution operations in each convolution layer, which are pooled and then convolved by continuous convolution and finally output through a fully connected layer. Figure 1 shows the structure of the VGG19 model.

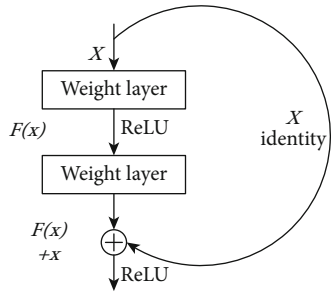

Figure 2: A residual block.

2.2. ResNet. The deeper neural network can extract more abundant image feature, but at the same time, the problem of gradient disappearance and explosion becomes more prominent. It leads to hard training. He et al. proposed ResNet [21] which solves this problem to some extent, and this model achieved first place in the ILSVRC2015. The shortcut structure in ResNet enables the image information of the front layer to be directly transmitted to the deeper layers, thus protecting the integrity of image feature information. Figure 2 shows a typical residual block used in ResNet.

2.3. DenseNet. The basic idea of DenseNet is similar to ResNet. That is, they both establish the connection between the previous layer and later layer. Compared to ResNet, DenseNet [22] proposed a dense link, by which all layers can connect. Explicitly, each layer accepts all of its previous layers as an extra input.

Figure 3 shows the dense connection of ResNet. Another significant characteristic of DenseNet is that it realizes the feature reuse through the connection of features on the channel.

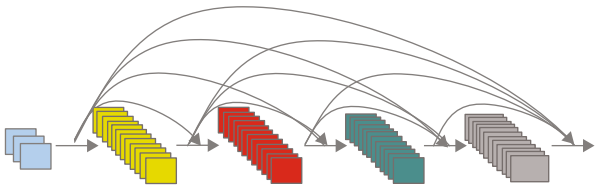

Figure 3: Structure of the DenseNet model.

These characteristics make DenseNet have fewer parameters and calculation costs than ResNet, but it has better performance than ResNet. Figure 3 shows a dense connection. In our experiment, DenseNet121 is used.

2.4. Ensemble Learning. In the supervised learning algorithm of machine learning, our goal is to learn a stable model that performs well in all aspects, but the actual situation is often not so ideal. Sometimes, we can only get models that perform better in some respects but may not perform well in other ways. Ensemble learning solves this problem very well. It combines several weak supervisory models to complete the learning task and finally get a more stable and comprehensive model. Due to the difference between the footprints of the image feature extraction of different models, there eventually will appear a variety of search results. To preserve and be compatible with these differences to obtain a more precious footprint image feature description, this paper refers to the idea of integrated learning. We trained the VGG, ResNet, and DenseNet networks on the footprint image dataset and, in the test phase, connected the outputs of the three models to form the final footprint image features. Experimental results show that the method is stable.

2.5. Transfer Learning. Massive data is a crucial factor when deep learning achieves such excellent performance in various fields. Without the support of a large amount of data, there are no attractive deep models. But in many cases, we cannot get a massive amount of data to train the model. In this case, transfer learning has solved this problem to some extent. The pretraining model is used as a checkpoint to start training to generate a neural network model to support new tasks. This method is usually called transfer learning. Its advantage is that it does not need to start designing and training a new network again. Instead, it is based on the trained network model, and the parameters and knowledge migration are performed on it. Only a small amount of computing resources and training time can support a new task. There are various feature data and weight information in the pretraining model, some are feature data and weight information that are strictly related to the object identified by the classification, and some are common feature data and information, which can be used for different tasks or shared between objects; transfer learning aim is to migrate those common feature data and information, so as to avoid learning this knowledge again and achieve fast learning. In our experiments, we found that high-quality footprint images that can be used in experiments are not particularly large after data cleaning, so we use migration learning methods to compensate for the lack of data. The experimental results show that transfer learning can be applied to our mission well.

3. Our Model

We have combined with some deep neural network model outstanding performance in face recognition and person reidentification [23–25]; VGGNet19, ResNet50, and DenseNet121 are used in our experiment. We extracted the footprint image by three models, respectively, and do some comparative experiments. Since the features of the footprint images obtained by different models will be various, we propose an ensemble deep neural network to fuse this difference.

We connect the features extracted by three different models to construct more special features of the footprint image for experiments. At the same time, we train models based on transfer learning with the footprint image dataset being built by us with PSBKC (Public Security Bureau of Kunshan City) in China.

Our approach includes four critical steps:

(1) Data cleaning and preprocessing

(2) Model structure fine-tuning and model training

(3) Extract the characteristics of footprint datasets through some deep neural network models and our ensemble neural network to establish the feature vector index

(4) Calculate the distance of footprint images and output retrieval results

The overall framework of the footprint image retrieval and matching method is shown in Figure 4.

3.1. Preprocessing. Preprocessing of data plays a vital role in computer vision. To obtain better experimental results, we combined with the experimental dataset of this project. The input picture needs to be preprocessed. First, to facilitate the extraction of feature parameters, an input footprint of approximately 1000×2000 pixels is adjusted to 128×256. Then, we pad the image and random crop the image into 128×256. On the one hand, this approach is indeed to increase the amount of sample data. On the other hand, because the footprint image in the dataset is rectangular, in order to maintain its aspect ratio for better feature extraction, we resize the original image to 128×256, which can save some graphic card memory during training.

At the same time, the data are normalized. For this dataset, a sample-by-sample mean reduction method is used for normalization. The specific operation method is to convert the input image into a matrix and subtract the statistical average of the data from each sample, that is, to calculate the average value of each image sample and then subtract the corresponding average cost of each sample and center on the average pixel.

For image data, this normalization method can remove the average brightness value of the image, thereby reducing the interference of the background effect of the image on

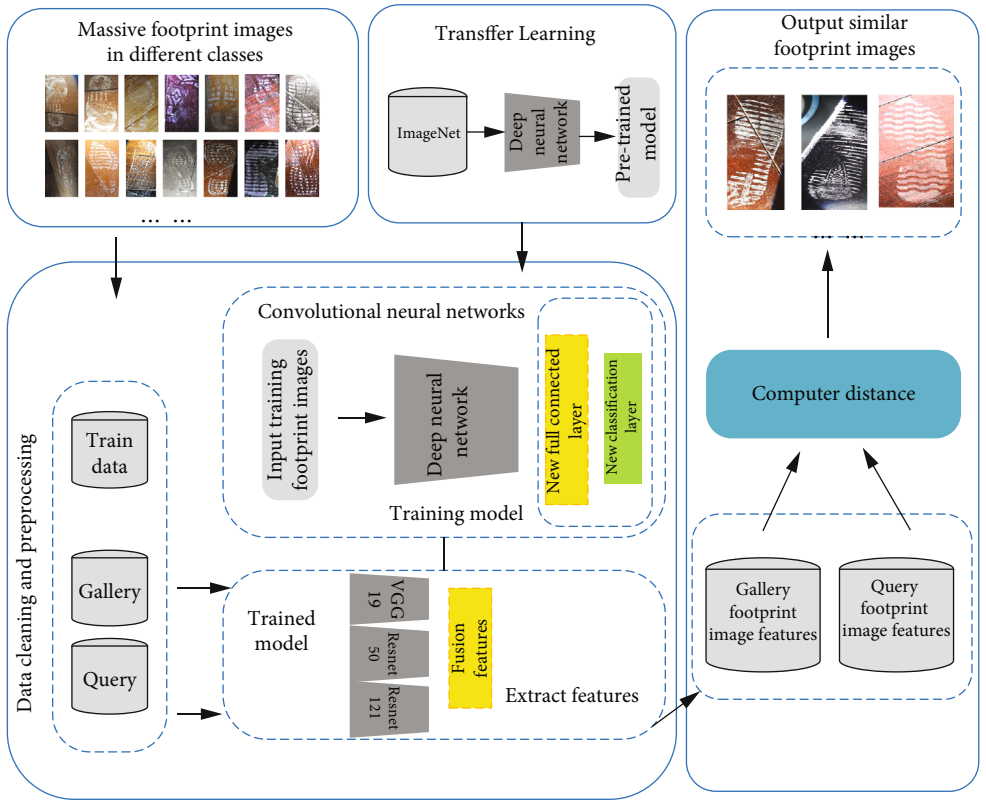

Figure 4: A framework of deep learning with application to footprint image retrieval.

the experiment. Finally, the image was converted to tensor and normalize to 0-1.

3.2. Model Fine-Tuning and Training. In this work, we modified the fully connected (FC) layers of the three models and the subsequent part of the fully connected layer. The new FC layer and classification layer are defined as the order of linear, batch normalization, ReLU, and linear. In ResNet50 and Densenet121, we modified the full connected layer as 512 and added a new classification layer. Here, the purpose and motivation of modifying the fully connected layers are to fit the two models. At the same time, we preserve the first fully connected layer of VGG19 and remove the second one; a new classification layer was also added. For the added layer, parameters are initialized with Kaiming normal [26]; other parameters of the model are pretrained on ImageNet. A training set of footprint images has trained the three models.

We also use the average adaptive pool. Due to the height of footprint images being larger than its width, we need to specify the pooling kernel. Thus, the average adaptive pool is more comfortable to implement. During the training process, we set up 60 epochs and used SGD [27] for backpropagation utilizing the parameters pretrained on ImageNet, as shown in the following equation:

$$\theta = \theta - \eta \bullet \nabla_\theta J\left(\theta; x^i, y^i\right), \tag{1}$$

where J is the objective function that SGD will optimize, x^i and y^i represent training sample and training label, and θ denotes model parameters, i.e., weights and biases.

We use dynamic learning rate in the experiment; the equation of learning rate is defined as follows:

$$\mathrm{lr} = \mathrm{lr}_0 \times \lambda^{\mathrm{epoch/step\_size,}} \tag{2}$$

where lr is the current learning rate; lr_0 is the initial learning rate; λ is the learning rate decay factor; step_size is the step of change learning rate; epoch is the number of the current epoch.

At the same time, the batch size was 32. The initial learning rate was 0.1, and the learning rate will be updated after every 40 iterations. The final model accuracy rate can reach more than 98%.

3.3. Feature Extracting. In image retrieval based on deep learning, the extraction of feature vectors is a critical step. Through the comparative analysis of a variety of current mainstream convolutional neural network models, combined with the actual situation of the subject, we use the three different models for experiments. After training these three models on our footprint images, we preserve the rest of the network structure except for the classification layer. Each image in the footprint gallery is extracted by the network model. When we use ResNet50 and DenseNet121 to extract features, the dimension of each image is 512, and the output of VGG19 is 4096.

Inspired by ensemble learning, while using these three models to extract features separately for comparison experiments, we performed an additional experiment to fuse these different essential information. We input a footprint image

and extract the features through three models at the same time. The feature vectors obtained by the models are connected and merged. The specific schematic diagram is shown in Figure 5. It was considered that the feature representations of the footprint images extracted by different models are different, but these different feature representations contain some essential information components. To fuse this different critical information and retain more features for the final retrieval, we connect the feature vectors extracted by these various models to get the final feature vector, and the dimension of it is 5120.

After all the features of the image are extracted, a feature vector index database for the footprint image database is established.

3.4. Metric Learning. Calculating the distance between samples is hot research [28]. At present, typical distance metrics such as Euclidean distance, cosine similarity distance, and Hamming distance have been widely used. In deep learning-based image retrieval, image features are extracted using CNN, feature vectors are established, and corresponding images are represented based on the feature vectors of the images.

The similarity between images is determined by calculating the distance between image feature vectors. In this paper, Euclidean distance and cosine distance will be used as the calculation method for the similar picture feature vector. The Euclidean distance equation was used to measure the absolute distance between points in a multidimensional space, and the comparison is shown as the following formula:

$$\mathbf{d}(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}. \tag{3}$$

The cosine distance uses the cosine of the angles between two vectors in the vector space as a measure of the difference between the two individuals. It focuses on the difference in direction between two vectors. The cosine distance equation is shown in the following formula:

$$\mathrm{sim}(\mathbf{x}, \mathbf{y}) = \cos\theta = \frac{\vec{x} \bullet \vec{y}}{\|x\| \bullet \|y\|}. \tag{4}$$

Here, assume that there are a total of N pictures in the picture library, as the following equation shows:

$$\mathbf{D}_i = \|q - T_i\|, i \in [1, N], \tag{5}$$

where q represents the feature vector of the image to be searched, T represents the feature vector of the ith footprint in the picture library, and D represents the distance difference between the feature vectors. The smaller value of D indicates there are two pictures with higher similarity. Then, the D value is compared by the sorting algorithm to find and output the most similar N pictures.

In this paper, three deep neural network models are used to extract features of each image in the image database and establish a feature vector index database. The image to be searched is used to extract feature vectors through the model, and the similarity is calculated by comparing the above distance algorithms.

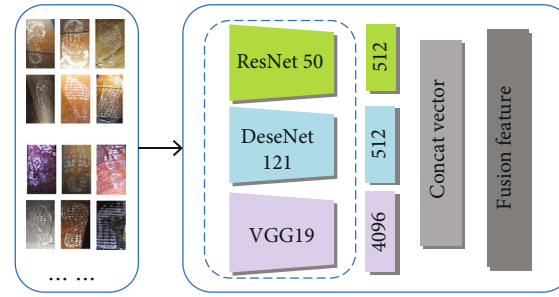

Figure 5: Feature fusion diagram.

4. Experiments

During the experiments, firstly, we clean the data and select these footprint images with high quality for our experiments. Due to the small number of each class of footprint image, we do some data augmentation and use the transfer learning method to get better models. We use pretrained models on the ImageNet dataset. Based on its parameters, we use the footprint dataset to train the model further; this can compensate for the problem of the accuracy of the model caused by the lack of data to a certain extent. In our experiments, the final training classification model can achieve an accuracy of more than 97%.

On the model structure, we define the new fully connected layer and classification layer of VGG19, ResNet50, and DenseNet121. To get better feature representations of footprint images, we first use the training set to train the model. Because we use the images in the same domain to train the model, it will be better when we use the model to extract features. After that, feature representations are extracted from the new fully connected layers of the trained model.

Based on the above experiments, we experimented with feature fusion and connected the feature vectors obtained from the three models to form a new feature vector for the last retrieval. The vector dimension is 5120 dimensions. The experimental results are shown in Table 1.

We extract features of footprint images both in the gallery and query to build two feature vector databases. We can compute the distance between the footprint images and output the ten most similar footprint images.

In this section, we use three different modified models and two different distance measurement functions for the experiments. We also conducted experiments on feature fusion and use the same distance measurement method as the above experiment.

4.1. Dataset and Experimental Setting. Our experiments run on Ubuntu 16.04 with NVIDIA RTX2080ti GPU, 11 G RAM. All the experiments are based on PyTorch (https://github.com/pytorch/pytorch). Experimental code is modified by a project of person reidentification (https://github.com/layumi/Person_reID_baseline_pytorch).

Table 1: Footprint image dataset.

	Gallery	Query	Train	Train all	Validation
Classes	51	51	40	40	40
Average number in each class	48	8	64	65	1
Total number of images	2490	432	2568	2608	40

The footprint dataset consists of two parts. One part was the actual situation of the crime scene footprint image provided by PSBKC. For these data, we first manually select the higher-quality footprints for easy research through data cleaning, and these data are labeled by crowdsourcing. The other part came from our footprint acquisition system. Based on edge computing technology, we developed the footprint acquisition system, which integrates ID card recognition and footprint shooting. With the help of this system deployed at the entrance and exit of our laboratory, some footprints and corresponding labels can be obtained automatically. Due to privacy and other reasons, we can only recruit a small number of volunteers. Thus, only a small number of footprint datasets are obtained.

For these footprint images, we also do some data augmentation, such as rotation, random cropping, gray value, and contrast transformation. The whole dataset has more than 5000 images. There are 40 classes in the train set and 51 classes in the gallery set. We randomly choose some images from each class of the gallery set to form a query set, and each class in the gallery contains more than six similar images with the query. We select one footprint image from each class and train all to form a validation dataset. The basic situation of the dataset is shown in Table 1. Of course, there are still some shortcomings in the process of constructing the footprint image dataset, such as the problem that the amount of data in a single sample is small and the total sample size is not large enough. Therefore, to compensate for the lack of data, we used transfer learning during the training of the model. At the same time, according to the provided footprint images, we use three different models to extract footprint image features in this paper and do some comparative experiments. Besides, we also performed experiments on feature fusion.

4.2. Performance Metrics. We evaluated the performance of three deep learning models on the footprint dataset based on five popular evaluation criteria: the recall rate, precision rate, F1-score, rank-n, and mean average precision (mAP). The precision rate indicates the ratio of the number of similar images in the search result (N_i) to the total number of images in the search results (N_t), as shown in the following equation:

$$\text{Precision} = \frac{N_i}{N_t}. \tag{6}$$

The recall rate indicates the ratio of the number of similar images (N_i) in the search results to the total number of similar images in the database (N_s), as shown in the following equation:

$$\text{Recall} = \frac{N_i}{N_s}. \tag{7}$$

Usually, the accuracy rate and recall rate are relatively contradictory, and it is difficult to achieve the optimal situation simultaneously. The recall rate can reflect the comprehensiveness of the search results, and the precision rate emphasizes the accuracy of the expression. F1-score combines the characteristics of recall rate and precision rate, which is defined as the following equation:

$$\text{F1-score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \tag{8}$$

Rank-n indicates the probability of correct result in the top n search result.

We also use mean average precision to evaluate comprehensive performance. The average precision (AP) is computed for each query. Mean average precision (mAP) indicates the correct average rate of the retrieval system and the equation as follows:

$$\text{mAP} = \frac{\sum_{i=1}^{N} \text{AP}_i}{N}. \tag{9}$$

5. Results and Discussion

5.1. Results. According to the above experiments and combined with five different evaluation methods, we can get experiment results as shown in Table 2.

In the experimental results of using a vector extracted by a separate model as a retrieval feature, it can be found that using ResNet50 + cosine can get the highest score of 83.43% on recall, while VGG19 + Euclidean get low scores on the precision. ResNet50 + cosine achieved the highest 82.78% and 83.10% score both in precision and F1-score. All the models except VGG19 + Euclidean can get nearly 100% rank10 score. It means that these models in the experiments can output similar footprint images to the input. ResNet50 + Euclidean achieve 87.02% on mAP.

In the feature fusion experiment, we can find that when the Euclidean distance is used as the measurement function, the experimental results are much higher than the VGG model under all the evaluation indicators, which is similar to ResNet50 and DenseNet121 alone but slightly lower than these two models. However, when combining cosine distances, the fusion feature is the highest in all indicators except

Table 2: Results in our experiments.

	Recall	Precision	F1-score	Rank1	Rank10	mAP
VGG19 + Euclidean	27.14%	26.69%	26.91%	30.79%	88.19%	49.50%
VGG19 + cosine	45.93%	45.46%	45.69%	96.30%	98.84%	50.49%
ResNet50 + Euclidean	78.17%	77.52%	77.85%	98.15%	100%	87.02%
ResNet50 + cosine	83.43%	82.78%	83.10%	99.77%	100%	86.46%
DenseNet121 + Euclidean	78.23%	77.59%	77.91%	95.60%	100%	84.13%
DenseNet121 + cosine	81.88%	81.23%	81.55%	100%	100%	84.12%
Fusion feature + Euclidean	76.62%	75.97%	76.29%	96.76%	100%	78.72%
Fusion feature + cosine	83.63%	82.99%	83.31%	100%	100%	85.60%

mAP and in the various indicators has a very stable performance. It also shows that the fusion feature method has practical effects. From the overall results of the experiment, DenseNet151 and ResNet50 are better than VGG19, and ResNet50 is slightly better than DensNet121. The fused features have excellent performance when paired with the cosine distance, but slightly lower than ResNet50 and DenseNet151 when paired with Euclidean distance. In terms of measuring distance algorithm, cosine distance is better than Euclidean in our experiment; fusion feature + cosine has the best performance.

5.2. Analysis and Discussion. In this paper, we explore three different deep convolutional neural networks for footprint image retrieval. Three typical depth models were modified and applied to this paper. All of them use Euclidean distance and cosine distance, respectively, for the retrieval experiments. At the same time, we propose a fusion method for the feature representation of the footprint image in this paper. We combine the feature outputs of the three models of VGG19, ResNet50, and DenseNet121 to form a more abundant footprint image feature. And experiments are also carried out with two measurement functions of Euclidean distance and cosine distance.

After experiments, we get good retrieval results. It can initially meet the actual needs of public security. To some extent, it can reduce the time required for manual screening and play a supporting role in the investigation of public security cases. We can see some retrieval results from Figure 6, where the green number means similar retrieval images, and the red number is the wrong output.

From the above experimental results, we found that VGG19 is the worst performing model in all models. It may be due to its shallow network structure, and the extracted footprint image features are not rich enough compared to the other two models.

ResNet has introduced a residual network structure, through which the residual network can make the network layer deeper and relatively improve its performance. Experiment results in our paper also show that ResNet50 achieves higher scores than DenseNet121. Though DenseNet121 has more deep layers to get richer footprint image features, it still has little distance with ResNet50 in the performance of footprint image retrieval. That is, the deeper network may not bring better results. As far as the separate model is concerned, ResNet50 is more suitable for the research content of this paper.

The fusion features we proposed are very useful in the experiment. When using the cosine distance as the distance measurement function, it has very stable and excellent performance. We can deduce that the reason why it can obtain such a good effect is that the feature vectors extracted by different models are different. The fusion features fully combine the footprint image information of three different dimensions, maximally retaining the image information obtained by each model.

However, these differences are not useless, which determine the characterization of the final footprint image. The fusion feature makes the final obtained footprint image feature information more abundant, which is more conducive to retrieval experiments. And the final experimental results prove that the method is significant and stable.

In this paper, we choose two common distance metric functions, cosine distance and Euclidean distance, for the similarity calculation between footprint images. These two distance metric functions are simple, but they work well in various tasks. As can be seen from our experiments, the use of cosine distance as a measure function is much better than the Euclidean distance. We believe that Euclidean distance is more focused on the numerical difference between vectors in low-dimensional space. In contrast, cosine distance is more concentrated on the difference between vectors in high-dimensional space. In our experiments, the footprint image has a higher feature dimension after passing through the depth model and is more suitable for processing using the cosine distance. It also indicates that the cosine distance is suitable for the recognition and retrieval of footprint images.

From the above experimental results, we can find that rank1 is very high; it means if the output is set to 10 images, the first one of the output results has a probability of more than 90% being similar images. The reason for this may be the lack of diversity in our data for each class of samples. In other words, we do not have enough data in each type of footprint image. Although we have done some data augmentation, just merely inverting the image and changing the gray value of footprint images cannot reach data diversity. In future research, we may focus on this problem and expand our footprint dataset via our footprint acquisition system.

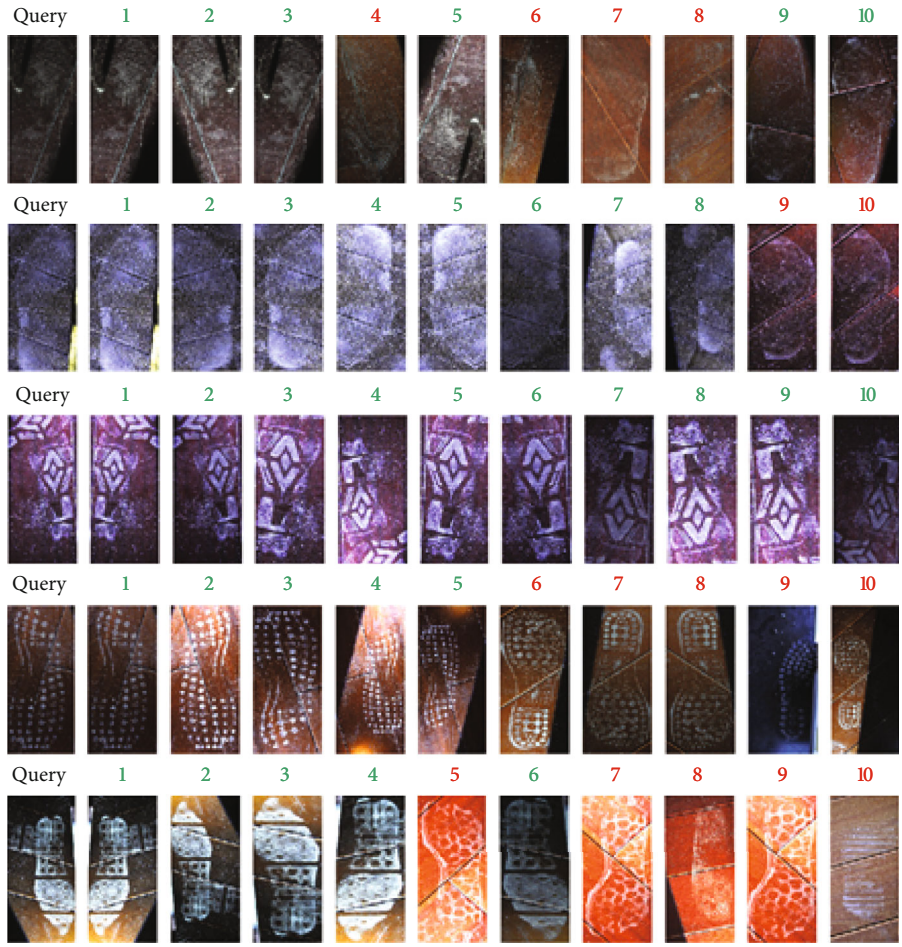

Figure 6: Some retrieval results.

6. Conclusions

This paper has explored some deep learning models used for footprint image retrieval and matching methods. We use the three different deep neural network models to extract features and two metric learning methods to calculate similarity. At the same time, based on the above research content, we propose a new feature representation method for the recognition and retrieval of footprint images. These experiments show that the method can efficiently be used for similar search footprints.

As far as we know, there is no standard and complete footprint image matching system at present. Our work can reduce the mass image matching to only requiring retrieval from a limited number of candidate images, thereby providing feasibility for further human judgment. By the experimental results, our approach can effectively help public security police in the detection of cases while significantly reducing the resources consumed in the manual retrieval process and has practical application value.

Although we have completed preliminary experiments, there are still some problems, such as the dataset is not large enough, which may affect the pretraining performance of the deep learning model based on CNN. Combining with the actual combat scene in the future, we will focus on the construction of sample datasets. At the same time, image background processing and different metric learning methods are still challenging tasks, which will also be the focus of our next step.

Data Availability

All data is available from the authors.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Authors' Contributions

Dechao Chen and Yang Chen contributed equally to this study and share first authorship.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (grant numbers 61876217, 61673290, 61672371, 61876121); the Innovative Team of Jiangsu Province under Grant (grant number XYDXX-086); the Science

& Technology Development Project of Suzhou (grant number SYG201817); and the Student Research Development Project of Suzhou University of Science and Technology (grant number SKSJ18_010).

References

[1] J. Sivic and A. Zisserman, "Video Google: a text retrieval approach to object matching in videos," in *ICCV*, p. 1470, IEEE, 2003.

[2] Y. Deng and B. S. Manjunath, "Unsupervised segmentation of color-texture regions in images and video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 8, pp. 800–810, 2001.

[3] R. Arandjelović and A. Zisserman, "Three things everyone should know to improve object retrieval," in *2012 IEEE conference on computer vision and pattern recognition*, pp. 2911–2918, Providence, RI, USA, 2012.

[4] M. M. Rahman, S. K. Antani, and G. R. Thoma, "A learning-based similarity fusion and filtering approach for biomedical image retrieval using SVM classification and relevance feedback," *IEEE Transactions on Information Technology in Biomedicine*, vol. 15, no. 4, pp. 640–646, 2011.

[5] J. Wan, D. Wang, S. C. H. Hoi et al., "Deep learning for content-based image retrieval: a comprehensive study," *Proceedings of the 22nd ACM international conference on multimedia*, 2014, pp. 157–166, ACM, 2014.

[6] Y. Liu, D. Zhang, G. Lu, and W. Y. Ma, "A survey of content-based image retrieval with high-level semantics," *Pattern Recognition*, vol. 40, no. 1, pp. 262–282, 2007.

[7] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning CNN image retrieval with no human annotation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 7, pp. 1655–1668, 2018.

[8] A. Gordo, J. Almazan, J. Revaud, and D. Larlus, "End-to-end learning of deep visual representations for image retrieval," *International Journal of Computer Vision*, vol. 124, no. 2, pp. 237–254, 2017.

[9] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han, "Large-scale image retrieval with attentive deep local features," in *Proceedings of the IEEE international conference on computer vision*, pp. 3456–3465, Venice, Italy, 2017.

[10] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "Imagenet: a large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Miami, FL, USA, 2009.

[11] V. Iglovikov, S. Mushinskiy, and V. Osin, "Satellite imagery feature detection using deep convolutional neural network: a Kaggle competition," 2017, https://arxiv.org/abs/1706.06169.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.

[13] C. Szegedy, W. Liu, Y. Jia et al., "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, Boston, USA, 2015.

[14] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: a unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, Boston, USA, 2015.

[15] T. N. Sainath, A. R. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," in *2013 IEEE international conference on acoustics, speech and signal processing*, pp. 8614–8618, Vancouver, BC, Canada, 2013.

[16] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, "Visualizing higher-layer features of a deep network," *University of Montreal*, vol. 1341, no. 3, p. 1, 2009.

[17] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*, pp. 818–833, Cham, 2014.

[18] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1717–1724, Columbus, USA, 2014.

[19] G. Tolias, R. Sicre, and H. Jégou, "Particular object retrieval with integral max-pooling of CNN activations," 2015, https://arxiv.org/abs/1511.05879.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, https://arxiv.org/abs/1409.1556.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, Las Vegas, USA, 2016.

[22] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, Hawaii, USA, 2017.

[23] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned cnn embedding for person reidentification," *ACM transactions on multimedia computing, communications, and applications (TOMM)*, vol. 14, no. 1, p. 13, 2018.

[24] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: deep filter pairing neural network for person reidentification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 152–159, Columbus, USA, 2014.

[25] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person reidentification," in *2014 22nd international conference on pattern recognition*, pp. 34–39, Stockholm, Sweden, 2014.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, Santiago, Chile, 2015.

[27] P. Goyal, P. Dollár, R. Girshick et al., "Accurate, large minibatch SGD: training imagenet in 1 hour," 2017, https://arxiv.org/abs/1706.02677.

[28] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *Journal of Machine Learning Research*, vol. 10, no. Feb, pp. 207–244, 2009.