Image Retrieval using CNN and Low-level Feature Fusion for Crime Scene Investigation Image Database

Ying Liu[123], Yanan Peng[1 *], Dan Hu[1], Daxiang Li[123], Keng-Pang Lim[13], Nam Ling[34]
1 Center for Image and Information Processing, Xi'an University of Posts and Telecommunications, Xi'an 710121, China
2 Key Laboratory of Electronic Information Application Technology for Scene Investigation, Ministry of Public Security, Xi'an, 710121, China
3 International Joint Research Center for Wireless Communication and Information Processing, Shaanxi, Xi'an, 710121, China
4 Department of Computer Engineering, Santa Clara University, California, 95053, USA

Abstract — **Crime scene investigation (CSI) image retrieval is used to search for crime evidences and is critical in helping in solving various crimes. In recent years, using Convolutional Neural Network (CNN) has demonstrated outstanding performances in large-scale image database retrieval. However, to prevent over-fitting in the training of CNN model due to limited number of CSI images, this paper proposes to cascade two CNN models obtained based on transfer learning and combine CNN features with low-level image feature to better describe CSI images. First, two pre-trained CNN models are fine-tuned using the target image set. CNN features are extracted from fully connected layer of each model and are concatenated as high-level features for the image. These concatenated CNN features are then fused with the low-level image features of the target image set. The final fused image features are used in the image retrieval. Experimental results on CSI image database proved the effectiveness of the proposed algorithm for limited number of training sets. In addition, experiments carried out on the GHIM-10K database proved the generalizability of the proposed algorithm.**

I. INTRODUCTION

Crime scene investigation (CSI) image is an important part of the information collected at crime scenes. Classification and retrieval of CSI images provide important clues and play an important role in solving serial crimes [1]. Therefore, there is an urgent need for an automatic and effective image classification and retrieval system to quickly find relevant images from a large number of CSI images to improve the efficiency of the investigation while saving human power and material resources.

Currently, there are few studies on CSI image retrieval. Existing CSI image retrieval technologies can be divided into two categories: CSI image retrieval based on low-level features and that based on high-level semantics. CSI image retrieval technology based on low-level features uses a content-based image retrieval (CBIR) framework to extract low-level features of the image (such as color histogram, gray level co-occurrence matrix, Gabor features, wavelet texture features, etc.) or to fuse different low-level features, which confirms the feasibility of CBIR technology in CSI image retrieval [2, 3]. In [4], the author proposes to combine low level features of image dominant color descriptors as color features, gray-level co-occurrence matrix as texture features and the edge feature obtained by gradient vector flow to improve CSI image retrieval performance. The disadvantage is that the computation is complex and slow. In [5], an image retrieval method based on regional semantic template is proposed. First, the user submits the query image and the region of interest, thereby constructing a regional semantic template and performing pre-classification. Finally, the image is sorted. Experiments show that the algorithm is effective to improve the accuracy of CSI image retrieval. Ref. [6] proposes a two-layer system for CSI image retrieval frameworks. First, the corresponding feature database of the CSI image database is computed, and a support vector machine (SVM) classifier model that can achieve multi-semantic classification is pre-trained. After the investigator submits the retrieved images, SVM automatically determines the semantic categories based on the image features, and then it performs matching retrieval on the image library containing only the semantics. Experimental results show that this method outperforms the Query By Example (QBE) method in multiple retrieval indexes, with significant reduction in retrieval time, by half. It is also an effective method to introduce relevant feedback (RF) into the CSI image retrieval. In [7], RF is used to automatically adjust the weights of shoe print features to improve precision. Although the above method achieved some good results, they lack the "semantic gap" which may improve the accuracy of image retrieval significantly.

With the pioneering work by Hinton et al. [8] in 2006, deep learning has developed rapidly in the recent decade. There are several types of deep learning frameworks such as convolutional neural networks (CNN) and deep belief networks (DBN), applied to digit recognition [9], image classification [10], face recognition [11], and other applications with unprecedented success. Deep learning has a wide range of applications in image classification and retrieval as well. For example, in the ImageNet competition, the accuracy of using traditional classifiers in 2010 (top 5 accuracy) was 71.8%, and in 2011 it was 74.3%. In 2012, Hinton and his student Alex et al. used deep learning to improve the accuracy rate to 84.7%. In the 2017 competition, the final accuracy rate was as high as 97.3%. A. Babenko and J. Donahue [12,13] extracted features of the CNN fully connected layer as high-level semantic features for retrieval, and also extracted image features from the convolutional layer for retrieval [14], and achieved good results. Juan A. Carvajal

[15] obtained good retrieval results on butterfly species images by fine-tuning the three pre-trained models and extracting the previous layer of the output layer as features. The excellent performance of CNN in image classification and retrieval depends on a large amount of training data. However, due to the unique and sensitivity of CSI images, there is no standard large-scale CSI image database in the academic community. Our Center has cooperated with the public security department for many years to sort out the CSI image database for academic use. It contains 19,363 images in various categories [1]. In order to apply the advantages of CNN in general image retrieval on CSI images, this paper combines the features extracted from CNN features with low-level image features. This algorithm makes full use of CNN's advantages in deep semantic feature extraction, and also considers the low-level image features of targeted image database resulting in improved effective CSI image retrieval and efficiency. The structure of this paper is as follows: Section II introduces the algorithm proposed in this paper, Section III presents the experimental results and analysis, and finally Section IV summarizes the work.

II. CSI IMAGE RETRIEVAL ALGORITHM BAESD ON FUSION OF CNN AND LOW-LEVEL FEATURES

In order to overcome the semantic gap in retrieval using low-level image features, this paper proposes the integration of CNN features with low-level image features. In addition, instead of relying on a single CNN model, the features extracted from the two CNN models are merged and then fused with the low-level features of targeted image database to represent the image more reliably. The algorithm is shown in Fig. 1.

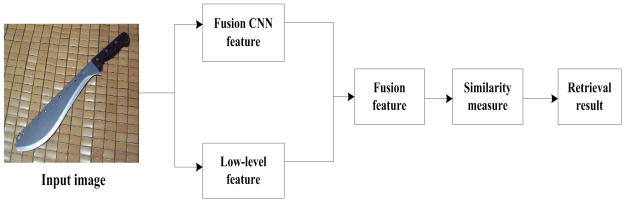

Fig. 1 CSI image retrieval algorithm based on fusion of CNN and low-level features.

A. CNN Feature Extraction

CNN feature extraction is mainly divided into three steps: transfer of pre-trained CNN model, fine-tuned CNN model, and CNN feature extraction and fusion.

(1) Transfer of Pre-trained CNN Models

CNN features have achieved great successes in image classification and retrieval. It performs so well by using a large amount of training data. However, in some special applications such as CSI image retrieval, because of insufficient data, the training of convolutional neural networks could easily overfit, resulting in unreliable results. For this reason, we extract the features from CNN models trained from the ImageNet dataset.

There are already many well-known pre-trained models such as AlexNet [16], VGG [17], and VGG-VD [18], which have been successfully applied in scene classification [19-20] and image retrieval [21-24]. The CNN models included in the VGG network are: VGG-F, VGG-M, and VGG-S. They have similar structures, and they differ only in the number and size of filters in the convolutional layer. In this experiment, VGG-F was selected as a pre-trained model. There are altogether 8 layers consisting of 5 convolutional layers, and 3 fully connected layers. The structure is shown in Table 1:

Table 1 The structure of VGG-F model.

Conv1	**Conv2**	**Conv3**	**Conv4**
64*11*11 stride 4	256*5*5 stride 1	256*3*3 stride 1	256*3*3 stride 1
Conv5	**Fc6**	**Fc7**	**Fc8**
256*3*3 stride 1	4096 dropout	4096 dropout	1000 softmax

VGG-VD (VGG Very Deep Convolutional Network) is a deep CNN network, including VD16 (16 layers in total, consisting of 13 convolutional layers and 3 fully connected layers) and VD19 (19 layers in total, consisting 16 convolutional layers and 3 fully connected layers). This experiment selected VGG-VD16 as a pre-trained model, and its structure is shown in Fig. 2.

One improvement of VGG compared with AlexNet is its simple structure. Several successive convolution kernels of 3x3 are used to replace the larger convolutional kernels in AlexNet (11x11,7x7,5x5). For a given receptive field, the small convolution kernel with stacking is better than the large convolution kernel, because multilayer nonlinear layers can increase the network depth to ensure that learning more complex patterns, and the cost is relatively small (less parameters).

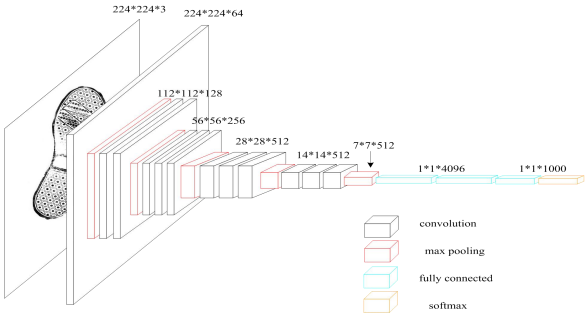

Fig. 2 The structure of VGG-VD 16.

(2) Fine-Tuning the CNN Model

Regardless of the number of classes in images, usually there is a big difference between the target image set and the pre-trained image set. Therefore the target image retrieval performance is often suboptimal by using the pre-trained CNN model [25] features directly. Hence, the images of the

target image set are often used to fine-tune the parameters of the pre-trained CNN model. The entire fine-tuning process is as follows:

Step 1: Each image in the image database is adjusted to 224*224 as the input to the CNN;

Step 2: For layers 1 to 7, i.e. conv1-fc7, initialize the parameters of the pre-trained model, and then modify the output category of the last layer of the fully connected layer of the pre-trained CNN model from 1000 to the targeted number of classes and retrain.

(3) CNN Feature Extraction and Fusion

The model has a total of 3 fully connected layers, where the last fully connected layer is used for classification. So the first two fully connected layers are used as feature extractors. We extract the output of the second fully connected layer, i.e. fc7 layer from the two CNN models as features of the image that are 4096 in dimension. If the fusion is performed directly, the fusion feature will have a very high dimension. High-dimensional features will lead to high computational complexity and high memory occupancy rate, which hinders real-time or near real-time retrieval. Since there is information redundancy between these features, the features extracted from the two models are reduced to 128 in dimension by using Principal Components Analysis (PCA) before fusion. The fusion feature F can be expressed as equation (1):

$$F = [w_1 f_1, w_2 f_2] \tag{1}$$

where f_1 represents feature extracted from the fc7 layer of VGG-VD 16 model, f_2 represents feature extracted from the fc7 layer of VGG-F model, $w_f = [w_1, w_2]$ is the feature weight, and the empirical values are derived from the experimental results in Section III. Cascading the two dimension-reduced features not only ensures the validity of features, but also greatly reduces the amount of computations and memory reduction. It can achieve real-time retrieval and improve user experience while ensuring the good performance.

B. Low-Level Feature Extraction

Referring to our studies on low-level features of CSI images in [1] and taking into account the common characteristics of natural images, this paper uses the following four low-level features:

(1) HSV Color Histogram

Color histogram is the basic means to describe color information of an image. It reflects the proportion of different colors in the entire image, i.e. the frequency of each color. The equation is as follows (2):

$$H(k) = \frac{n(k)}{N} \tag{2}$$

where, k = 0,1,.....,L-1, represents the feature value of the image, L is the number of possible feature values, and $n(k)$ is the number of pixels having the feature value k in the image. N is the total number of image pixels.

Since the HSV color space is more in line with human eye perception, this experiment converts the image to the HSV color space, and performs quantification of 20, 10, and 5 on the H, S, and V channels respectively, to form a 1000-dimension color histogram. [26].

(2) Gabor Features

The Gabor wavelet is very similar to the visual stimuli response of simple cells in the human visual system. It has good characteristics in extracting the local space and frequency domain information of the target image. Gabor filtering can be used to extract the texture information of the image [27]. Gabor feature extraction steps are as follows:

(1) Set up Gabor filter group: 4 scales, 6 directions, consisting a total of 24 Gabor filters.

(2) The Gabor filter bank is convolved with each image in the spatial domain, and each image can be obtained with 24 filter outputs.

(3) Extract the mean and variance of each filter output as the texture features of the image with a total of 48 dimensions.

(3) DCT Wavelet

Discrete Cosine Transform (DCT) [28-31] is a classical image frequency domain information analysis tool and is often used for image texture feature extraction. Wavelet is a mathematical function that decomposes data or signals into different time-frequency components through time-frequency analysis. Discontinuous and sharp signals can be analyzed by wavelet transform [32]. Kekre's Wavelet (abbreviated as Kekre wave) generated by the Kekre transformation matrix is applied to the DCT coefficient matrix, and DCT wavelet (abbreviated as DCT wave) can be obtained [32]. The Kekre transformation matrix can be any size as follows:

$$K_{N\times N} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 & 1 \\ -N+1 & 1 & 1 & \cdots & 1 & 1 \\ 0 & -N+2 & 1 & \cdots & 1 & 1 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 1 \\ 0 & 0 & 0 & \cdots & -N+(N-1) & 1 \end{bmatrix} \tag{3}$$

The $16^2 \times 16^2$ Kekre wave matrix is constructed by a 16×16 Kekre transformation matrix in paper. After the Kekre wave matrix is obtained, the Kekre wave matrix is computed with the orthogonally DCT transform coefficients to obtain the DCT wave, as shown in equation (4):

$$[F]=[kw]d[kw]^T \quad (4)$$

where, kw is Kekre wave matrix, d is the DCT transform coefficient and F is the final DCT wave.

In this paper, the CSI image is divided into four blocks, and the three channels R, G, and B of each block are DCT transformed; the DCT transform matrix and the Kekre wave matrix are calculated to obtain the DCT wave coefficients; the mean value and variance of the DCT wave coefficients are calculated which constitute the feature vector of each small block; finally, the feature vectors of all the small blocks are connected to form the texture features of the entire CSI image, with a total of 24 in dimension [1].

(4) GIST Features

The GIST model [33] was originally proposed by Olive. The GIST feature is a bio-inspired feature that simulates human visual extraction of rough but concise contextual information in an image. The GIST descriptor uses a series of statistical attributes to quantify the image scene, such as naturalness, openness, roughness, degree of expansion, and ruggedness. The extraction procedure for GIST features is as follows:

(1) Divide an image of size $h\times w$ into $n_a\times n_b$ grid regions, each of which has a size of $h'\times w'$, where $h'=h/n_a$, $w'=w/n_b$.

(2) Each area is convolved with a Gabor filter with *m*-scale *n*-direction to obtain the Gabor response results of $n_c=mn$ channels. Then, the Gabor features are obtained by connecting the result levels of each channel. The equation is shown as follows:

$$G_i(x,y)=\underset{n_c}{cat}(f(x,y)*g_{mn}(x,y)) \quad (5)$$

where i = 1,2,..., $n_a\times n_b$, $g_{mn}(x,y)$ denotes a Gabor filter with *m* dimensions in *n* directions and *cat* denotes cascading the response results for each channel.

(3) Calculate the mean values of Gabor features in each region and cascade them to obtain GIST features.

C. Fusion Feature

In order to obtain more distinguishing features and a more comprehensive representation of the image, we fuse the above features to obtain the fusion feature *F*, which is expressed by equation (6):

$$F=\{w_1f_{CNN1},w_2f_{CNN2},w_3f_{GIST},w_4f_{HSV},\ w_5f_{Gabor},w_6f_{DCT-W}\} \quad (6)$$

where f_{CNN1} denotes the CNN feature extracted from the model VGG-F, f_{CNN2} denotes the CNN feature extracted from the model VGG-VD 16, f_{GIST} denotes the GIST feature, f_{HSV} denotes the HSV color histogram, f_{Gabor} denotes the Gabor feature, and f_{DCT-W} denotes the DCT wave feature. $w_f=[w_1,w_2,w_3,w_4,w_5,w_6]$ represents the feature weight vector, and the specific weight will be displayed in the experimental results section.

III. EXPERIMENTAL RESULTS AND ANALYSIS

A. Experimental Database

The experiment used two databases to test the proposed algorithm and used MatConvNet [34] to implement our algorithm. MatConvNet is a MATLAB toolbox that provides CNN function, which is simple and efficient. Many pre-trained CNN networks can be used.

The first one is Crime Scene Investigation Image Database (CSID), which was acquired from the Center for Image and Information Processing (CIIP) of Xi'an University of Post and Telecommunications. The experiment selected 12 semantic categories of images, including: biological evidence, blood stains, cars, doors, fingerprints, site plans, shoe prints, skin, tattoos, tools, tires, and windows, a total of 10,500 images. Some examples are shown in Fig. 3.

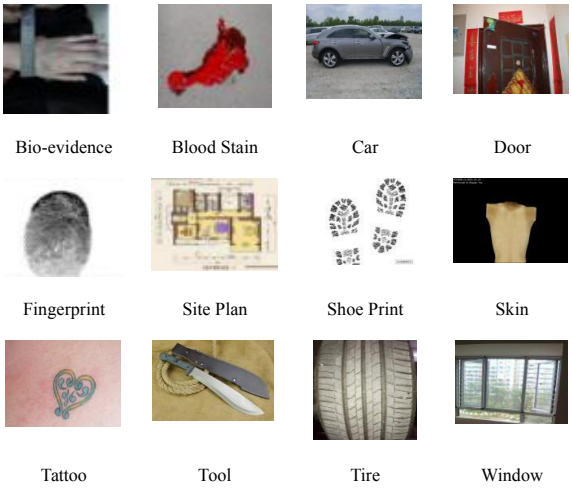

Fig. 3 Samples in CSID.

The second database is the GHIM-10K database [35], which contains a total of 20 types of images, such as firework, building, car, flower, butterfly, etc. Each category contains 500 pictures, a total of 10,000 pictures. Some example images are shown in Fig. 4.

Fig. 4 Samples in GHIM-10K Database.

B. Evaluation Criteria

This paper uses precision to measure retrieval performance. The larger the value, the more the number of relevant images returned, reflecting more accurate retrieval. The equation is shown as follows:

$$P = \frac{S}{K} \tag{7}$$

where S is the number of correct related images contained in the result returned in one query, and K is the total number of images returned in a single query.

C. Experimental Results

(1) Selection of Low-Level Features

The low-level features in Section II.B and the precision of their combinations were tested on the CSID and GHIM-10K databases, respectively. The results are shown in Table 2 and Table 3, respectively.

Table 2 Average precision of various low-level features and combinations on CSID (K=10).

Single feature	Average precision	Fusion feature	Average precision
HSV	55.40%	HSV+GIST+DCT wave	63.97%
Gabor	62.32%	GIST+DCT wave	66.62%
DCT wave	45.38%	HSV+Gabor+GIST	67.37%
GIST	65.18%	Gabor+GIST	70.07%

The experimental results show that the combination of Gabor and GIST has the best effect as the low-level features of the CSI images. For CSI images, many categories do not have unique color information, such as biological evidence, shoe prints, site plans, etc. Therefore, the introduction of HSV color histograms in the experiment reduced the precision rate.

Table 3 Average precision of various low-level features and their combinations on the GHIM-10K database (K=10).

Single feature	Average precision	Fusion feature	Average precision
HSV	43.06%	DCT wave+GIST	57.20%
Gabor	44.70%	HSV+DCT wave+GIST	63.48%
DCT wave	29.88%	Gabor+GIST	58.08%
GIST	52.98%	HSV+Gabor+GIST	65.83%

The color features are effective for describing natural images. The experimental results show that the combination of HSV color histogram, Gabor features, and GIST features can better represent natural images more reliably.

(2) Average Precision of CNN Features and Fusion Features

In order to further improve the performance of image features, we cascaded the features of the fc7 layers of the two CNN networks, and then fused with the low-level features that are suitable for representing the images. Similarly, the test is conducted on the CSI image database and the GHIM 10K database. As shown in the following experiments (10 images returned for each search), the best retrieval results for the CSI image database uses the combination of CNN1, CNN2 (CNN1 and CNN2 represent features extracted from both models VGG-F and VGG-VD 16 respectively), GIST, Gabor features with a weighting ratio of 0.2:0.6:0.1:0.1. For GHIM 10K database, the best result is the combination of the features of CNN1, CNN2, HSV, Gabor, and GIST with the weighting ratio of 0.1:0.7:0.1:0.05:0.05. The full results are shown in Tables 4 and 5.

Table 4 Average precision of CNN features and fusion features on CSID (K=10).

Feature	Average precision	Average Retrieval Time/per image (s)
GIST+Gabor	70.07%	0.044
CNN1	87.83%	0.024
CNN2	91.94%	0.028
CNN1+CNN2	92.33%	0.030
CNN1+CNN2+GIST + Gabor	93.21%	0.065

Table 5 Average precision of CNN features and fusion features on the GHIM-10K database (K=10).

Feature	Average precision	Average Retrieval Time/per image (s)
HSV+Gabor+GIST	65.83%	0.120
CNN1	94.40%	0.020
CNN2	98.40%	0.021
CNN1+CNN2+HSV +Gabor+GIST	98.49%	0.191

The above experimental results of Table 4 show that the fusion of two CNN features and low-level features can effectively improve the accuracy of CSI image retrieval. And the experimental results of Table 5 show that the proposed method is suitable not only for CSI images but also for general images (although the retrieval precision is only slightly improved). After combining low-level features, there is little increase of the average retrieval time of per image. It can still maintain real-time retrieval.

(3) Average Precision of Images in Each Category

The above results are the average precision of each database. To understand the average precision of images in each category, we presented the average precision of each category image in the two databases. The results are shown in Fig. 5 and Fig. 6:

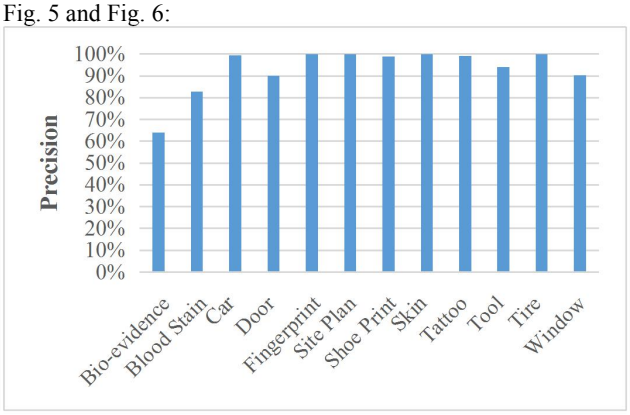

Fig. 5. Average precision of various categories of images on CSID (K=10).

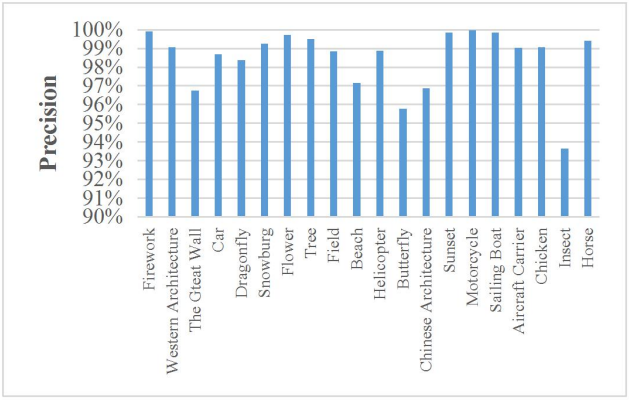

Fig. 6. Average precision of various categories of images on the GHIM-10K database (K=10).

From the above results, it can be seen that the accuracy rate of each category image in the same database is different. For example, in the CSID database, the precision rate of biological evidence is lower than that of the other categories. In the GHIM-10K database, the average precision of insects is low. The reason is that the target image is not obvious and is greatly affected by the background. In our future work, we will extract more effective features for image classes with low precision.

IV. CONCLUSION

In order to alleviate overfitting problems arising from the limited CSI images during CNN network training, this paper proposes to use the CNN model trained on large-scale ImageNet and use the CSI image database for fine-tuning. Firstly, pre-trained models VGG-F and VGG-VD 16 are fine-tuned by using CSI images and the fc7 layer features are extracted. CNN features dimensions are reduced and are merged before fusing with the low-level image features suitable for CSI images. Final experimental results show that the algorithm can effectively describe the content of CSI image and maintain a high average precision. In addition, this algorithm is applicable for other types of image databases as well. In order to further increase average precision rate, future work will incorporate a more effective scheme for individual categories with low precision.

ACKNOWLEDGMENT

This work was supported by Science and Technology Project Fund (No. 2016GABJC51) under the Ministry of Public Security of China.

REFERENCES

[1] Y. Liu, D. Hu, J. L. Fan, F. P. Wang. Multi-feature fusion for crime scene investigation image retrieval. The International Conference on Digital Image Computing: Techniques and Applications (DICTA), Sydney, Australia, Nov 30 – Dec 2, 2017.

[2] C. Y. Wen, C. C. Yu . Image Retrieval of Digital Crime Scene images. Forensic Science Journal, vol. 4, pp. 37-45, 2005.

[3] Y. Liu, J. L. Fan, Z. Li. Exploration of crime scene investigation image database retrieval techniques. Journal of Xi'an University of Posts and Telecommunications, vol. 20, 2015.

[4] Gulhane S. A. Content based image retrieval from forensic image databases. International Journal of Engineering Research & Applications, vol. 5, pp. 66-70, 2015.

[5] Y. Liu, Y. Huang, S. Zhang, D. S. Zhang, N. Ling. Integrating object ontology and region semantic template for crime scene investigation image retrieval. International Conference on Industrial Engineering and Applications (ICIEA2017), Nagoya, Japan April, 2017.

[6] W. Liu, Y. Liu. A two-phase hierarchical approach for crime scene investigation image retrieval. Journal of Xi'an University of Posts and Telecommunications, vol. 21, no. 6, 2016.

[7] Y. Yu,. Research and System Implementation of Sole Pattern Retrieval Algorithm. Dalian Maritime University, 2009.

[8] G. E. Hinton, R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. Science, vol. 313, pp. 504-507, 2006.

[9] Y. LeCun, L. Jackel, L. Bottou, C. Cortes, J.S. Denker, H. Drucker. Learning algorithms for classification: A comparison on handwritten digit recognition. Neural networks: The statistical mechanics perspective, pp. 261–276, 1995.

[10] A. Krizhevsky, I. Sutskever, G.E. Hinton. Imagenet classification with deep convolutional neural networks. Proceeding of the 25th International Conference on Neural Information Processing Systems, pp. 1097–1105, 2012,

[11] Y. Taigman, M. Yang, M. Ranzato, L. Wolf. Deep face: closing the gap to human-level performance in face verification. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1701–1708, 2014.

[12] A. Babenko, A. Slesarev, A. Chigorin. Neural Codes for Image Retrieval. European Conference on Computer Vision", pp.

584–599, 2014.

[13] J. Donahue, Y. Jia, O. Vinyals et al. DeCAF: A deep convolutional activation feature for generic visual recognition. The 31st International Conference on Machine Learning, vol. 32, no.1, pp. 647-655, 2014.

[14] G. Tolias, R. Sicre, H. Jégou. Particular object retrieval with integral max-pooling of CNN activations. Proceedings of the 4th International Conference on Learning Representations (ICLR), 2016.

[15] A. Juan Carvajal1, G. Dennis Romero. Fine-tuning based deep convolutional networks for lepidopterous genus recognition. Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications (CIARP), pp.467 – 475, 2016.

[16] A. Krizhevsky, I. Sutskever, G.E. Hinton. Imagenet classification with deep convolutional neural networks. Proceedings of the 25th International Conference on Neural Information Processing Systems, pp. 1097–1105, 2012.

[17] K. Chatfield, K. Simonyan, A.Vedaldi, A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. British Machine Vision Conference, 2014.

[18] K. Simonyan, A. Zisserman. Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations, 2015.

[19] D. Marmanis, M. Datcu, T. Esch, U. Stilla. Deep learning earth observation classification using Imagenet Pretrained Networks. IEEE Geosci. Remote Sens. Lett. vol. 13, pp. 1–5, 2015.

[20] M. Castelluccio, G. Poggi, C. Sansone, L. Verdoliva. Land use classification in remote sensing images by convolutional neural networks. arXiv 2015, arXiv:1508.00092.

[21] V. Chandrasekhar, J. Lin, O. Morère, H. Goh, A. Veillard. A practical guide to CNNs and Fisher Vectors for image instance retrieval. Signal Process. vol. 128, pp. 426–439, 2016.

[22] A.B. Yandex, V. Lempitsky. Aggregating local deep features for image retrieval. Proceedings of the IEEE International Conference on Computer Vision, Santiago, December 2015, pp. 1269–1277.

[23] A. Babenko, A. Slesarev, A. Chigorin. Neural codes for image retrieval. European Conference on Computer Vision, pp. 584–599, 2014.

[24] P. Napoletano. Visual descriptors for content-based retrieval of remote sensing images. International Journal of Remote Sensing, Published online: 24 Nov 2017 (in press).

[25] Z. Li. Image retrieval of CNN visual features. Journal of Beijing University of Posts and Telecommunications, 2015, 38.

[26] Liang Zheng, Shengjin Wang, Qi Tian. Coupled Binary Embedding for Large-Scale Image Retrieval.IEEE Transactions on Image Processing, vol. 23, no. 8, pp. 3368-3380, June, 2014.

[27] Priyadarsan Parida, Nilamani Bhoi. 2-D Gabor filter based transition region extraction and morphological operation for image segmentation. Computers & Electrical Engineering, vol. 62, pp. 119-134, 2017.

[28] B. Vibha, B. P. Sandeep. CBIR using DCT for feature vector generation. International Journal of Application or Innovation in Engineering & Management (IJAIEM), vol. 1, no. 2, pp. 196-200, 2012.

[29] J. Qin, G. M. Luo. A Fast Image Retrieval Technique Based on DCT Domain. Computer System Application, 2005, 14(5):29-31. doi: 10.3969/j.issn.1003-3254.

[30] Reeves A. R., Kubik K., Osberger W. M. Texture characterization of compressed aerial images using DCT coefficients. Proc. SPIE, 1997, 3022 (3022):398-407.doi: 10.1117/12.263428.

[31] Lay J. A., Guan L. Image retrieval based on energy histograms of the low frequency DCT coefficients. IEEE International Conference on Acoustics, Speech, and Signal Processing, Phoenix, AZ, USA 1999: 3009-3012.

[32] Kekre D. H. B., Athawale A., Sadavarti D. Algorithm to generate Kekre's wavelet transform from Kekre's transform. International Journal of Engineering Science & Technology, vol. 2, no. 5, pp. 756-767, 2010.

[33] Oliva A, Torralba A: Modeling the shape of the scene: a holistic representation of the spatial envelope. International Journal in Computer Vision, vol. 42, pp.145–175, 2001.

[34] Vedaldi, A., Lenc, K: Matconvnet: Convolutional neural networks for MATLAB. Proceedings of the 23rd Annual ACM Conference on Multimedia Conference, pp: 689–692, 2015.

[35] Guang-Hai Liu, Jing-Yu Yang, et al. Content-based image retrieval using computational visual attention model. Pattern Recognition, vol. 48, pp.2554–2566, 2015.