ISSN 1746-7659, England, UK
Journal of Information and Computing Science
Vol. 14, No. 3, 2019, *pp*.195-202

Image Retrieval Method based on Integration of Principal Component Analysis and Multiple Features

Jingji Zhao

School of Mathematics and Statistics, Nanjing University of Information Science & Technology, Nanjing, 210044, China

(Received May 11 2019, accepted July 20 2019)

Abstract: Existing content-based image retrieval methods exist some drawbacks, such as low retrieval precision, unstable performance. To address these drawbacks, in this paper a content-based image retrieval method is presented based on multi-feature fusion of principal component, oriented-gradient and color histogram. The idea for the proposed method is: firstly, input image is grayscale and flattened into a one-dimensional vector, and the first n principal components from the vector yielded by the PCA algorithm are extracted, in other word, input image is represented as a n×1 dimensional PCA feature vector. Secondly, to remedy color and orientation information missed by PCA, oriented-gradient and color histograms are used to extract orientation and color features respectively. Thirdly, extracted oriented-gradient and color histograms are merged with PCA features to generate the multi-feature representation of the input image. This paper confirms that the proposed multi-feature method can better represent an input image and can easily measure the similarity between images. The experiments are carried out and evaluated based on Corel-1000 , the target method is significantly better than the four popular methods**.**

Keywords: contented-based image retrieval(CBIR); principal component analysis(PCA); histogram of oriented gradient(HOG); histogram of color(HoC); similarity

1. Introduction

In the past two decades, content-based image retrieval (CBIR) technology has been developed rapidly, and new technologies and methods have emerged in endlessly. As the name implies, CBIR realizes image matching and retrieval based on the inherent features of the image itself. Here, that features of the images generally refer to natural features such as the image color, the texture and the shape. For a given image, the CBIR method pre-extracts the extracted features and is effectively characterized by the extracted features, using such a feature to identify a given image and measure the similarity of the features between the images.

Color feature is one of the basic features of image. Histogram of Color(HoC) [1-2]is the most common representation of image color feature. Because it has feature invariance in terms of geometric changes such as translation, rotation, scaling and so on, it shows good robustness. The texture is another important feature of the image, and it is generally considered that the texture is the regular arrangement combination of the texture elements, and the region with the repeatability, the simple shape and the intensity is regarded as a texture element. Since Dalal et al proposed Histogram Oriented Gradient (HOG) [3-4], HOG has become one of the most important texture feature extraction methods. HOG has been widely used in face recognition, image retrieval and other fields because of its excellent ability to depict local targets.

In recent years, the image retrieval technology based on the combined features has become a hot spot in the research field of CBIR. In 2017, Pavithra et al. [5]proposed a hybrid framework of CBIR, which first uses color moment features for preliminary retrieval, then further uses LBP and Canny to extract texture and edge features and carries out secondary retrieval of preliminary retrieval results. In 2018, Liu et al. [6]proposed a CBIR system based on the fusion of texture features and shape features based on non-downsampling shear transformation and low-order quaternion polar coordinate transformation. In 2019, Pavithra et al.[7]proposed an effective seed point selection method with Dominant Color Descriptor(DCD)which improves the retrieval accuracy of CBIR system based on DCD. In fact, in these methods, different underlying visual features are extracted and combined, but the combination of such features does not always ensure better retrieval accuracy[6] [8].

Since the Eigenface algorithm [9] was proposed last century, the research of face recognition algorithm based on principal component analysis (PCA) [10-12] has been continuing. Chan et al. [13]proposed the PCANet method of applying PCA to CNN, using PCA to learn multi-level filter banks, which accelerates the

DOI https://doi.org/2024-JICS-22413 | Generated on 2025-04-09 07:43:40

training speed of neural network, and the accuracy of face recognition reaches the most advanced level at that time. Yao et al. [14] proposed a PCA-based face recognition algorithm, which uses Support Vector Machine (SVM) and Adaptive Boosting as a classifier and has a significant effect. The PCA algorithm is also widely used in the fields of data compression and image compression transmission[15-16].These applications are based on the fact of compression, such a fact shows that the compression process of PCA algorithm is essentially to extract the key feature information of the compressed object, and the PCA principal component is actually a feature representation of the compressed object. At the same time, as a classical compression dimension reduction algorithm, PCA principal component has the advantages of low computational complexity, fast computing speed, strong ability to extract image spatial structure features, which will be verified in section 2.Inspired by this fact, this paper studies a new method different from the traditional CBIR technology. This method is based on the principal component feature extracted by PCA and combines the HOG and HoC as the auxiliary image feature representation. For the convenience of description, it is abbreviated as the PHH method. It is proved that the PHH method is more effective and robust and has higher computational performance than the latest CBIR method.

2. PHH Method

2.1 The Feature Extraction And Feature Representation of PCA

Set as $I_1, I_2, \dots, I_n$ are images taken from the image data set. In order to extract the principal components of the given image I_i and their feature representations, the processing steps are as follows:

(1) The image I_i（$i = 1,2,\cdots,n$）is transformed into a one - dimensional column vector and will be the column(i) of the image matrix P,that is:

$$\mathbf{P} = \left(\mathbf{I}_1^{(1)}, \mathbf{I}_2^{(1)}, \dots, \mathbf{I}_n^{(1)}\right) \qquad （1）$$

Here, Matrix P is $m \times n$, m is the number of pixels of the image I_i and n is the number of amplitudes of the image.

(2) Zero-centering the image set matrix $P = \begin{pmatrix} \boldsymbol{p}_1 \\ \boldsymbol{p}_2 \\ \vdots \\ \boldsymbol{p}_m \end{pmatrix}$, of which $\boldsymbol{p}_j$ is the j-th row of the matrix P, and the zero-center of the j-th row is as follows:

$$\boldsymbol{q}_j = (p_{j1} - \bar{p}_j, p_{j2} - \bar{p}_j, \cdots, p_{jn} - \bar{p}_j) \qquad （2）$$

Here p_{jk} is the k-th element of $\boldsymbol{p}_j$ and $\bar{p}_j = \frac{1}{m}\sum_{k=1}^{m} p_{jk}$ is the mean of the k-th line of matrix P. The zero-centered matrix of the image set matrix P is obtained by row zero-center calculation:

$$\mathbf{Q} = \begin{pmatrix} \boldsymbol{q}_1 \\ \boldsymbol{q}_2 \\ \vdots \\ \boldsymbol{q}_n \end{pmatrix}$$

(3) Calculate the eigenvalues of the covariance matrix $C = \frac{1}{m} Q Q^T$ of Q and its corresponding eigenvectors, and arrange the n eigenvalues in descending order, take the eigenvectors corresponding to the first k largest eigenvalues and construct the feature matrix $\boldsymbol{F}$ in order.

(4) Transform the zero-cenetred matrix Q into the feature space $\boldsymbol{F}$ to obtain the reduced-dimensional data set $\boldsymbol{P^F}$:

$$\boldsymbol{P^F} = \mathbf{F}^T \mathbf{Q} \qquad （3）$$

Each column corresponds to the PCA principal component feature of a particular image in the date set after dimension reduction $\boldsymbol{P^F}$.The PCA principal component feature of image I_i is the column vector i of the matrix $\boldsymbol{P^F}$,recorded as $\boldsymbol{P_i}^{\boldsymbol{F}}$.Conversely,the approximate restored image I_i with PCA feature $\boldsymbol{P_i}^{\boldsymbol{F}}$ can be obtained using PCA feature $\boldsymbol{P_i}^{\boldsymbol{F}}$ and feature matrix $\boldsymbol{F}$,the reduction formula is:

$$\mathbf{I}_i \approx \boldsymbol{F}\boldsymbol{P}_i^{\boldsymbol{F}} \qquad （4）$$

For any one of the query images I_{query}, the extraction process of principal components is as follows:

(1) Transform I_{query} into a one-dimensional column vector $I_{query}^{(1)}$;

(2)Zero-centering $I_{query}^{(1)}$, and record the zero-centered vector as $\boldsymbol{Q}_{query}$, that is:

$$\boldsymbol{Q}_{query} = \begin{pmatrix} i_1 - \bar{p}_1 \\ i_2 - p_2 \\ \vdots \\ i_n - p_n \end{pmatrix} \tag{5}$$

Here i_j is the j-th element of $I_{query}^{(1)}$.

(3) Calculate the principal components of I_{query}:

$$\boldsymbol{P}_{query}^{\mathrm{F}} = \boldsymbol{F}^{\mathrm{T}} \boldsymbol{Q}_{query} \tag{6}$$

Here $\boldsymbol{P}_{query}^{F}$ is the first k PCA principal components of the query image I_{query}.

The formula (3) and (6) respectively give k principal component feature representations of all n images $I_1, I_2, \dots, I_n$ and arbitrary query images I_{query} in the image data set..

2.2 The relationship between the number of Principal components of PCA and the representation of Image Features

In order to analyze the influence of the number of principal components of image PCA on the validity of image feature representation, Fig.1 shows that from left to right columns (a) to (e) are original images and corresponding images that are restored in the number k of principal component is 10,20,50,100, respectively. It is easy to see that when the number of principal components k=10, the restored image is the most blurred, and the spatial structure and edge details of the image are seriously lost. As the number of principal components increases.At k=50, the restored image has good spatial structure and edge details.It can be seen that the selection of the number of principal components k has a crucial impact on the image feature retention. In other words, the more the number of principal components, the

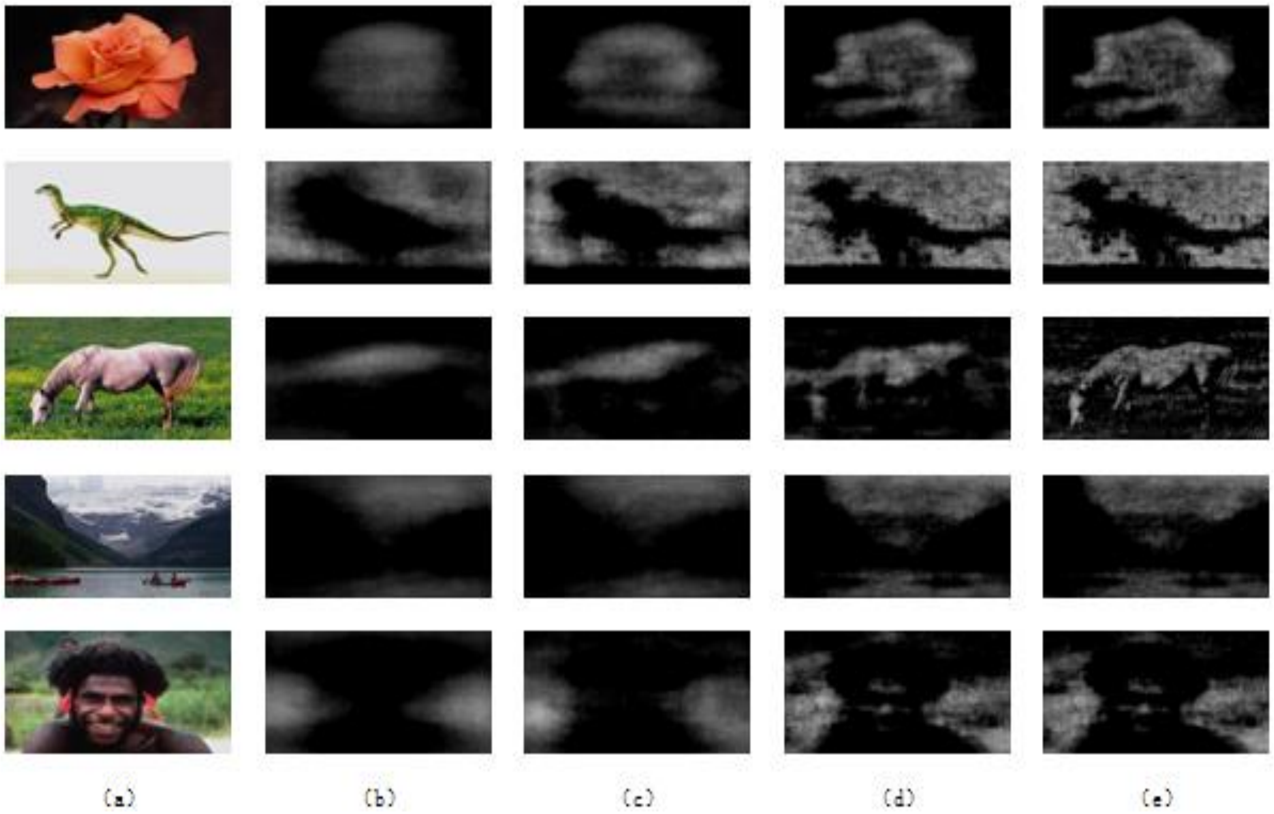

Fig.1 Impaction on image restoration while number of PCA components n is taken different value: (a) original images;（b）images restored when n=10;（c）Images restored when n=20;（d）Images restored when n=50;（e）Image restored when n=100

richer the spatial structure features extracted from the original image will be, and the more effective the image feature representation will be.However, the value of the number of principal components k should not be too large, otherwise the true meaning of PCA feature image representation would be lost. Therefore, the number of principal components k should be selected with a moderate size.A large number of experimental observations and analysis show that it is a good balance to select the number of principal components k=50. Under such a number of principal components, the recovered images lose less information, the principal components have a higher image feature identification effect, and the calculation performance is more efficient. The number of principal components k in the subsequent work of this paper is all 50.

In order to identify the image in order to prove the image PCA features, that is similar to the image characteristics of PCA is similar, with the characteristics of the geometric invariance, needs to be defined between two images based on PCA feature similarity measure, if two images are similar means that the measurement on the distance is short, on the contrary, if the two images in human visual sense difference is big, then the measurement in the distance is big.

In order to test whether PCA features can depict the differences between images, the absolute difference of principal component vectors of two images is simply performed and the image is restored. Fig.2,Fig.3 shows some experimental results, from left to right, column (a) is the original image, column (b) is the PCA principal component number n = 50 reduction of grayscale, column (c) as the PCA principal component absolute difference of two images generated grayscale image restoration, column (d) for the grayscale in column (c) after the threshold value of binary figure.It can be seen from Fig.2, two visual images PCA principal component feature vectors are similar to that of absolute difference, in absolute deviation by their principal component grayscale, partial light pixels in the minority, in order to be able to clearly observe the distribution of the non-zero pixels for gray image binarization processing, in the column (d) of the corresponding binary image can be more clearly observed in the white pixels in the whole image of less absolute proportion, this suggests that their corresponding visual meaning similar image principal component eigenvectors of absolute difference in recovery after binary image white pixels is sparse,If the percentage of white

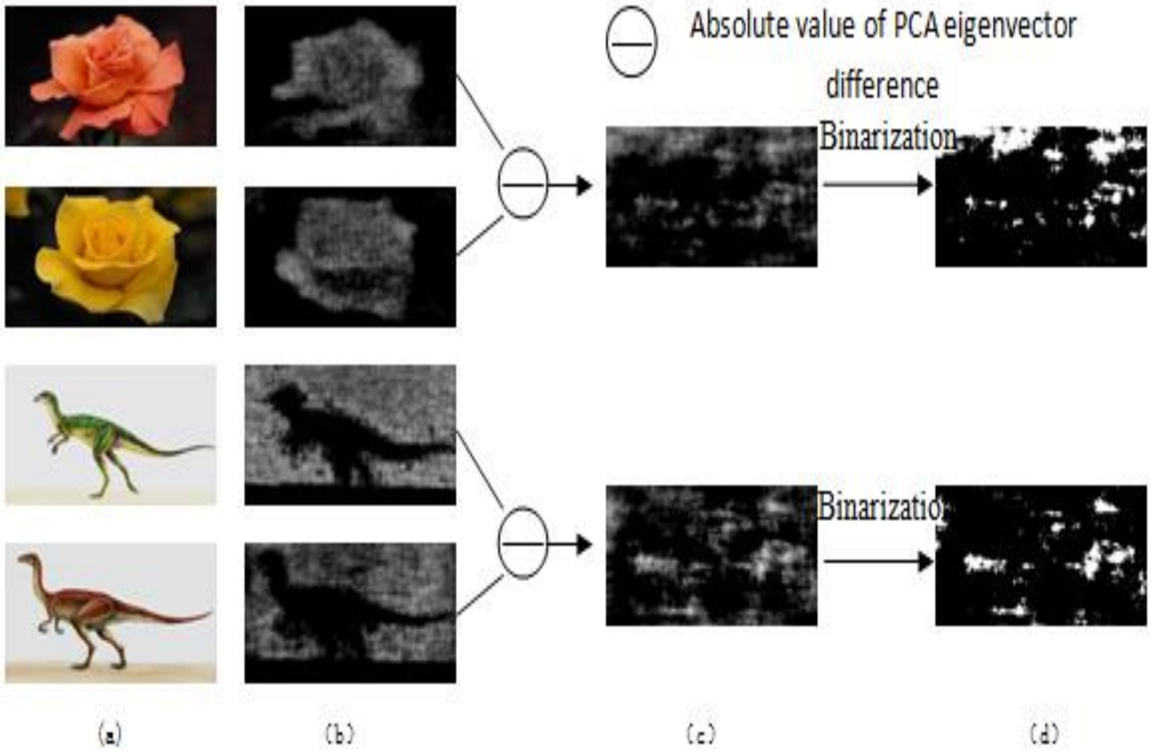

Fig.2 Illustration of image restored from absolute difference between principal components of similar images: (a) original images; (b) images restored from corresponding PCA features when n=50; (c) images restored from absolute difference between principal components of similar images; (d) images yielded by binarizing (c)

DOI https://doi.org/2024-JICS-22413 | Generated on 2025-04-09 07:43:40

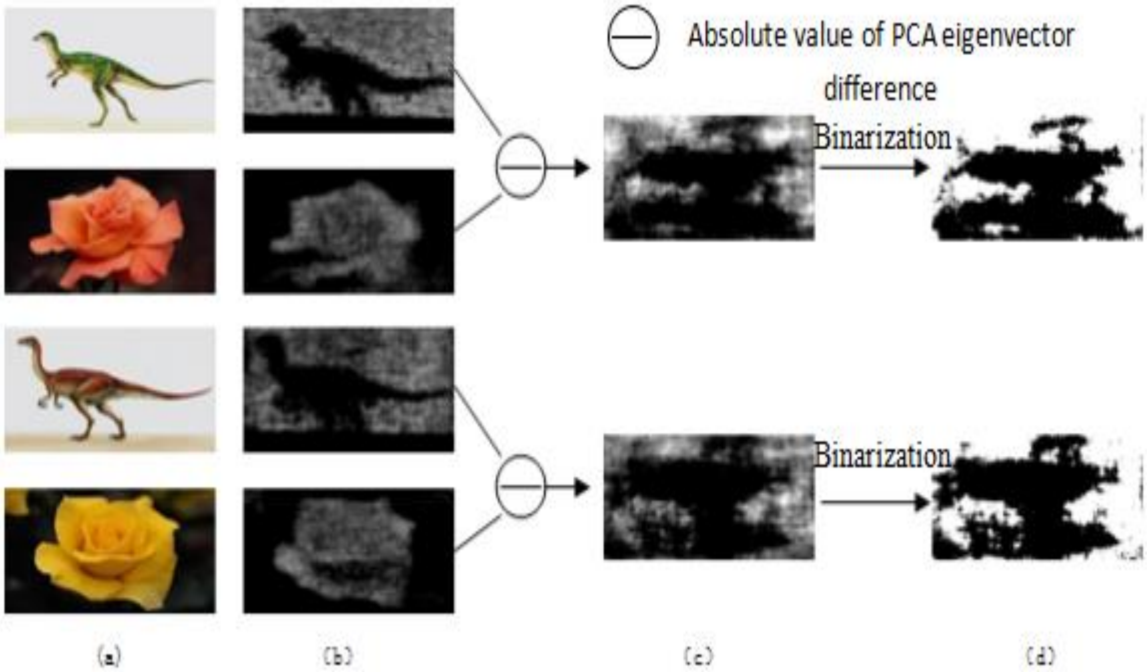

Fig. 3 Illustration of images restored from absolute differences between principal components of significantly unlike images in human vision: (a) original images; (b) images restored from corresponding PCA features when n=50;（c）images restored from absolute difference between principal components of unlike images;（d）images yielded by binarizing (c)

pixels in the whole image is used to measure the similarity, the ratio is a small value.On the contrary, it can be seen from Fig. 3 that for two images with significant difference in visual significance, the proportion of the gray-scale images recovered by the absolute difference of their principal components is relatively bright, and the proportion of white pixels in the corresponding binary images is significantly increased.The above observation and analysis results show that the PCA feature of the image is an important feature expression of the spatial structure and distribution feature of the image, which can perfectly depict the spatial structure and distribution feature of the image, and the simple absolute difference of principal component feature vectors between images can be used as a measure of the similarity between images.

2.3 Feature Fusion

Figure 1,2 and 3 also shows that PCA principal component of the image characteristics in record of the image spatial structure and its distribution is extremely effective, however, what needs to be especially emphasized is restored image is a gray image, which means that the PCA features can't keep color information and keeping the texture feature is also weak, in order to make up the color and texture feature, integrate HOG and HoC features into PCA principal component to construct a more comprehensive and more robust image characteristics.

$$\mathbf{F}^{\mathbf{PHH}} = \left(\mathbf{F}^{\mathbf{PCA}}, \mathbf{F}^{\mathbf{HOG}}, \mathbf{F}^{\mathbf{HoC}}\right) \tag{7}$$

3. Experiment

3.1 Experimental parameters and methods

It is known from section 2.2 that it is a better choice to take the PCA eigenvector F^{PCA} composed of the first 50 PCA components of each image, so the PCA eigenvectors of the images in the experiment are all 50-dimensional column vectors. The selection of image HOG feature vector is mainly dependent on the partition of the number of directions, blocks and cells. In this paper, the HOG feature vector is divided into 9 directions, 1 block and 4*4 cells. Therefore, the HOG feature vector $F^{HOG}$is a 144-dimensional column vector.The selection of image HoC feature vector depends on the quantization level. In this experiment, the image HoC feature vector adopts a 16*4*4 quantization scheme, so HoC feature vector F^{HoC} is a 256-dimensional column vector. The image feature vector F^{PHH}after the fusion of the three feature vectors F^{PCA}、F^{HOG}、F^{HoC} is a 450-dimensional column vector.

DOI https://doi.org/2024-JICS-22413 | Generated on 2025-04-09 07:43:40

For the image data set Corel-1000,30 images are randomly extracted from each class, and a total of 10 classes of 300 images are extracted as the query image sets. For each query image, calculate the F_{query}^{PHH} characteristic vector, and each image feature vector and feature data set $F_{target}^{\mathrm{PHH}}$ each matching query images and retrieve images, the similarity between $F_{query}^{PHH}$and $F_{target}^{\mathrm{PHH}}$, here using Pearson correlation coefficient[17-18] measurement, will all be matching similarity in descending order, take the top 20 of the characteristics of the highest similarity as the query result sets, then evaluate the precision and recall rate[19-20] of the query result set.

3.2 Evaluation of experimental results

Tab.1 and Fig.4 shows that the PHH method on the Corel-1000 data set is compared with the four best CBIR methods CDDWT[21], SHWAS[22], LFDSR[23] and CIRCCODF[24] based on the index Precision in the past three years, and the comparison is based on classification and average. In the classification, the CDDWT method and the PHH method in this paper show outstanding performance. The CDDWT method is superior to the PHH method in classifying africa, cars, elephants, mountains and food, and the PHH method is superior to the CDDWT method in architecture, flowers and horses. But on the beach, the dinosaur's precision was even. The enemy. However, on the overall average of 10 classes, the average accuracy of PHH method is 73.8%, and that of CDDWT method is 73.5%, which indicates that PHH method is slightly superior to CDDWT method.

In order to compare the recall rate, the amplitude of the query result set image is changed from 20, 30, 40 to 100. Fig.5 shows the Recall diagram of each method on the data set Corel-1000. The PHH and CDDWT methods are obviously superior to the other three methods. Especially after the first 40 images, the average recall rate advantage of PHH

Tab.1 Precision comparison of top20 between PHH method and other methods on Corel1000

Class	CDDWT[21]	LFDSR[22]	CIRCCODF[23]	SHWAS[24	PHH
africa	75.0%	75.0%	54.0%	55.0%	72.0%
beach	60.0%	60.0%	38.0%	53.0%	60.0%
building	50.0%	42.0%	30.0%	61.0%	52.0%
car	90.0%	68.0%	64.0%	90.0%	84.0%
dinosaur	100%	95.0%	96.0%	98.0%	100%
elephant	80.0%	68.0%	61.0%	58.0%	68.0%
flowers	85.0%	95.0%	68.0%	90.0%	94.0%
horse	70.0%	92.0%	76.0%	78.0%	94.0%
mountain	55.0%	36.0%	44.0%	52.0%	48.0%
food	70.0%	54.0%	52.0%	70.0%	66.0%

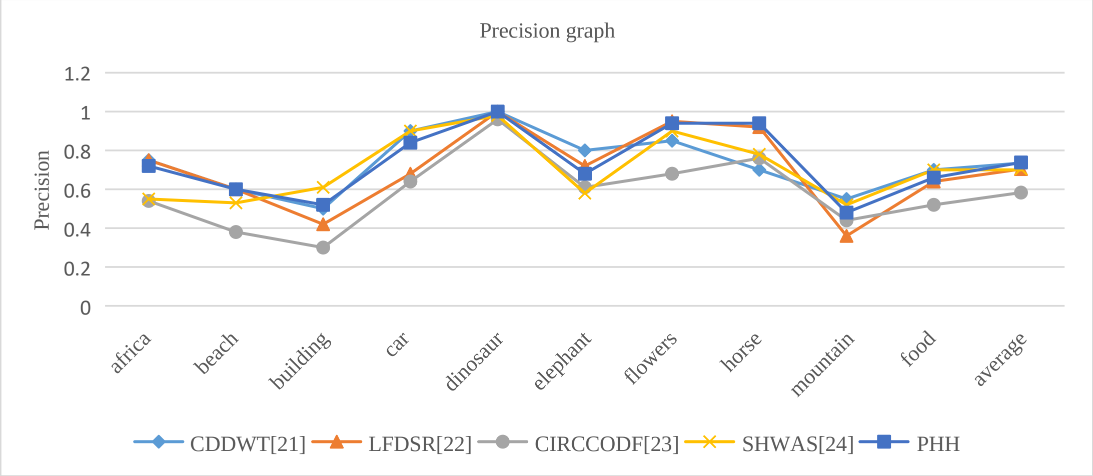

Fig. 4 Retrieval precision comparison between PHH and other methods based on dataset Corel1000

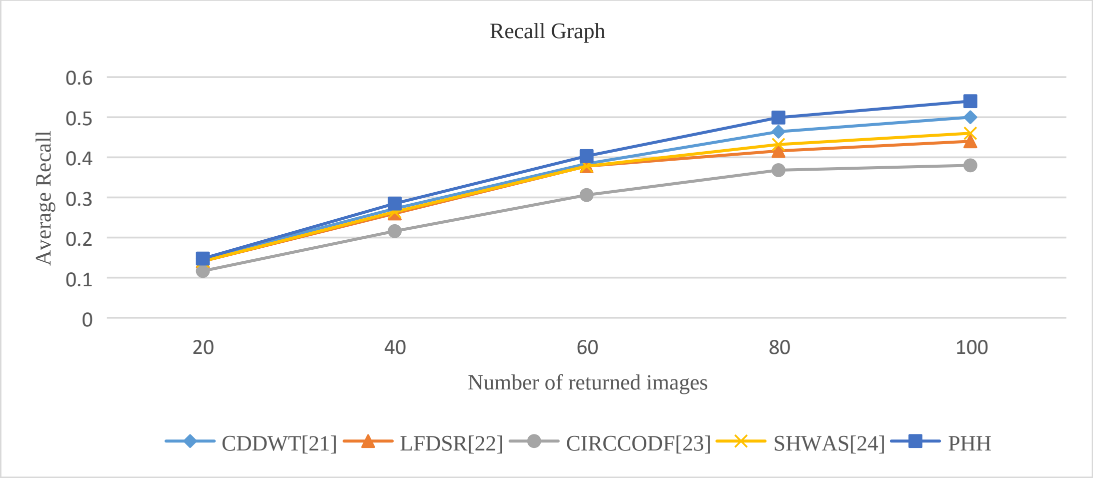

Fig. 5 Average recall comparison between PHH and other methods based on Corel-1000

method begins to increase significantly.

Based on Corel-1000 data set and two evaluation indexes of accuracy and recall, the PHH method proposed in this paper has significant advantages. Moreover, the CBIR system based on PHH method has excellent robustness on the whole, and its computational performance is obviously better than the four methods compared.

4. Conclusion

This paper proposes a new CBIR method based on PCA, HOG and HoC , the method uses the PCA algorithm to extract the main components of the image and uses it as the main feature component of the image feature representation, however, because of the PCA principal component will not be able to keep the image color and texture information, therefore, HOG and HoC feature extraction model is constructed and the direction of the image information and information into PCA principal component. Pearson correlation coefficient method is used in similarity measurement. Based on the common data set of Corel-1000 and two evaluation indexes, the PHH proposed in this paper was comprehensively analyzed and evaluated, and compared with the most excellent recent related methods, which verified that PHH method is a CBIR method with high precision, stable performance and low calculation cost. Future research efforts will focus on the field of image retrieval in deep hash networks.

References

[1] J. Chauveau, D. Rousseau, P Richard, et al. Multifractal analysis of three-dimensional histogram from color images[J]. Chaos Solitons & Fractals, 2010, 43(1-12):57-67.

[2] P Liu , J Guo , K Chamnongthal, et al. Fusion of color histogram and LBP-based features for texture image retrieval and classification[J]. Information Sciences, 2017, 390:95-111.

[3] N Dalal, B Triggs. Histograms of Oriented Gradients for Human Detection[C]// null. IEEE Computer Society, 2005.

[4] N Dalal. Finding people in images and videos[D]. Institut National Polytechnique de Grenoble-INPG, 2006.

[5] L K Pavithra, T S Sharmila. An efficient framework for image retrieval using color, texture and edge features[J]. Computers & Electrical Engineering, 2017:S0045790616308229.

[6] Y Liu, S Zhang, Y Sang , et al. Improving image retrieval by integrating shape and texture features[J]. Multimedia Tools and Applications, 2018.

[7] L K Pavithra, T Sharmila. An efficient seed points selection approach in dominant color descriptors (DCD)[J]. Cluster Computing, 2019.

[8] X Wang, L Liang, Y Li, et al. Image retrieval based on exponent moments descriptor and localized angular phase histogram[J]. Multimedia Tools and Applications, 2017, 76(6):7633-7659.

[9] M Turk . Eigenfaces for recognition[J]. J. Cognitive Neuroscience, 1991, 3.

[10] P Wagh, R Thakare, J Chaudhari, et al. Attendance system based on face recognition using eigen face and PCA algorithms[C]// International Conference on Green Computing & Internet of Things. 2015.

[11] I T Jolliffe. Pincipal Component Analysis[J]. Journal of Marketing Research, 2002, 25(4):513.

[12] A Herve, L J Williams. Principal component analysis[J]. Wiley Interdisciplinary Reviews Computational Statistics, 2010, 2(4):433-459.

[13] T H Chan, K Jia , S Gao , et al. PCANet: A Simple Deep Learning Baseline for Image Classification?[J]. IEEE Transactions on Image Processing, 2015, 24(12):5017-5032.

[14] M Yao, C Zhu . SVM and adaboost-based classifiers with fast PCA for face reocognition[C]// 2016 IEEE International Conference on Consumer Electronics-China (ICCE-China). IEEE, 2016.

[15] J Guo, R Xie, H Liu. A Hybrid Method for NMR Data Compression Based on Window Averaging (WA) and Principal Component Analysis (PCA)[J]. Applied Magnetic Resonance, 2018.

[16] H David , B Laura , J A Fessler. Asymptotic performance of PCA for high-dimensional heteroscedastic data[J]. Journal of Multivariate Analysis, 2018:S0047259X17304852-.

[17] P C Coefficient. Pearson’s correlation coefficient[J]. New Zealand Medical Journal, 1996, 109(1015):38.

[18] J Ahlgren. Requirement for a Cocitation Similarity Measure, with Special Reference to Pearson's Correlation Coefficient[J]. Journal of the American Society for Information Science & Technology, 2003, 54(6):550-560.

[19] V Raghavan, P Bollmann, G S Jung. A critical investigation of recall and precision as measures of retrieval system performance[J]. Acm Transactions on Information Systems, 1990, 7(3):205-229.

[20] M Buckland , F Gey . The relationship between recall and precision[M]. John Wiley & Sons, Inc. 1994.

[21] R Ashraf, M Ahmed, S Jabbar, et al. Content Based Image Retrieval by Using Color Descriptor and Discrete Wavelet Transform[J]. Journal of Medical Systems, 2018, 42(3):44.

[22] C Celik, H S Bilge. Content based image retrieval with sparse representations and local feature descriptors : A comparative study[J]. Pattern Recognition, 2017, 68:1-13.

[23] H H Tsai, B M Chang, P S Lo, et al. [IEEE 2016 First IEEE International Conference on Computer Communication and the Internet (ICCCI) - Wuhan, China (2016.10.13-2016.10.15)] 2016 First IEEE International Conference on Computer Communication and the Internet (ICCCI) - On the design of a color image retrieval method based on combined color descriptors and features[J]. 2016:392-395.

[24] S Yu, D Niu, L Zhang, et al. Colour image retrieval based on the hypergraph combined with a weighted adjacent structure[J]. Iet Computer Vision, 2018, 12(5):563-569.