2021 International Conference on Communications, Information System and Computer Engineering (CISCE) | 978-1-6654-0352-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/CISCE52179.2021.9445921

Image Feature Fusion and Fisher Coding based Method for CBIR

Dayou Jiang[1*]
Department of Computer Science and Technology
Anhui University of Finance and Economics
Bengbu, China
ybdxgxy13529@163.com

Abstract—**The paper proposed a new method for content-based image retrieval (CBIR) based on image feature fusion and fisher encoding (FV). Firstly, low-level image content features such as hue-saturation-value (HSV) histogram, uniform local binary patterns (LBP), Dual-Tree complex wavelet transform (DTCWT) are extracted based on image blocks. In contrast, high-level features are extracted by using the AlexNet convolutional neural network (CNN). The singular value decomposition (SVD) was applied to the LBP and DTCWT. Secondly, low-level features are fused using normalization and weights. Lastly, after using the FV encoding, the fused fisher vectors are used to measure the similarity of image pairs. The experimental results on the benchmark Corel-1k show that the accuracy on the top 10, 12, and 20 images returned are 93.4%, 92.8%, and 91.4%, respectively.**

Keywords- image feature fusion; fisher encoding; image retrieval; Corel-1k

I. Introduction

As the amount of available data grows, the explosion of images has brought more severe challenges to automated image search and retrieval. The accuracy of the results obtained by the image search engine query based on text annotation is often not ideal, so the research community favors content-based image retrieval.

The visual characteristics of images such as color, texture, and shape are widely used for image indexing and retrieval. In recent years, various image retrieval methods have sprung up. Color features such as HSV histogram [1] are extensively used in CBIR, which may be endorsed to better-colored images over the grayscale images. The modified color motif co-occurrence matrix (MCMCM) [2] collects the inter-correlation between the red, green, and blue color planes which is absent in the color motif co-occurrence matrix. Texture features can represent internal spatial structure information (repetitive patterns) of images. The most commonly used is the LBP [3]. Inspired by the recognition accuracy and simplicity, several variants of LBPs have been proposed, such as local tetra patterns (LTrPs) [4], Directional local extrema patterns (DLEP) [5]. Besides, Wavelet-based texture features from DTCWT [6], shape adaptive discrete wavelet transform (SADWT) [7] are studied. The DLEP algorithm allows extracting directional edge information based on local extrema. The SADWT can work on each image region separately and preserve its spatial and spectral properties. The extraction of shape features is mainly performed to capture the shape attributes (such as bending moment, area and boundary, etc.) of the image item. Shape-based methods are grouped into region-based and counter-based techniques. Methods such as Fourier descriptors [8], geometric moments [9] are often used. Some ways extract image feature descriptors by using the compression scheme, such as Vector Quantization (VQ) [10], Ordered Dither Block Truncation Coding (ODBTC) [11]. ODBTC compresses an image block into corresponding quantizers and bitmap images. The dither array in ODBTC method substitutes the fixed average value as the threshold value for the bitmap image generation. Image retrieval based on deep learning has become a hot research field with the increase of image data sets. The deep learning techniques allow the system to learn features with CNN architecture. CNN-based techniques are also incorporated to extract image features at various scales and encoded using BoW or vector of locally aggregated descriptors (VLAD). Recently, Some CNN-based CBIR methods have been proposed, such as embedded neural networks with band letized regions (ENN-BR) [12], LeNetF6 [13], Shape-Based Filtering, and Spatial Mapping Integrated with CNN (SBF-SMI-CNN) [14]. Using a low-level function to distinguish between different images usually has some limitations. Therefore, many methods based on the fusion of several low-level features have been proposed, such as CCM and difference between pixels of scan pattern (CCM-DBPSP) [15], color histogram and local directional pattern (CH-LDP) [16], Color-Shape and Bag of Word(C-S-BOW) [17], micro-structure descriptor and uniform local binary patterns (MSD-LDP) [18], Correlated Microstructure Descriptor (CMSD) [19]. The CMSD is used for correlating color, texture orientation, and intensity information. The Fused Information Feature-based Image Retrieval System (FIF-IRS) [20], which is composed of 8-Directional Gray Level Co-occurrence Matrix (8D-GLCM) and HSV Color Moments (HSVCM).

The proposed method uses the FV encoding [22] on low-level features, and CNN features extracted from AlexNet architecture [23]. The low-level fused features are based on HSV, LBP, and DTCWT. The SVD is applied to LBP and DTCWT for dimension reduction [24]. The Fisher encoding uses a Gaussian Mixture Model (GMM) to construct a visual word dictionary. In the experiments, the fused encoded features are used for the CBIR task. The remainder of this paper is organized as follows. Section 2 described the proposed method. Section 3 presented the experimental results. Conclusions are derived in Section 4.

978-0-7381-1215-2/21/$31.00 ©2021 IEEE

II. Methods and Meterials

A. *Fisher vectors encoded fused low-level feature HLD-FV*

The process of the HLD-FV feature extraction is declared in Fig. 1.

Step 1: Resize the query image with the size 256×256. Divide the image into 4×4 non-overlapping sub-bands of the same size. Regroup the image into six new image blocks (as shown in Fig. 2).

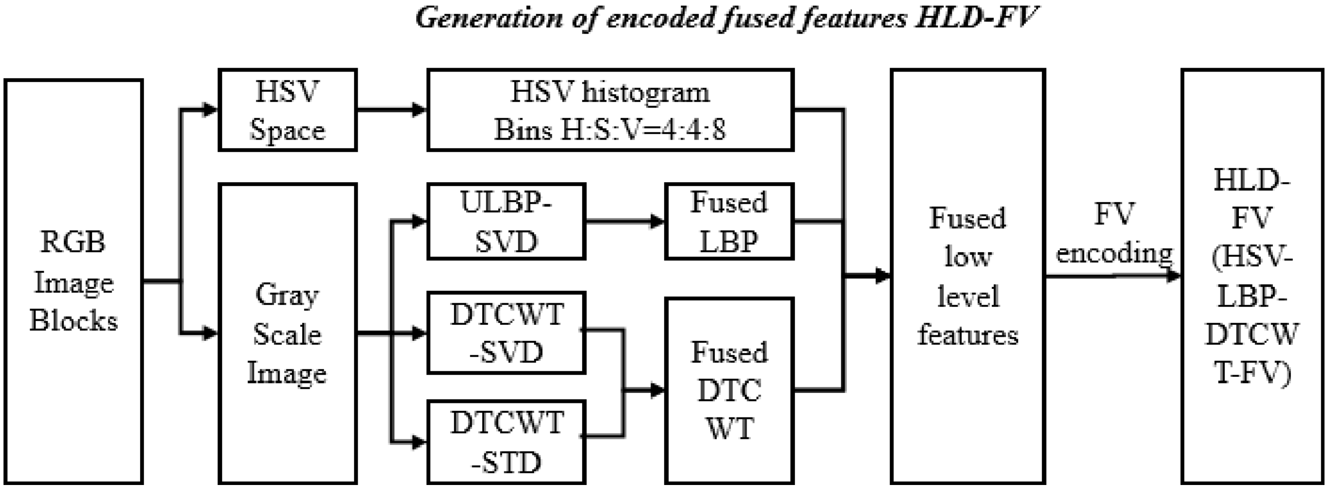
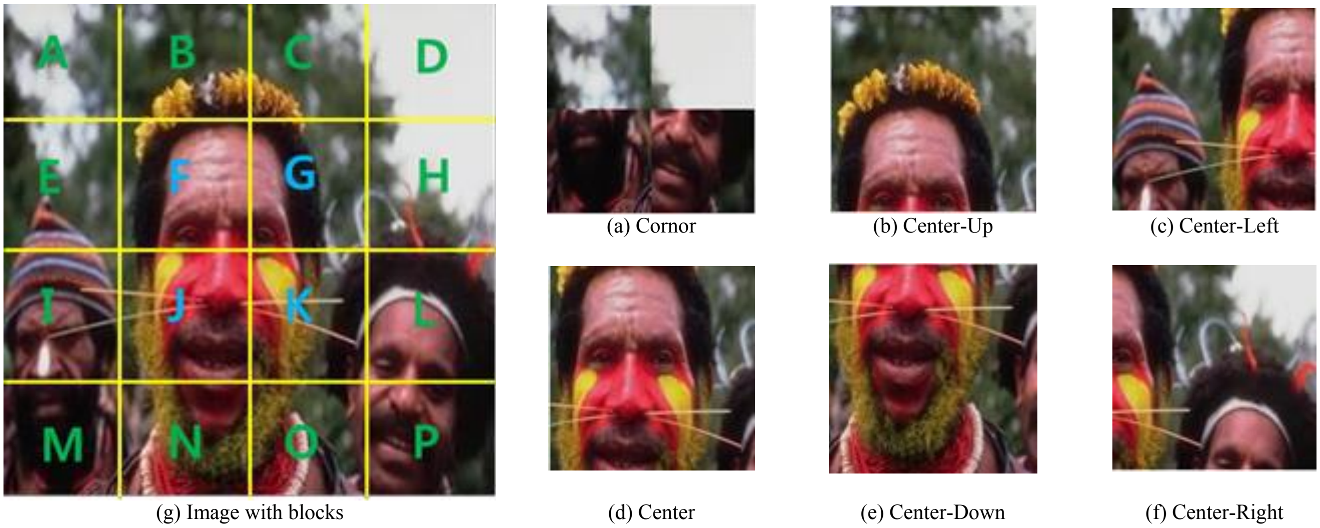


Figure 1. Generation of encoded fused features HLD-FV.



(a) Cornor

(b) Center-Up

(c) Center-Left

(g) Image with blocks

(d) Center

(e) Center-Down

(f) Center-Right

Figure 2. Generating image blocks of an sample image by using crossed-blocking.

Step 2: For each image block, generate LBP features (ULBP-SVD) and DTCWT features (DTCWT-SVD and DTCWT-STD) refer to Ref [24]. The lowest real coefficients and complex coefficients of three different scales are all selected to generate feature vectors. The maximum singular value of wavelet coefficients in each scale is chosen. The Standard Deviation (STD) variances of coefficients are also calculated. To reduce the LBP features' dimensions, the SVD is applied to the first two dimensions of ULBP histograms.

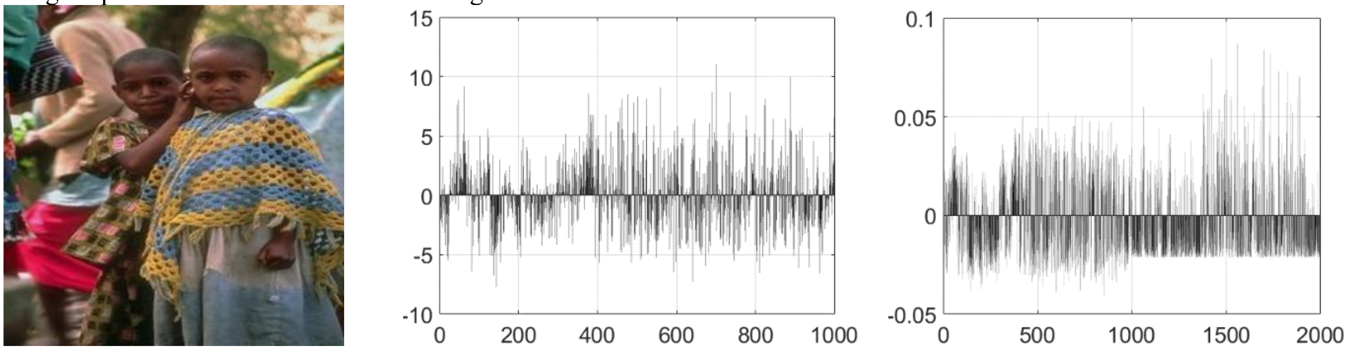
Step 3: Normalize each feature vector to sum as 1, and fuse the features by using the weights (w_1: w_2: w_3=2:6:2).

$$HLD = [w_1 * F_{HSV}, w_2 * F_{LBP}, w_3 * F_{DTCWT}) \qquad (1)$$

Step 4: Use FV encoding to encoded the fused feature HLD. The dimension of the feature vector HLD-FV is 2856 ($D=2_{Encoded}\times 6_{Blocks}\times(128_{HSV}+58_{LBP}+52_{DTCWT})$).

B. *CNN features extraction*

AlexNet achieved a top-5 error of 15.3% in the ImageNet Large Scale Visual Recognition Challenge in 2012. Its success has dramatically enhanced the enthusiasm of deep learning in various fields. AlexNet contained eight layers, the first five were convolutional layers, some of them followed by max-pooling layers, and the last three were fully connected layers. The first convolutional layer filters the 224×224×3 input image with 96 kernels of size 11×11×3 with a stride of 4 pixels [22]. The output of AlexNet is a 1000-way softmax. The fully connected layer Fc7 of size 4096 and Fc8 contains 1000 dimensions. The features of Fc8 are extracted from pre-trained AlexNet architecture. For the CBIR task, the paper uses FV encoding to process the Fc8 features. The length of encoded extracted Fc8-FV is 2000. Fig. 3 shows the CNN features of a sample query image.



(a) Image Q_1 (b) Fc8 of image Q_1 (c) FV encoded Fc8 of image Q_1

Figure 3. Generation of encoded fused features HLD-FV.

C. Method CHLD-FV

Fig. 4 illustrates the schematic diagram of the proposed CHLD-FV image retrieval method. The CHLD-FV is generated by fusing the low-level feature vectors LDH-FV and CNN feature vectors Fc8-FV. The LDH-FV is conducted on image blocks, and Fc8-FV is generated from the full resized image.

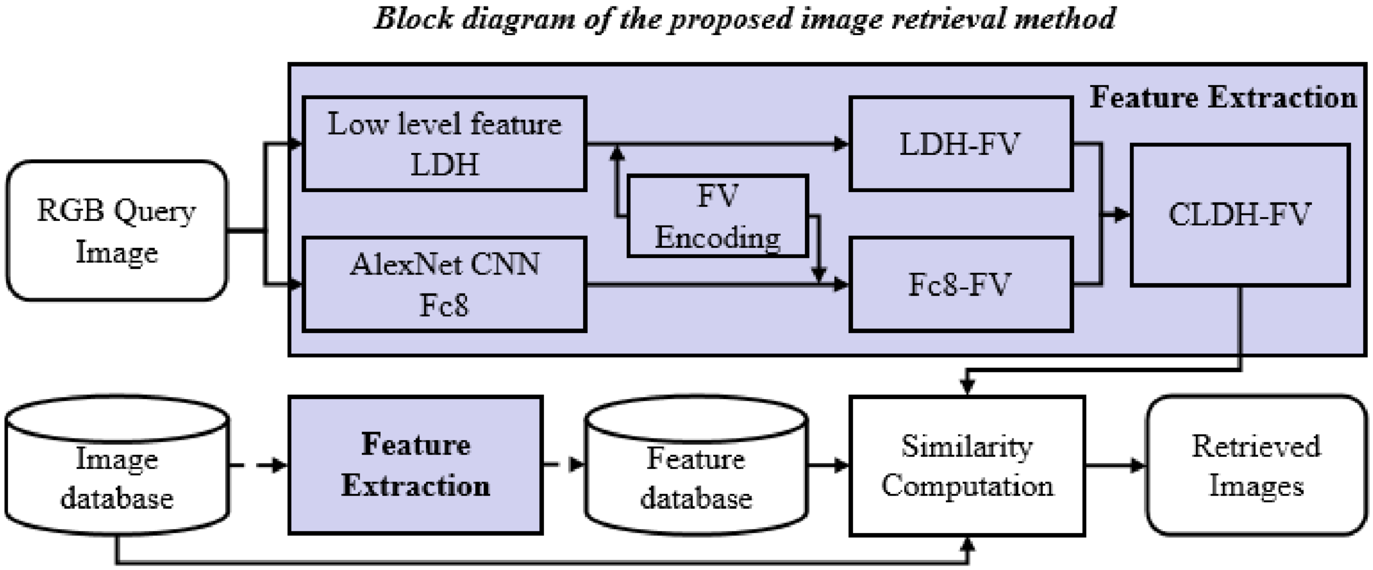


Figure 4. Block diagram of the proposed image retrieval method

D. Performance evaluation metrics

The L1 norm (Manhattan distance) is used for similarity computation. The average retrieval precision (ARP) is used to evaluate the performance of the different methods. Here, ARP(q, n) represents the nth image's average retrieval rate in the qth category. The sign |DB| denotes the total number of images. The sign q_nNR denotes the number of relevant images of the nth query image in the qth category in the number of retrieved images (NR). The M is the number of images in each category.

$$Mean(q) = \frac{1}{M}\sum_{n=1}^{M} ARP(q,n) \qquad (2)$$

$$ARP(q,n) = \frac{1}{N}\sum_{i=1}^{|DB|} q_n NR. \qquad (3)$$

III. Experimental Result

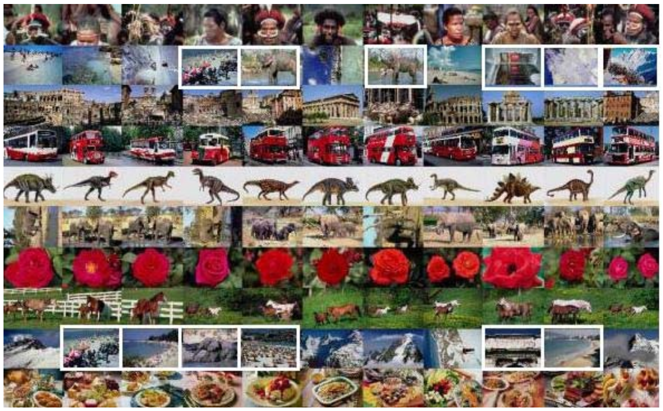
In this section, the proposed method's performance is tested using Corel 1k dataset [21] compared to state-of-the-art methods. The dataset has ten categories, and each category has 100 images with the size 256*384 or 384*256. The ten categories are Africans, Beaches, Buildings, Buses, Dinosaurs, Elephants, Flowers, Horses, Mountains, and Food. All images in the database are turned as query images, and then the retrieval precision of each image, each category, and the total dataset can be computed. To prove that the proposed method can perform image retrieval tasks and outperform some state-of-the-art methods. Three comparison experiments with different NR are considered. The NR is set as 20, 12, and 10, respectively. Table 1 lists the ARPs generated by different methods on each category of the Corel-1k dataset with the

NR=20. The bold values indicate the best results. The method HLD-FV performed better than DELP [5], MCMCM [2], VQ [10], ODBTC [11], CCM-DBPSP [15], CDH-LDP [16] and relatively worse than SADWT [7], C-S-BoW [17]. Using the CNN features, the performance of CHLD-FV has improved too much compared to HLD-FV, achieved the highest ARPs on Buses, Dinosaurs, Elephants, Flowers, Mountains, and Food categories.

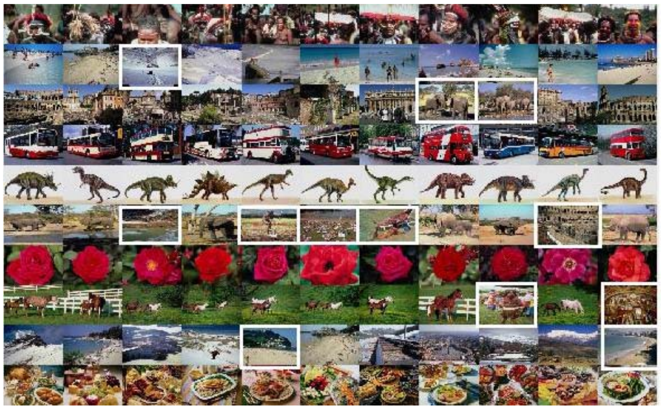
TABLE I. COMPARISON OF DIFFERENT IMAGE RETRIEVAL METHODS ON THE COREL-1K DATASET(NR IS 20).

Category	**Average Precision (%), NR=20**									
	DELP [5]	**MCMCM [2]**	**VQ [10]**	**ODBTC [11]**	**CCM-DBPSP [15]**	**SADWT [7]**	**CDH-LDP [16]**	**C-S-BoW [17]**	**Proposed HLD-FV**	**Proposed CHLD-FV**
Africans	74.3	69.7	70.2	84.7	70	80.1	77.9	**90**	69.2	74.9
Beaches	65.6	54.2	44.4	46.6	56	**75.3**	60.1	60	58.3	72.1
Buildings	75.4	63.9	70.8	68.2	57	77.1	69.1	**90**	73.2	90.3
Buses	98.2	89.6	76.3	88.5	87	99.2	87.6	75	96.6	**100**
Dinosaurs	99.1	98.7	**100**	99.2	97	99.5	99.4	**100**	99.8	**100**
Elephants	63.3	48.8	63.8	73.3	67	67.8	59.2	70	77.6	**98.9**
Flowers	94.1	92.3	92.4	96.4	91	96.0	95.8	90	93.4	**99.6**
Horses	78.3	89.4	94.7	93.9	83	88.2	91.8	**100**	92.2	99.0
Mountains	51.3	47.3	56.2	47.4	53	59.2	64	70	47.6	**81.7**
Food	85.0	70.9	74.5	80.6	74	82.3	78.1	90	81.4	**97.3**
Mean	78.46	72.5	74.3	77.9	74	82.47	78.31	83	78.93	**91.4**

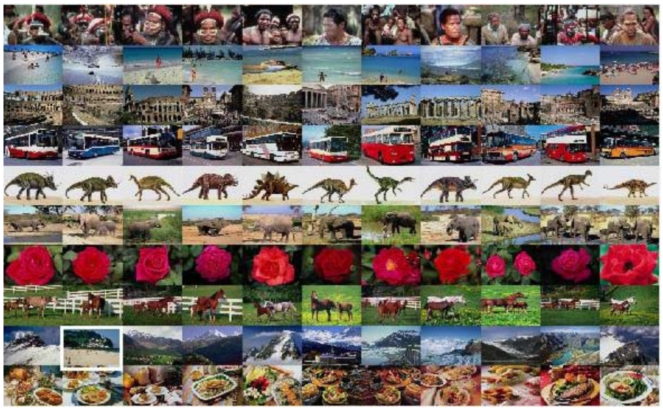
Fig.5 shows the top 10 retrieved images for each class of the Corel-1k dataset. Fig. 5 (a), (b), and (c) present the results generated by the ODBTC method and the proposed HLD-FV, CHLD-FV methods, respectively. The images highlighted with white boxes are the incorrectly retrieved ones. The ODBTC method performed relatively worse when handling the images within "Beaches" and "Mountains." The proposed HLD-FV method retrieves errors in beaches, buildings, elephants, horses, and mountains. In comparison, the CHLD-FV method performed much better.



(a) ODBTC



(b) HLD-FV



(C) CHLD-FV

Figure 5. Example of top 10 retrieved images for each class

Fig. 6 shows the ARPs of the Corel-1k dataset with the NR=12. Here, the methods such as CMSD [18] and MSD-LBP [19] are based on low-level feature fusion. The methods of

ENN [12] and LeNetF6 [13] are based on CNN features. However, those CNN-based methods' performance is not yet ideal and is lower than the HLD-FV method on average. The performance has related to the structure of the neural network. The MSD-LBP performed better than HLD-FV on categories such as Africans, Buildings, Buses, Flowers, Mountains, and Food. The CHLD-FV has the best performance except in the Africans category

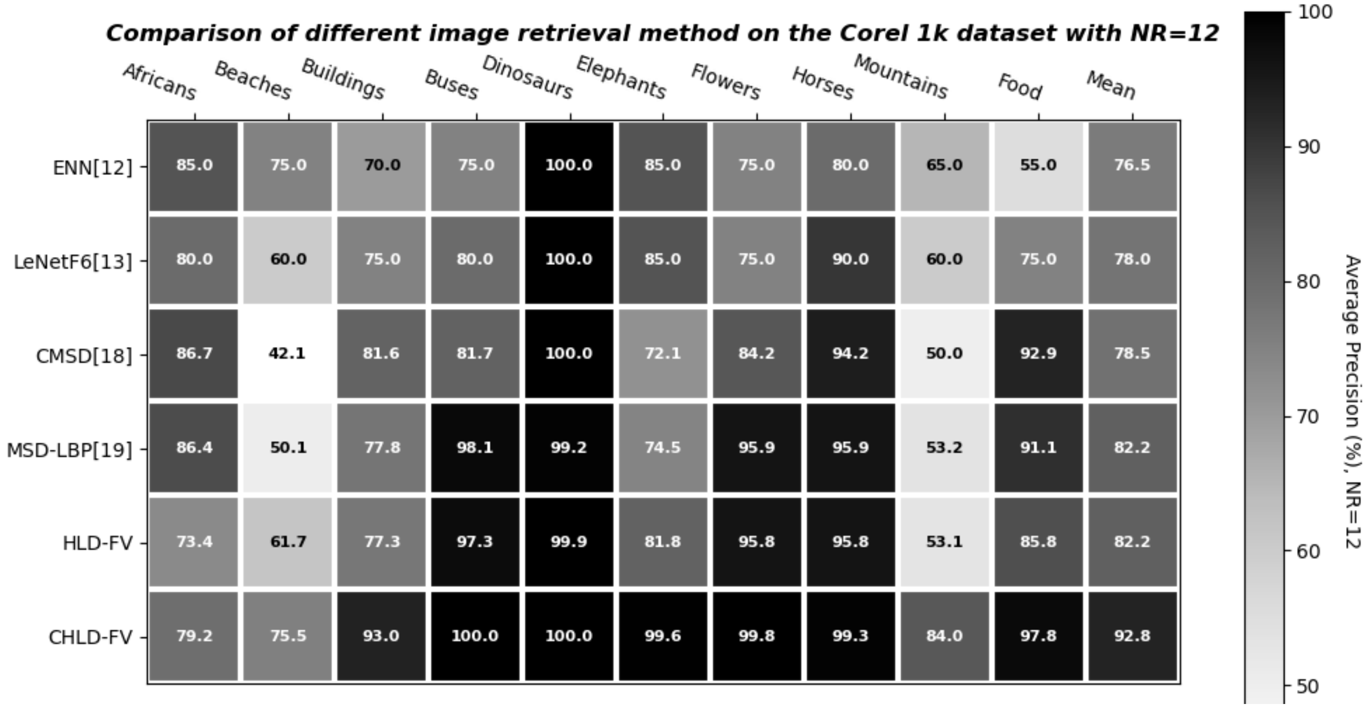


Figure 6. Heatmap of ARPs by using different image retrieval methods (NR=12).

Fig. 7 shows the ARPs with the NR=12 on Corel-1k dataset. Here, FIF-IRS [20] and SBF-SMI-CNN [14] are both methods based on image feature fusion, and the latter uses CNN features in combination. The HLD-FV process has better performance than FIF-IRS, and CHLD-FV has the best performance. The SBF-SMI-CNN showed significant performance in categories such as beaches and mountains than the proposed methods. This benefits from image sampling, scaling, integration, shape-based filtering, RGB coefficients, and spatial mapping with CNN capabilities.

In summary, Due to the neural network structure's limitations, the acquired features cannot achieve good results in all categories. But the proposed feature fusion method based on CNN and Fisher encoding have more advantages in overall performance.

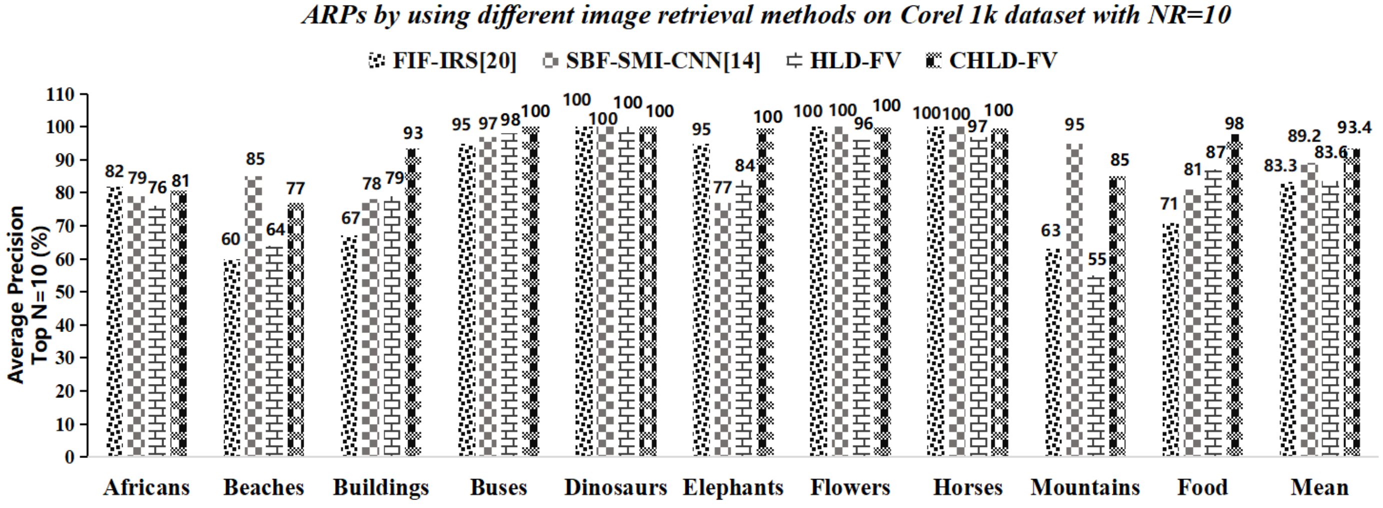


Figure 7. Bar plot of ARPs by using different image retrieval methods (NR=10).

IV. Conclusions

The proposed approach based on image features fusion and FV encoding has outstanding performance for CBIR task. The proposed low-level feature fusion-based method HLD-FV using the idea of image block. The CHLD-FV method combined the CNN features, which has improved much in recognizing beaches, buildings, elephants, mountains, and food. However, the proposed method still has difficulty retrieving categories such as Africans, beaches, and mountains. Besides, the scheme is relatively complex and time-consuming. Therefore, future work should be focused on scenario recognition and algorithm optimization.

References

[1] S. Sural, G. Qian, and S. Pramanik, “Segmentation and histogram generation using the HSV color space for image retrieval,” In Proceedings. International Conference on Image Processing. 2002. vol. 2, pp. II-II). IEEE.

[2] M. Subrahmanyam, Q.J. Wu, R.P. Maheshwari, and R. Balasubramanian, “Modified color motif co-occurrence matrix for image indexing and retrieval,” Computers & Electrical Engineering. 2013. 39(3), 762-774.

[3] T. Ahonen, A. Hadid, and M. Pietikäinen, “Face recognition with local binary patterns,” In European conference on computer vision. 2004. pp. 469-481. Springer, Berlin, Heidelberg.

[4] S. Murala, R. P. Maheshwari, and R. Balasubramanian, “Local tetra patterns: a new feature descriptor for content-based image retrieval,” IEEE transactions on image processing. 2012. 21(5), 2874-2886.

[5] S. Murala, R. P. Maheshwari, and R. Balasubramanian, “Directional local extrema patterns: a new descriptor for content based image retrieval,” International journal of multimedia information retrieval. 2012. 1(3), 191-203.

[6] S. Vetova, “Comparative analysis between two search algorithms using DTCWT for content-based image retrieval,” In 3rd International Conference on Circuits, Systems, Communications, Computers and Applications. 2014. pp. 113-120.

[7] L. Belhallouche, K. Belloulata, and K. Kpalma, “A new approach to region based image retrieval using shape adaptive discrete wavelet transform,” IJ Image, Graphics and Signal Processing. 2016. 1, 1-14.

[8] D. Zhang, and G. Lu, “Shape-based image retrieval using generic Fourier descriptor,” Signal Processing: Image Communication. 2002. 17(10), 825-848.

[9] K. T. Ahmed, A. Irtaza, and M. A. Iqbal, “Fusion of local and global features for effective image extraction,” Applied Intelligence. 2017. 47(2), 526-543.

[10] P. Poursistani, H. Nezamabadi-pour, R. A. Moghadam, and M. Saeed, “Image indexing and retrieval in JPEG compressed domain based on vector quantization,” Mathematical and Computer Modelling. 2013. 57(5-6), 1005-1017.

[11] J. M. Guo, and H. Prasetyo, “Content-based image retrieval using features extracted from halftoning-based block truncation coding,” IEEE Transactions on image processing. 2014. 24(3), 1010-1024.

[12] R. Ashraf, K. Bashir, A. Irtaza, and M. T. Mahmood, “Content based image retrieval using embedded neural networks with bandletized regions,” Entropy. 2015. 17(6), 3552-3580.

[13] H. Liu, B. Li, X. Lv, and Y. Huang, “Image retrieval using fused deep convolutional features,” Procedia Computer Science. 2017.107, 749-754.

[14] K. Kanwal, K. T. Ahmad, R. Khan, A. T. Abbasi, and J. Li, “Deep Learning Using Symmetry, FAST Scores, Shape-Based Filtering and Spatial Mapping Integrated with CNN for Large Scale Image Retrieval,” Symmetry. 2020. 12(4), 612.

[15] M. E. ElAlami, “A new matching strategy for content based image retrieval system,” Applied Soft Computing. 2014. 14, 407-418.

[16] J. X. Zhou, X. D. Liu, T. W. Xu, J. H. Gan, and W. Q. Liu, “A new fusion approach for content based image retrieval with color histogram and local directional pattern,” International Journal of Machine Learning and Cybernetics. 2018. 9(4), 677-689.

[17] K. T. Ahmed, S. Ummesafi, and A. Iqbal, “Content based image retrieval using image features information fusion,” Information Fusion. 2019. 51, 76-99.

[18] H. Dawood, M. H. Alkinani, A. Raza, H. Dawood, R. Mehboob, and S. Shabbir, “Correlated microstructure descriptor for image retrieval,” IEEE Access. 2019. 7, 55206-55228.

[19] D. Niu, X. Zhao, X. Lin, and C. Zhang, “A novel image retrieval method based on multi-features fusion,” Signal Processing: Image Communication. 2020. 87, 115911

[20] M. I. T. Bella, and A. Vasuki, “An efficient image retrieval framework using fused information feature,” Computers & Electrical Engineering. 2019. 75, 46-60.

[21] J. Z. Wang, J. Li, and G. Wiederhold, “SIMPLIcity: Semantics-sensitive integrated matching for picture libraries,” IEEE Transactions on pattern analysis and machine intelligence. 2001. 23(9), 947-963.

[22] G. Wimmer, A. Vécsei, M. Häfner, and A. Uhl, “Fisher encoding of convolutional neural network features for endoscopic image classification,” Journal of Medical Imaging. 2018. 5(3), 034504.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” Advances in neural information processing systems. 2012. 25, 1097-1105.

[24] D.Y. Jiang, and J.W. Kim, “Texture Image Retrieval Using DTCWT-SVD and Local Binary Pattern Features,” JIPS. 2017. 13(6), 1628-1639.