Combining HSV Color and rootSIFT for Image Retrieval

Ahmad Alzu'bi, Abbes Amira, Naeem Ramzan, and Tareq Jaber

The proposed model introduces an optimized image descriptor that combines color and local features for image retrieval. Color histograms in HSV space are extracted as global features, while root scale-invariant feature transform (rootSIFT) descriptors are densely extracted as local descriptors. The extracted features are fused and encoded by the visual locally aggregated features (VLAD) approach. The Corel image dataset is used for performance

Introduction

Content-based image retrieval (CBIR) [1] is a popular technique that has been widely applied to address the problems of traditional text-based image retrieval systems. CBIR is mainly based on the extraction process of low-level image features.

On one hand, local image descriptors describe local information using key points of some image parts, e.g. regions and corner points. The scale-invariant feature transform (SIFT) [2] is one of the most successful and commonly applied image descriptors. This feature is invariant to image scale and rotation and provides a robust matching across a range of fundamental image variations, e.g. noise addition, affine distortion, and change in illumination and 3D viewpoint. Some efficient variants of SIFT have been proposed such as dense SIFT (D-SIFT) [3] and rootSIFT [4].

On the other hand, color feature is one of the most extensive vision characteristic. In this work, HSV color space is used because the components of hue and saturation are closely related to the pattern of human visual perception. Consequently, combining local descriptors with global features is an essential part of the presented model [5].

Image Retrieval Framework

As shown in Fig. 1, the retrieval model is implemented as correlated blocks that represent images using the extracted descriptors as vectors. Firstly, D-SIFT descriptors are densely extracted from every image in the dataset at 7 scales by a factor $\sqrt{2}$ between successive scales, bin size of 8 pixels wide and a step of 4 pixels. Then, Hellinger kernel is applied to form rootSIFT descriptors and measure the similarity between SIFT descriptors, which yields superior performance in most cases without increasing processing or storage requirements. The 128-*D* rootSIFT descriptor is reduced to 100-*D* vector by the principal component analysis (PCA) projection.

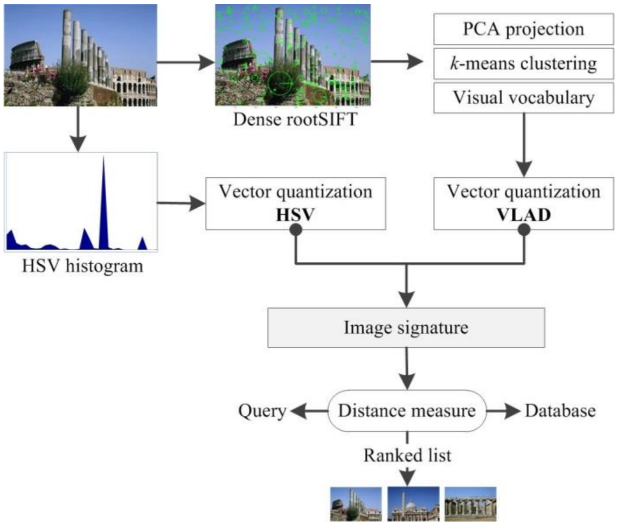


Fig. 6. The general framework of retrieval model.

The vector of locally aggregated descriptors (VLAD) [6] encoding is applied to quantize descriptors into vectors. It is a simplified non-probabilistic version of Fishers kernels, which is trained using *k*-means to accumulate the local descriptors and then normalized by L2 norm. A visual vocabulary of 256 clusters built by *k*-means approach is used in all experiments as a moderate size to keep balance between high discriminative image signature and low computation time.

The color value *C* is determined in the quantization by the equation $C = 8H + 2S + 2V$, where *C* is an integer between zero and 31.

Secondly, HSV histograms are extracted and quantized from the whole image as a global color feature. The *H* component is quantized into eight ranks non-uniformly and the *S* and *V* components are quantized into two ranks uniformly. Consequently, *H* and *S* are combined into a histogram of only 32 bins which is the representing color feature of each image.

Thirdly, the quantized HSV vector is combined with the VLAD vector to form the final signature of each encoded image in the dataset. Finally, the query image is matched with all dataset images using the Euclidean distance measure and the returned images are ranked based on the similarity scores obtained.

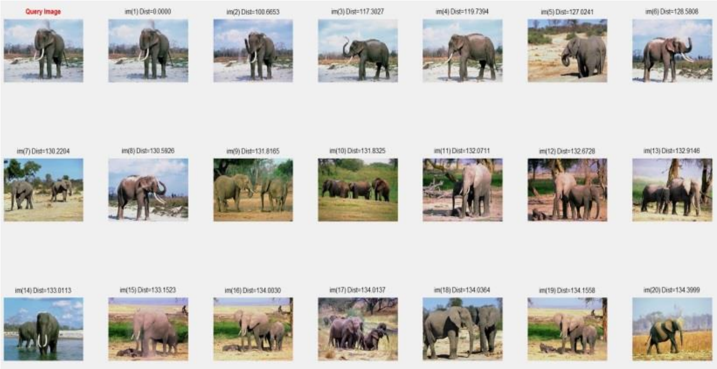


Fig. 2. Sample of query image and top 20 relevant images obtained.

Experiments

A subset of Corel image dataset[2] is used for experiments, which is widely utilized in computer vision applications. It consists of 1000 colorful images with 10 different semantic categories and each category contains 100 different images.

The Euclidean distance measure is computed between features transformed into *n*-dimensional vector of the query image and each vector in the image dataset in order to retrieve the relevant images. The mean average precision (mAP) is then obtained at different ranking positions to evaluate the retrieval accuracy.

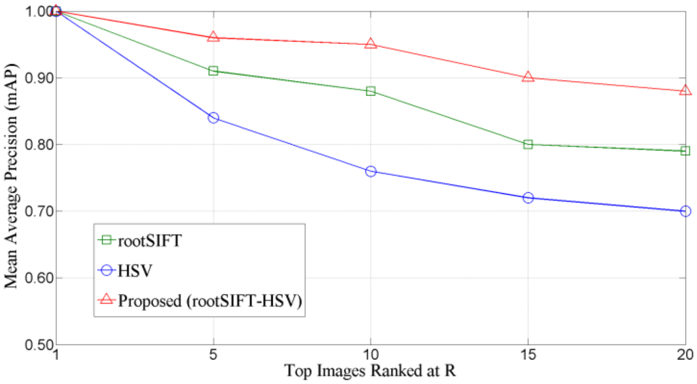


Fig. 3. The retrieval accuracy of three different methods rootSIFT, HSV, and the combined feature (i.e. Global and Local).

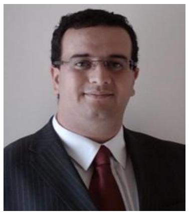
In every testing iteration, five query images are randomly selected from each dataset category and the process is repeated 10 times. Fig. 2 presents a sample of retrieval results. The retrieval accuracy (mAP) is reported over all queries at a range of ranking positions (*R*), i.e. top *R* at 1, 5, 10,

[2] http://wang.ist.psu.edu/docs/related/

15, and 20. Fig. 3 shows the best results of retrieval accuracy achieved using different image representations. It is clear that the proposed model outperforms both HSV and rootSIFT at all positions of rank images. At the top 10, 15, and 20 ranked images the retrieval accuracy improved by approximately 20% and 10% over HSV and rootSIFT, respectively.

The retrieval model achieves an efficient performance in terms of retrieval time and memory usage as follows: the actual memory size of each image vector is 100 KB, the average time (AT) elapsed to formulate the image vector is 0.445 seconds, the AT elapsed to search the whole image dataset and show the top 20 relevant images to the submitted query image is 0.795 seconds.

Ahmad Alzu'bi *is a PhD researcher at school of engineering and computing, University of the West of Scotland, Paisley, UK.*



Abbes Amira *is a Professor in visual communications, school of engineering and computing, University of the West of Scotland, Paisley, UK.*

Naeem Ramzan *is a Reader in visual communications, school of engineering and computing, University of the West of Scotland, Paisley, UK.*

Tareq Jaber *is an Assistant Professor, department of information systems, University of Jeddah, Saudi Arabia.*

References

[1] A. Alzu'bi, A. Amira, and N. Ramzan, "Semantic Content-based Image Retrieval: A Comprehensive Study," *Journal of Visual Communication and Image Representation*, vol. 30, pp. 20-54, 2015.

[2] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004.

[3] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," In *ACM Proceedings of the International Conference on Multimedia*, pp. 1469-1472, 2010.

[4] R. Arandjelovic and A. Zisserman, "Three things everyone should know to improve object retrieval," In *CVPR*, pp. 2911-2918, 2012.

[5] A. Alzu'bi, A. Amira, N. Ramzan, and T. Jaber, "Robust Fusion of Color and Local Descriptors for Image Retrieval and Classification," *In IEEE Proceeding of 22nd conference On Systems, Signals, and Image Proceesing (IWSSIP)*, 2015.

[6] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid, "Aggregating local image descriptors into compact codes," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 34, no. 9, pp. 1704-1716, 2012.