Optik 126 (2015) 1695–1699

Contents lists available at ScienceDirect

Optik

journal homepage: www.elsevier.de/ijleo

Content based image retrieval system using clustered scale invariant feature transforms

Gholam Ali Montazer [a,b,*,1], Davar Giveki [c,2]

[a] Information Technology Engineering Department, School of Engineering, Tarbiat Modares University, Tehran, Iran
[b] Visiting Lecturer; Iranian Research Institute for Information Science and Technology (IranDoc), Tehran, Iran
[c] Iranian Research Institute for Information Science and Technology (IranDoc), Tehran, Iran

ARTICLE INFO

Article history:
Received 23 March 2014
Accepted 5 May 2015

Keywords:
Content based image retrieval
Scale invariant feature transform (SIFT)
Cardinality matrix of cluster-sets (CMCS)
Resultant vector of clustered SIFT features (RVCSF)

ABSTRACT

The large amounts of image collections available from a variety of sources have posed increasing technical challenges to computer systems to store/transmit and index/manage the image data to make such collections easily accessible. To search and retrieve the expected images from the database a content-based image retrieval (CBIR) system is highly demanded. CBIR extracts features of a query image and try to match them with extracted features from images in the database. This paper introduces two novel methods as image descriptors. The basis of the proposed methods is built upon scale invariant feature transform (SIFT) algorithm. After extracting image features using SIFT, k-means clustering is applied on feature matrix extracted by SIFT, and then two new kinds of dimensionality reductions are applied to make SIFT features more efficient and realistic for image retrieval problem. Using the proposed strategies we cannot only take the advantage of SIFT features but also we can highly decrease the memory storage used by SIFT features. As well as in order to compare images we do not need to run the time-consuming matching algorithm of SIFT. Finally, proposed methods are compared with two popular methods namely, color auto-correlogram and wavelet transform. As a result, our proposed retrieval system is fast and accurate and it can efficiently manage large databases. Experimental results on two popular databases, Caltech 101 (with 9144 images) and Li database (with 2360) images, show the superiority and efficiency of the proposed methods.

1. Introduction

Research on multimedia systems and content-based image retrieval has been given tremendous importance during the last decade. The reason behind this is the fact that multimedia databases deal with text, audio, video and image data which could provide us with enormous amount of information and which has affected our life style for the better. Content-based image retrieval (CBIR) is a challenge of the access of multimedia databases simply because there still exists vast differences in the perception capacity between a human and a computer known as Semantic Gap.

Plenty of research works had been undertaken in the past decade to design efficient image retrieval techniques from the image or multimedia databases. Although large number of indexing and retrieval techniques have been developed, there is still no universally accepted feature extraction, indexing and retrieval techniques available. The most important factor that highly affects the performance of an image retrieval system, in terms of its speed and accuracy, is image descriptor. The image descriptor is responsible for extracting the content image as image features. In this paper we introduce a novel approach using scale invariant feature transform (SIFT) algorithm to design an image retrieval system. SIFT is an algorithm in computer vision to extract and describe local features in images [1]. SIFT algorithm has been successfully used in many computer vision and image processing tasks. Applications include object recognition, robotic mapping and navigation, image stitching, 3D modeling, gesture recognition, video tracking, individual identification of wildlife and match moving [2–8]. So, in this paper we decided to use SIFT as our image descriptor. Actually, in the case of image retrieval problem there are two important drawbacks with SIFT namely, time complexity and memory usage. In order to tackle these problems we propose a new kind of clustering method that highly decreases a dimension of the features extracted by SIFT while at the same time keeps the discriminative power of

* Corresponding author at: P.O. Box: 14115-179, Tehran, Iran. Tel.: +98 9123230540.
E-mail addresses: montazer@modares.ac.ir (G.A. Montazer), Giveki@students.irandoc.ac.ir (D. Giveki).
[1] Tel.: +98 9163652138.
[2] Tel.: +98 9127038683.

http://dx.doi.org/10.1016/j.ijleo.2015.05.002
0030-4026/

the image features. The rest of the paper is organized as follows. In Section 2, SIFT method and its drawbacks in image retrieval problem is briefly discussed. In Section 3 proposed approaches are described in details. Section 4 deals with experimental results and implementations. Finally, future work and conclusion are drawn in Section 5.

2. Scale invariant feature transform (SIFT)

For any object in an image, interesting points on the object can be extracted to provide a "feature description" of the object. This description, extracted from a training image, can then be used to identify the object when attempting to locate the object in a test image containing many other objects. To perform reliable recognition, it is important that the features extracted from the training image be extractable even under changes in image scale, noise and illumination. Such points usually lie on high-contrast regions of the image, such as object edges. Another important characteristic of these features is that the relative positions between them in the original scene should not change from one image to another. Similarly, features located in articulated or flexible objects would typically not work if any change in their internal geometry happens between two images in the set being processed. Scale invariant feature transform is a method for extracting distinctive invariant features from images that can be used to perform reliable matching between different views of an object or scene. The features are invariant to image scale and rotation, and are shown to provide robust matching across a substantial range of affine distortion, change in 3D viewpoint, addition of noise, and change in illumination. The features are highly distinctive, in the sense that a single feature can be correctly matched with high probability against a large database of features from many images. The major stages of computation used to generate the set of image features using SIFT method are shortly described in the following [1]:

2.1. Scale-space extrema extraction

The first phase of the computation seeks to identify potential interest points. It searches over all scales and image locations. The computation is accomplished by using a difference of Gaussian (DoG) function. The resulting interest points are invariant to scale and rotation, meaning that they are persistent across image scales and rotation.

2.2. Keypoint localization

For all interest points found in phase 1, a detailed model is created to determine location and scale. Keypoints are selected based on their stability. A stable keypoint is thus a keypoint resistant to image distortion.

2.3. Orientation assignment

One or more orientations are assigned to each keypoint location based on local image gradient directions. All future operations are performed on image data that has been transformed relative to the assigned orientation, scale, and location for each feature, thereby providing invariance to these transformations.

2.4. Keypoint descriptor

The local image gradients are measured at the selected scale in the region around each keypoint. These are transformed into a representation that allows for significant levels of local shape distortion and change in illumination. A keypoint descriptor is a 128-dimensional vector that describes a keypoint. The reason for this high dimension is that each keypoint descriptor contains a lot of information about the point it describes.

For image matching and recognition, SIFT features are first extracted from a set of reference images and stored in a database. A new image is matched by individually comparing each feature from the new image to this previous database and finding candidate matching features based on Euclidean distance of their feature vectors. The keypoint descriptors are highly distinctive, which allows a single feature to find its correct match with good probability in a large database of features. However, in a cluttered image, many features from the background will not have any correct match in the database, giving rise to many false matches in addition to the correct ones. The correct matches can be filtered from the full set of matches by identifying subsets of keypoints that agree on the object and its location, scale, and orientation in the new image. The probability that several features will agree on these parameters by chance is much lower than the probability that any individual feature match will be in error. The determination of these consistent clusters can be performed rapidly by using an efficient hash table implementation of the generalized Hough transform. Each cluster of 3 or more features that agree on an object and its pose is then subject to further detailed verification. First, a least-squared estimate is made for an affine approximation to the object pose. Any other image features consistent with this pose are identified, and outliers are discarded. Finally, a detailed computation is made of the probability that a particular set of features indicates the presence of an object, given the accuracy of fit and number of probable false matches. Object matches that pass all these tests can be identified as correct with high confidence.

In order to get useful insights toward the image retrieval problem using SIFT, one can refer to [10]. In this paper Bakken did some useful experiments for evaluating the power of SIFT in CBIR. For instance he measured the running time of keypoint generation and keypoint matching. Also, he examined the results of applying SIFT on different kind of images like images with high rate of variation or simple images with flat area of just one color. In the end he concluded that SIFT is not suitable for CBIR problem. Two of the most important drawbacks of SIFT that prevent it from having a good performance in CBIR are its memory usage and running time of the matching algorithm. Here we address these two drawbacks in such a way that SIFT can be efficiently used for retrieving images from large databases in real time performance while the proposed approaches can reach high accuracy.

3. Proposed approaches

In this section we propose two approaches for using SIFT features in CBIR. SIFT features are local features that are highly distinctive and suitable for image matching. However, in case of CBIR, a problem arises from the large number of keypoints extracted by SIFT algorithm. Storing these features and then matching them with images in the large databases is really impractical [10]. Therefore, it seems that new ways for decreasing the size of SIFT descriptor is needed. Data clustering is one way for achieving this goal.

Here we use clustering for decreasing the size of SIFT feature descriptor. By increasing the number of keypoints in images, SIFT feature descriptor gets too big. Suppose that an image A contains a set of k keypoints $P = \{p_1, p_2, \ldots, p_k\}$ in which feature vector of each point p_i is $v_i = \{v_{i1}, v_{i2}, v_{i3}, \ldots, v_{i128}\}$. Now, we define the following concepts.

3.1. Cardinality matrix of cluster-sets (CMCS)

In our proposed methods k-means clustering is used to decrease the dimension of feature matrix. To this end, in our first method,

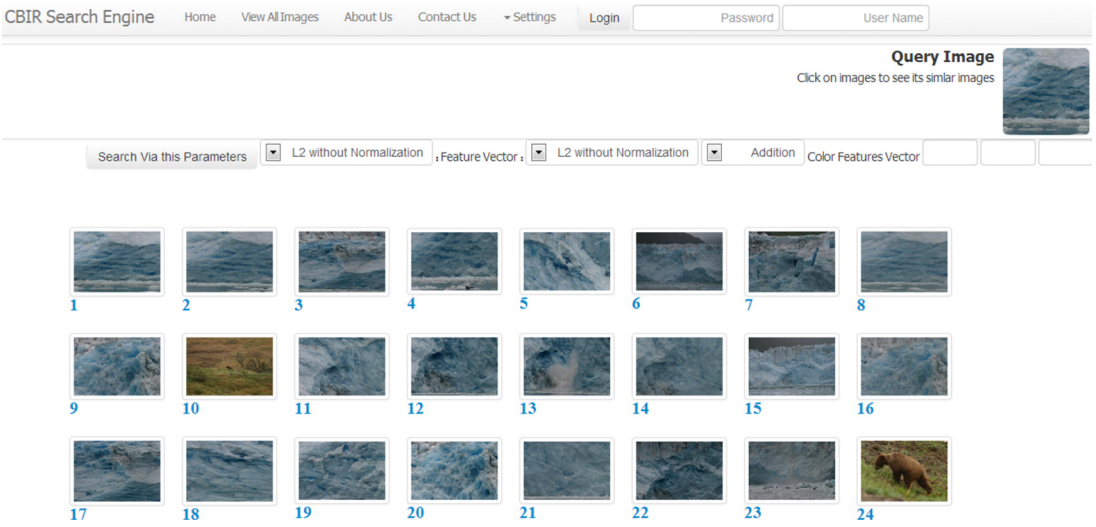

Fig. 1. The retrieval results of the mountain image from Li database [9].

feature vector of each keypoint is divided into 16 parts so that each part includes eight keypoints. Clustering these parts yields a new feature vector. So, the feature vector v_j is divided into 16 parts as following:

$$v_j = \{v_{j1}, v_{j2}, v_{j3}, \ldots, v_{j128}\} \rightarrow \text{division of Feature vector}$$
$$v'_{j1} = \{v_{j1}, v_{j2}, v_{j3}, \ldots, v_{j8}\}, \ldots, v'_{j16} = \{v_{j121}, v_{j122}, v_{j123}, \ldots, v_{j128}\} \quad (1)$$

We construct a k-means clustering named K with eight clusters so as $\mathbf{K}(\mathbf{v'_{jl}}) = \mathbf{n}$ if and only if $\mathbf{v'_{jl}}$ belongs to the nth cluster. Therefore, each $\mathbf{v'_{jl}}$ belongs to one cluster and feature vector $\mathbf{v_j}$ can be represented in a more compact way using 8 clusters.

However, for each image, like A, the number of keypoints vary which causes a big for comparing images. Therefore, we have to aggregate extracted features from points. In order to do that, we define the following Cluster-Set:

$$M'_{lj} = \{v'_{lj} | K(v'_{lj}) = i, v_l \in P\} \quad (2)$$

where M'_{ij} is a set of parts belongs to the ith cluster. This set helps to construct a global matrix G as following:

$$G = \begin{bmatrix} |M'_{11}| & \cdots & |M'_{1,16}| \\ \vdots & \ddots & \vdots \\ |M'_{81}| & \cdots & |M'_{8,16}| \end{bmatrix}_{8\times 16} \quad (3)$$

where $|M'_{ij}|$ is cardinality of the set $|M'_{ij}|$. Fig. 1 shows the experimental results of the first proposed method.

3.2. Resultant vector of clustered SIFT features (RVCSF)

In the second method, we compress local feature vectors based on the concept of SIFT features. To do so, resultant of histogram bins in the same directions is computed. Hence, a new compact feature vector $\bar{v}_j$ is built as following:

$$v_j = \{v_{j1}, v_{j2}, v_{j3}, \ldots, v_{j128}\} \rightarrow \bar{v}_j$$
$$= \left\{ \sum_{l=0}^{15} v_{j(l*8+1)}, \sum_{l=0}^{15} v_{j(l*8+2)}, \ldots, \sum_{l=0}^{15} v_{j(l*8+8)} \right\} \quad (4)$$

As a result of this compression, the local features tend to be more global while they are highly distinctive yet. This improvement results in better performance in the scope of CBIR. To make more improvement we decided to cluster new compact feature vector using k-means, K, with 16 clusters. Hence, $K(\bar{v}_j) = n$ if and only if $\bar{v}_j$ belongs to the nth cluster. By using clustering K, each feature descriptor can be referred as a cluster. Therefore, with aggregating clustering results, we build the global feature vector G_2 as following:

$$G_5 = \left[\left|\bar{M}_1\right|, \left|\bar{M}_2\right|, \ldots, \left|\bar{M}_{16}\right| \right] \quad (5)$$

where $\bar{M}_k = \{p | K(\bar{v}_j) = k, \quad p \in P\}, \quad k = 1, 2, \ldots, 16.$

In this method we used 16 clusters. Experimental results show that by increasing the number of clusters the efficiency of the proposed method is highly increased. Fig. 2 shows the result of implementing the second method with 16 clusters. We should mention that the best results were achieved when we used the second method with 16 clusters and dividing images into 16 parts.

We also noticed that the locations of the keypoints in images play an important role in the SIFT algorithm. Therefore, to take this point into account, we decided to divide images into some parts and compute their G2 feature. In this case the distance between two images is sum of the distances between corresponding parts.

4. Experimental results and comparisons

In this section we make some comparisons between our proposed methods with some popular methods like color auto-correlogram [12] and wavelet transform [13]. As it can be seen from Fig. 3 the performance of the first and second proposed methods are better than color auto-correlogram and Wavelet.

For comparing retrieval results on Li database, we used the name of the images in the database which are created by the database collector and the fact that similar images are given sequential numbers so the sum of Euclidean distances between the name of the query image and retrieved images can be a realistic criteria for comparing the methods.

As it can be seen from Fig. 3 the most promising results were achieved using proposed approaches.

The performance of the proposed scheme is also tested on the benchmark Caltech 101 which contains 9146 images grouped into 101 distinct object classes and a background class. The number of images in an image category ranges from 31 to 800. Most images in the Caltech 101 database have a centered object, which covers

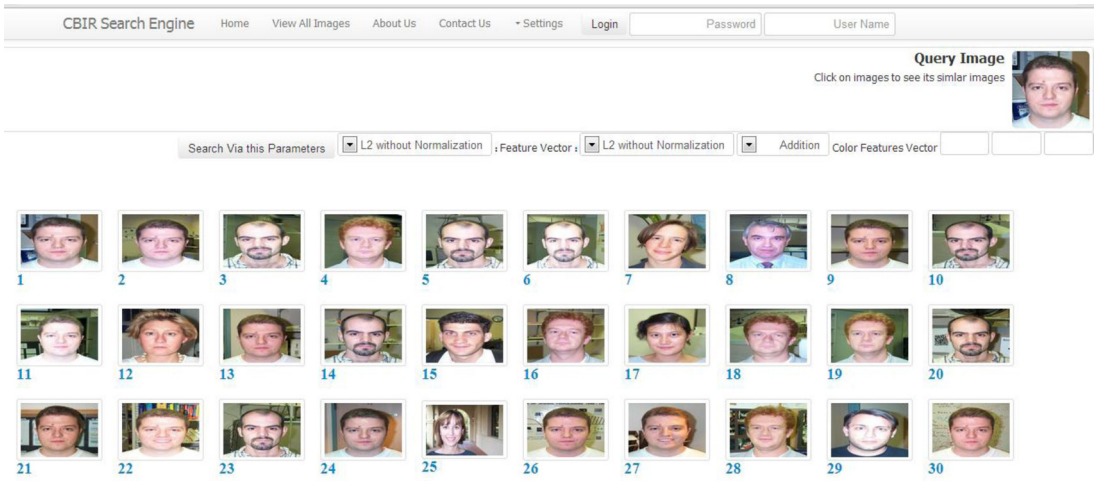

Fig. 2. Retrieval results of the face image from Caltech 101 database [11].

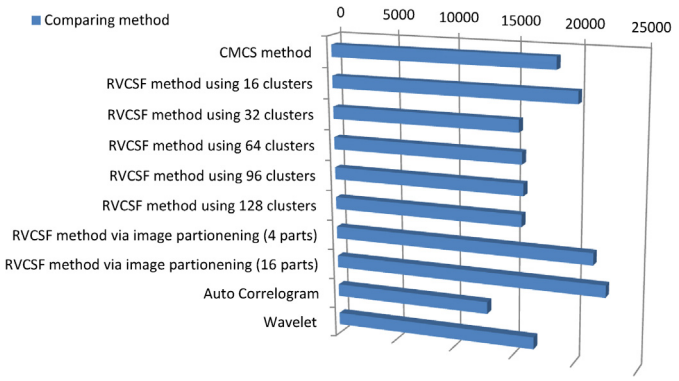

Fig. 3. The results of comparison between our proposed methods, auto-correlogram and wavelet transform.

the entire image. Hence, the problem of geometrical variance of object representation is minimal as in the case of ideal object based image retrieval. The experiments were carried out by using a set of ten different, randomly sampled images from each image category as query images. Thus, the recall experiments were performed for 1010 query images. All experiments are performed on grayscale images. The effectiveness of our scheme is demonstrated by comparing the results with the traditional Bag of words (BoW) approach [14] as well as the Spatial pyramid matching (SPM) based scheme [15].

The mean Average Precision (mAP) values for the three schemes are calculated over all the query images. Average precision of a single query is defined as the mean of precision scores after each relevant image retrieved. mAP is defined as the mean of the individual average precision scores for all query images. Our proposed method achieved 92.69% mAP while BoW method and SPM achieved 82.00% and 88.80% mAP, respectively.

Therefore, the proposed scheme produce an improvement of 10.69% over the traditional BoW scheme, while the SPM based scheme produces an improvement of 6.8%.

In general, image sets, which had distinct visual patterns and less background occlusion responded well to the proposed scheme like the 'camera' image set in the Caltech 101 database. Image sets where the position or orientation of the central object of interest varies significantly within the images in the set, are seen to show relatively poor results when used as query images for the proposed scheme. This could be attributed to the fact that the proposed scheme does not focus on geometrically invariant representation of the object structure.

5. Conclusion and future work

This paper proposes two new approaches for content-based image retrieval using Scale invariant feature transform method. The main goals of these two approaches are tackling two important drawbacks of SIFT method namely, its memory usage and its matching time, that prevents it to be used as a reliable method for CBIR problem. In the first approach using *k*-means clustering, the feature matrix extracted by SIFT method is clustered to 16 clusters so that instead of having a huge matrix with dimension of $k \times 128$, which k is the number of keypoints, we will have a reduced matrix of size 8×16. In the second approach using the fact that each keypoint is described by eight dominant directions, the feature matrix is reduced to a vector of size 1×18. In both methods there is no loss of discriminative power of SIFT while the size of feature matrix is highly reduced. The proposed approach shows high performance when facing with object images namely the images with an object in them. This fact is because of the concept of SIFT. However, for more complex images like scene images or images with cluttered and occluded background the performance of SIFT needs more improvement.

Acknowledgements

The authors are grateful to the anonymous reviewers for the insightful comments and constructive suggestions. Part of this research was funded by the Iranian Research Institute for Information Science and Technology (IranDoc) (No. TMU92-03-44).

References

[1] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[2] J. Beis, D.G. Lowe, Shape indexing using approximate nearest-neighbour search in high-dimensional spaces, in: Conference on Computer Vision and Pattern Recognition, Puerto Rico, 1997, pp. 1000–1006.
[3] S. Se, D.G. Lowe, J. Little, Vision-based mobile robot localization and mapping using scale-invariant features, in: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), 2001.
[4] M. Brown, D.G. Lowe, Recognising panoramas Proceedings of the Ninth IEEE International Conference on Computer Vision, 2, 2003, pp. 1218–1225.
[5] I. Gordon, D.G. Lowe, What and where: 3D object recognition with accurate pose, in: Toward Category-Level Object Recognition, Springer-Verlag, 2006, pp. 67–82.

[6] G.T. Flitton, T. Breckon, Object Recognition using 3D SIFT in complex CT volumes, in: Proceedings of the British Machine Vision Conference, 2010, pp. 11.1–12.
[7] M. Toews, W.M. Wells, D.L. Collins, T.T. Arbel, Feature-based morphometry: discovering group-related anatomical patterns, Neuroimage 49 (3) (2010) 2318–2327.
[8] G. Flitton, T.P. Breckon, N. Megherbi, A comparison of 3D interest point descriptors with application to airport baggage object detection in complex CT imagery, Pattern Recognit. 46 (9) (2013) 2420–2436.
[9] Li Database. http://sites.stat.psu.edu/~jiali/index.download.html
[10] T. Bakken, An evaluation of the SIFT algorithm for CBIR, in: Telenor R & I Research Note, 2007.
[11] F.F. Li, R. Fergus, P. Perona, Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories, in: IEEE, CVPR, Workshop on Generative-Model Based Vision, 2004.
[12] B.S. Manjunath, J.R. Ohm, V.V. Vasudevan, A.A. Yamada, Color and Texture Descriptors, IEEE Trans. Circuits Syst. Video Technol. 11 (6) (2001) 703–715.
[13] M. Kokareh, P.K. Biswas, B.N. Chatterji, Texture image retrieval using new rotated complex wavelet filters, IEEE Trans. Syst. Man Cybern. B: Cybern. 35 (6) (2005) 1168–1178.
[14] J. Sivic, A. Zisserman, Video Google: a text retrieval approach to object matching in videos, in: Proceedings IEEE International Conference on Computer Vision (ICCV), 2003, pp. 1470–1477.
[15] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, in: Proceedings of the International Conference Computer Vision and Pattern Recognition (CVPR), volume 2, 2006, pp. 2169–2178.