Improved Retrieval of Medical Images Using K-means Clustering Algorithm

Farsad. Zamani Boroujeni [1], Seyed Masoud. Khademi [2], Simindokht. Jahangard [3]
1- Department of Computer Engineering, Islamic Azad University, Isfahan (Khorasgan) Branch, Isfahan, Iran
f.zamani@khuisf.ac.ir
2- Department of Computer Engineering, Islamic Azad University, Isfahan (Khorasgan) Branch, Isfahan, Iran
khademi_136@yahoo.com
3- Department of Computer Engineering, Islamic Azad University, Isfahan (Khorasgan) Branch, Isfahan, Iran
s_jahangard@aut.ac.ir

Received March 2016 Revised April 2016 Accepted May 2016

ABSTRACT:
Due to the rapid and increasing progress of medical equipment and medical imaging machines, a large number of digital images are produced in therapeutic centers and stored in the large databases. Retrieving a set of images which are most similar to a query image is a major challenge in this field. A popular way to address this problems is to apply image retrieval based on a bag of visual words. A popular technique for producing visual words is to use K-means clustering algorithm. However, the effectiveness of K-means algorithm highly depends on the initial selection of cluster centroids which are selected randomly in the original K-means algorithm. Therefore, the image retrieval task is affected by poor clustering solutions due to random selection of cluster centroids. The goal of this paper is to overcome this problem and improving the accuracy of content based image retrieval from medical image databases using the optimal selection of initial points in K-means clustering algorithm. In the proposed method, after extracting SIFT features from the color images, the optimal centroid points are selected based on different ranges, weights and means of samples and visual words are created based on the selected centroid. The result of experiments on applying three different algorithms for selecting initial cluster centroids show that selecting the optimized initial points results in producing more discriminative visual words and provide favorable accuracy for medical image retrieval systems.

KEYWORDS: Image retrieval, Bag-of-words, K-means clustering, Initial centroids, SIFT descriptors.

1. INTRODUCTION

With the development of technology and prevalence of digital images in the medical field, a large number of medical images is produced in hospitals and medical research centers including X-ray, MRI, and CT scan. Image retrieval has become one of the most active research area in the field of computer vision especially in the applications that require browsing the large multimedia repositories. In the past decade, content-based medical image retrieval have had a huge impact on the diagnosis and treatment of diseases.

Content-based image retrieval (CBIR) is the application of computer vision techniques to search for images that have the most visual similarity to a query image in a large database. The interest in CBIR has grown because of its large range of possible applications for efficient image retrieval. It made CBIR a prosperous field of study in the past few years. The main objective of the content-based image retrieval is extracting low level features for grasping semantic similarity to find those images which have the most similarity in visual content. Nowadays, the bag-of-features (BOF) is a common method of interest in the large-scale image retrieval systems which is an extension of bag-of-words in text mining [2]. It has proved to be an efficient method in both image retrieval and classification. In BOF model each image is described as a collection of local features. The BOF model has four main phases that must be taken into consideration: 1) feature extraction 2) constructing the visual words which is the phase required to create a visual dictionary by using the clustering algorithms and defining the cluster centers as visual words 3) assigning the collected feature examples to the clusters by calculating the similarity between the examples extracted from a given image and the visual word defined in the previous step 4) retrieving and classifying the query image on the basis of a histogram matching technique [1]. In the most of the recent studies, the K-means clustering algorithm is commonly used as a major component to create the visual words, i.e. the cluster centers [1].

The K-means algorithm proceeds by alternating between the two steps: In the first step, k data points are randomly selected as initial cluster centers. Then, the remaining tuples in the dataset are allocated to the

clusters based on their (Euclidean) distance to the cluster centers. In the second step, after allocating the tuples to the cluster centroids, a new center is calculated for each cluster are by averaging the vectors of tuples included in the cluster. It continues until no data point is changed.

However, the K-means algorithm is significantly sensitive to the selection of cluster centers. Since the initial centers are selected randomly, the K-means algorithm does not necessarily find the optimal clusters. Therefore, the visual words does not have sufficient accuracy to describe the image features which affects the performance of the retrieval algorithm. In addition, various performances results is achieved for the same input data [4]-[5]. In this paper, to gain optimal visual words, three methods have been involved for selecting the optimal initial cluster centers providing more accurate retrieval of the medical images.

The rest of this paper is organized as follows: In the second section, related works are presented. In the third section, the proposed method to retrieve medical images by using optimal visual words are expressed. The experimental results and their discussion are shown in section four and section five provides conclusions and future works.

2. RELATED WORKS

Content-based image retrieval system (CBIR) has been raised for the first time in 1992, when Kato used it for automatically retrieving from a database of images based on color and shape [13]. Since then, the term has been used to describe the process of retrieving desired images from a large collection of images based on image features. However, recently, content-based image retrieval systems are widely used in applications such as digital library, facial recognition and fingerprints, online shopping, trademark search, and online search. But a small number of these systems have been designed and implemented in certain medical applications.

Caicedo et al. (2010) designed a recovery system based on multiple features that combine the features of the tissue and the SIFT, DCM descriptors used for histogram features. The SVM classifier has been employed to recover the image from the likely proximity of the input image to the appropriate category. They found that their test results in SIFT is less efficient than DCT in classification [6]. In 2011, Wang et al. provided a method of retrieving medical images based on the BOF method in which each image is represented by a set of local features such as split image and important point using SIFT descriptor. They employed multiple assignment of a local descriptor which make neighboring visual words by using K-means clustering and simultaneously assign weights to visual words to increase the performance and accuracy of the classification. Finally, in order to extract the similarities and to measure the distance, sub histogram similarity and the sub distance between visual words are used [1].

In 2013, Ravindran and shakila presented a CBIR method in which visual content of images was represented by bag of words (BOF). In their study, two main tasks were used to explore the visual features: 1) finding the correlation between visual patterns and high-level concepts that means finding the lowest redundancy and choose the most relevant features and clustering analysis 2) performing automatic annotation image. This method was tested on histology images and the accuracy of the method has been about 80% [7].

In 2014, Vanegas et al. employed the combination of unsupervised learning feature (ULF) and the classic descriptors such as SURF, SIFT, DCT. They used UFL to learn global features like color, scale and rotate automatically. Their test results showed excellent performance not only for recovery features but also for the performance which surpassed the standard descriptors [8].

Abdonazeer and Kumar [10] calculated the difference between the maximum and minimum value for each column of the data set. After defining numerical range of each column, the highest and maximum range is identified among the columns and the tuples are sorted based on selected column in ascending order. The data is divided into k equal parts and the average of each part is considered as initial centers of clusters.

S. Mahmud et al. [14] presented a technique that calculates the average value for each point by multiplying each component to a given weight. Then, it sorts the data according to average score and divides it into k sub-sets and calculates their average. The data point that is closer to the average value is considered as an initial cluster centers. Ran Vijay et al, [3] presented a method for selecting the initial centroids on a given frequency. In their approach, each attribute of the dataset is assigned a minimum and maximum value and its range is calculated by (1). Then, it divides the data points into k groups with wide range and calculates the frequency for data points on each group. The initial cluster centers are selected by identifying the high frequency data points.

$$(MAX + MIN) / K \quad (1)$$

Goyal and Kumar [2] provided a method for selecting the initial centers which is based on the distance from the origin. In their method the distance of all data points, i.e. Euclidean distance, are calculated from the origin, zero point, and then the data is sorted according to this distance. Then sorted data are divided into k categories and their averages are considered as initial centers.

In comparison with basis K-means clustering, Kumar Abdolnzeer and kumar method perform better in complexity and accuracy of their method which is suited for low values of k. In their method, choosing a large number for k, e.g. more than 10, results in producing empty clusters.

Sohrab's method reduces the number of repetitions in K-means algorithm which decreases the computational complexity [14]. Their experiments showed that, by increasing the value of k, the accuracy of their method is decreased. Ran Vijay and colleagues [3] suggested an efficient method using K-means algorithm with randomly selected centers. But, the main drawback of their method is that by increasing the number of dimensions, the computational complexity and time are increased exponentially (k * k * d). They found that appropriate results can be achieved in when k <= 10. Kumar and Goyal method for random selection enjoys high efficiency but it remains challenging when k is large enough that has negative effect on robustness.

In the previous works discussed above, the K-means algorithm has been used to construct bags of visual words without addressing the problem of random selection of initial cluster centers. In our study, we propose to determine the optimal initial centers of clusters in the first step of K-means clustering algorithm and investigate its effects on the performance of the retrieval algorithm.

3. THE PROPOSED APPROACH

In this section, we introduce the proposed method in more detail. The flowchart presented in fig. 1 shows the overall procedure and its main stages which are described in the following sections.

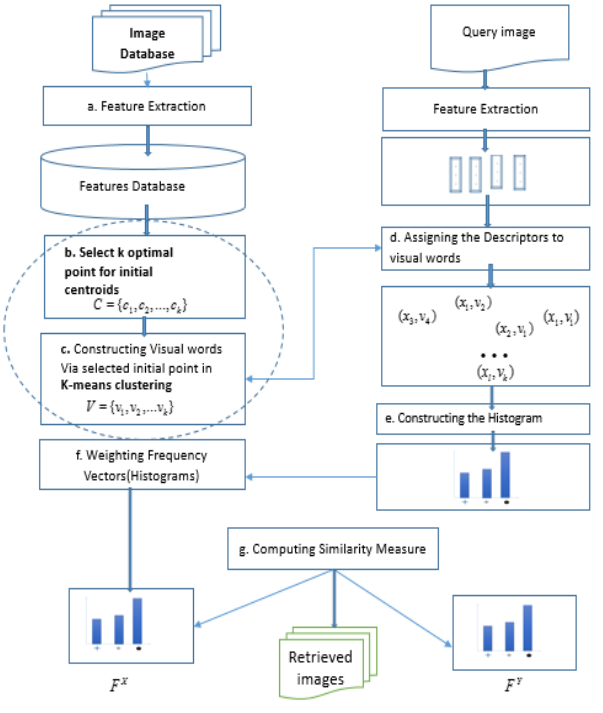

Fig. 1. The flowchart of the proposed method. The circle indicates the steps of interest

3.1. Feature Extraction

Our method provides extracting features which considers the input image as a series of small patches. Then, local features or interesting points from the patches are extracted using SIFT descriptor suggested in [10].

3.2. Claculating the visual words

The next step is creating the visual words by using K-means clustering method. The visual words refer to the feature vectors calculated as the representative of a cluster. In order to perform a deterministic selection of initial centroids,

In our study, three main approaches are applied to select the initial centroids. The inputs of the algorithms include k which is a constant value as the number of centers and a n by m matrix D where n denotes the number of rows or data points that are indicated by (d_1, d_2, d_3, ..., d_n) and m refers to the number of columns or features that are denoted by (x_1 , x_2 , x_3 , ..., x_m).

The first approach is based on Nazeer et al, methods [10] that consists of five steps. Firstly, the maximum and minimum of data points are defined for each column and the difference between the maximum and the minimum is calculated for each column. Then, the column that has the highest range will be detected. Next, the set of data points for selected column are sorted, as non-additive. In the fourth step, the sorted set of data points are divided into k classes where k is the number of clusters. Finally, the average weight of each class of data points calculated separately and the obtained averages are considered as initial centers indicated by (c_1, c_2, c_3... c_k).

The second approach follows the Goyal's method [11]. At the first step, the distance from the origin, zero point is origin, is calculated for each data point in the set of data points D. In the second step, the distance that obtained in the previous stage sorted and according to the sorted distance, data set D can be arranged. The sorted collection of data points are divided into k classes, k is the number of clusters, in the third step. At the fourth step, the average of each category is obtained which considered as an initial centers.

The third approach employed a strategy called weighted score that adopted from the method suggested by Mostafazir et al. [12]. The procedure of obtaining initial of stages is described below:

In the first step, the maximum value of each column is found, then for each row, the value of each attribute is divided by the maximum value and the resulting values are added together. The result is considered as the score of weight of this rows. The weight is calculated by:

$$(WS)T_n = \sum_{j=1}^{m} \frac{x_j}{x_{j(max)}} \quad (2)$$

Where x_j is the value of column j in the current row, $x_j(max)$ is the maximum value of column j and m is the

number of columns in the dataset. In the second step, all of the data points are sorted in ascending order according to their weight values calculated in the first step. In the third step, the sorted data is divided into equal k groups, where k is the number of clusters. In the next step the average weight of the data points in each group is calculated separately. Finally, the data point that has closer weight to the average that calculated in the previous step is found and considered as the optimal cluster centers.

The most important challenge in these approaches is that, by increasing the value of k, the number of the empty clusters will be increased. In order to address this problem, the nearest point to the average of the data is considered as initial centers. After optimal selection of initial centers, the centers are considered as visual words.

3.3. Assigning features to visual words

In this section we explain how features are assigned to visual words. Let x_l denotes a feature vector that is assigned to the nearest word in the dictionary V_k. So the image X is represented by the local features $\{(x_l, a_l)\}$. The a_l is calculated as follows:

$$a_l = argmin_k (D (v_k , x_l)) \quad (3)$$

Where $D(v_k, x_l)$ indicates the distance between the word V_k and feature vector x_l . Then, after normalization events of visual words with L_1 norm frequency vector image X displayed as $HX = [h_1^x \ h_2^x . . . h_k^x]$ [1]. In addition to the strategy of nearest neighbor (NN) employed in this research, other techniques such as soft assign strategy [6] and multiple assign strategy [1] are also studied.

3.4. Weighting

Image retrieval method based on BOF is similar to the retrieval of the text. Weighting the textual words has a significant impact on data recovery. Weighting to a word provides information about the relevance of the word (visual word) in the document (picture). Weighting method checks the distribution and frequency of words in the document (image) and determines a numerical weight for the word. This weight means grade and credibility of a word in different documents. Also the weights are dependent on the frequency of occurrence of a word in the document and the frequency of occurrence of a word in a particular document.

In this paper we used TF * IDF weighting method. It is as follows:

$$W_{i,j} = TF_{i,j} \frac{log\ N}{n_i} \quad (4)$$

$TF_{i,j}$ Frequency visual word i in the image j.

N: The total number of documents.

n_i: The number of documents that contain the word i.

3.5. Measure of similarity

To measure of similarity between the histogram of query image and database images usually HIK and soft L1 and JSD [1] is suggested. Here we provide sharing core measurement that have excellent results in image retrieval and is calculated as follows:

$$s_k^{HIK}(H^X, H^Y) = min(h_K^X, h_K^Y) \quad (5)$$

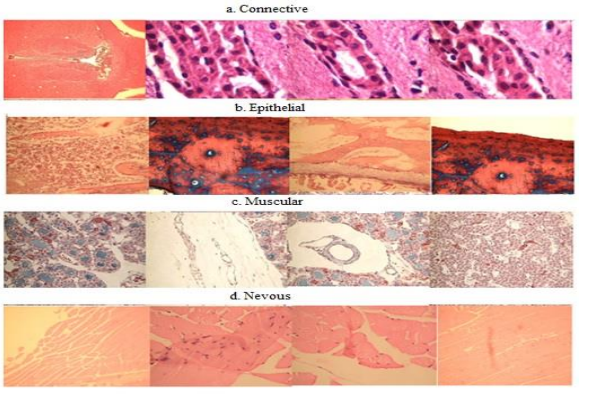

Fig. 2. Histology Images

4. THE EXPERIMENTS

In order to evaluate the image retrieval performance, we used a set of standard histopathology images. This collection consists of images acquired from various body organs that are representative of the four basic tissues. This dataset contains 2828 images that are labeled in four different categories shown in fig. 2. The main components of this data set are:

-484 connective tissue images.

-804 mucosal tissues images.

-514 muscle tissue images.

-1026 nerve tissue images.

In our experiments images were divided into training and testing categories. 80 percent of total images were randomly selected for training and the remaining images were assigned for testing. To test the precision of the retrieval, we selected 5 images randomly from each category and evaluated them. Finally, the average of the collected results were considered as the final result.

In this research, we conducted three experiments to evaluate the performance of the proposed method. In each experiments all methods of visual words including Random, Nazeer, Rahman,and Goyal were used. Each time the methods of assigning tuples to clusters have been changed including NN assign, soft assign (Gemert) and multiple assign (Multiple) that were used in first, second and third experiment respectively. In our first experiment, the tuples were assigned by soft method

(Gemert), Rahman and Random methods had highest and lowest precision, respectively (Fig. 3). In the second experiment, Nazeer had maximum accuracy and Goyal had the poor accuracy (Fig. 4). In the third experiment, the multiple assignment method, the method of Nazeer, had the most accuracy and Goyal had the lowest accuracy (Fig. 5). Consequently, it can be seen in experiments that used non-random selection centers have higher accuracy than random methods. Assigning features also effect on retrieved accuracy.

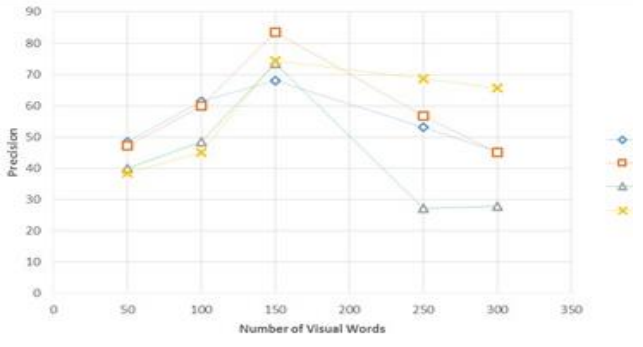

Fig. 3. Result of NN assignment for methods with random selection of initial points, [12], [11] and [10].

5. CONCLUSION

Image retrieval based on bag-of-visual words approach is one of the state-of-the-art methods for image retrieval that uses K-means clustering to create visual words; but the basis of this clustering algorithm is selecting initial cluster centers randomly. Therefore, each time the clustering algorithm is applied on same data, different clustering results will be obtained leading to poor performance results.

In our new method, we performed retrieval of medical images by estimating optimal selection of initial centers in K-means clustering which employed three different methods to select the initial cluster centers and also three methods to assign data points to the visual words. As we have shown non-random methods of selection of initial centers have superior performance in compare to random methods.

Fig. 4. Result of Gemert Assignment for methods of Random, [12],[11] and [10].

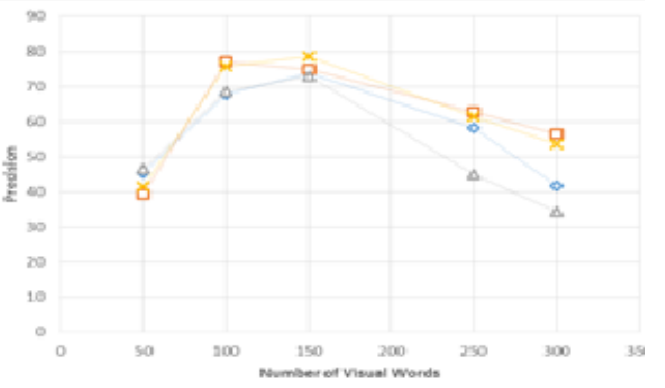

Fig. 5. Result of Multiple Assignment for methods of Random, [12],[11] and [10].

REFERENCES

[1] X. Wang, M. Yang, T. Cour , Sh. Zhu, K. Yu and T. X. Han, **"Contextual weighting for vocabulary tree based image retrieval,"** *Proceedings of the IEEE International Conference on Computer Vision*, pp. 209-216, Nov. 2011.

[2] J. Yu, Z. Qin, T. Wan, and Z. Xi, **"Feature integration analysis of bag-of-features model for image retrieval,"** *Neurocomputing*, vol. 120, pp. 355-364, Mar. 2013.

[3] U. Avni, H. Greenspan, M. Sharon, E. Konen, and J. Goldberger, **"X-ray image categorization and retrieval using patch-based visualwords representation,"** *Biomedical Imaging: From Nano to Macro. ISBI'09. IEEE International Symposium*, pp. 350-353, Agu. 2009.

[4] M. Erisoglu , N. Calis , and S. Sakallioglu, **"A new algorithm for initial cluster centers in K-means algorithm, "**, *Pattern Recognition Letters*, pp. 1701-1705, 2011.

[5] A. K. Jain, **"Data clustering: 50 years beyond K-means. Pattern recognition letters,"**, *Pattern Recognition Letters*, vol. 31, pp. 651-666, Sep 2010.

[6] J. C. Caicedo, and E. Izquierdo, **"Combining Low-level Features for Improved Classification and Retrieval of Histology Images, "**, *Transactions on Mass-Data Analysis of Images and Signals*, vol. 2, pp. 68-82, 2010.

[7] U. Ravindran, and T. Shakila, **"Content based image retrieval for histology image collection using visual pattern mining, "** , *International Journal of Scientific & Engineering Research*, vol. 4, Apr 2013.

[8] J. A. Vanegas, J. Arevalo, F. A. González, **"Unsupervised feature learning for content-based histopathology image retrieval, "**, *Proceedings - International Workshop on Content-Based Multimedia Indexing*, 2014.

[9] D. G. Lowe, **"Distinctive image features from scale-invariant key points, "**, *International Journal of Computer Vision*, vol. 60, pp. 91-110, Jan 2004.

[10] K. A. A. Nazeer, S. D. M. Kumar, and M. P. Sebastian, **"Enhancing the K-means clustering algorithm by**

using a O (n logn) heuristic method for finding better initial centroids, " , *Emerging Applications of Information Technology (EAIT), Second International Conference on IEEE*, pp. 261-264, 2011.

[11] M. Goyal, S. Kumar, **"Improving the Initial Centroids of K-means Clustering Algorithm to Generalize its Applicability, "**, pp. 345-350, Jul 2014.

[12] D. Ebehard and E. Voges, **"An Approach for Selecting Optimal Initial Centroids to Enhance the Performance of K-means, "** *International Journal of Advances in Computer Science and its Applications – IJCSIA*, 2014.

[13] T. Kato, **"Influence of harmonics on power distribution system protection, "**, *Proceedings of SPIE Conference on Image Storage and Retrieval Systems*, vol. 1662, pp. 112-123, 1992.

[14] Md. S. Mahmud, Md. M.Rahman, and Md.N. Akhtar, **"Improvement of K-means clustering algorithm with better initial centroids based on weighted average, "**, *Electrical & Computer Engineering (ICECE), 7th International Conference on IEEE*, 2012.