Proceedings of the 9th WSEAS International Conference on Applied Mathematics, Istanbul, Turkey, May 27-29, 2006 (pp332-337)

Fuzzy C-Means Initialized by Fixed Threshold Clustering for Improving Image Retrieval

NAWARA CHANSIRI[1] , SIRIPORN SUPRATID[2] ,CHOM KIMPAN[3]
Faculty of Information Technology
Rangsit University Muang-Ake, Paholyotin Road, Patumtani, 12000
THAILAND

Abstract: - Fuzzy C-Mean (FCM) algorithm is one of the well-known unsupervised clustering techniques. Such an algorithm can be used for unsupervised image clustering. Then, images can be indexed in databases. The different initializations cause different evolutions of the algorithm. Random initializations may lead to improper convergence. This paper proposes FCM initialized by fixed threshold clustering. The case study regards to retrieve from the database the color JPEG images, indexed by color histogram vectors. The result shows that the proposed method gives more accurate results than FCM with random initialization and color histogram clustering do.

Key-Words:-Fuzzy-C Mean, Color histogram, Image retrieval, Image clustering

1 Introduction

The growing demand for accurate access and retrieval of information has extended to visual information as well. The Internet and the World Wide Web are certainly part of this evolution. The search for images that are similar to a given query example, from the point of view of their content, has become a very important part of research. To retrieve similar images from an image database for a given query image, i.e., a pattern image, image indexing is utilized [1, 2, 3, 4].

Image indexing became color-oriented, since most of the images of interests are in colors. Many of the previous researches used the color composition of an image [4, 5, 6]. Using color histogram is one way to represent or index an image. The color histogram vector is obtained by discretizing the image colors and counting the number of times each discrete color occurs in the image. The main idea is to compute a color distribution from the query image and to compute it with the same distribution computed for each image within the targeted database. The advantages of the histogram are that it is invariant for translation and rotation of the viewing axis. With this method, the histogram is changed very little when comparing the images taken, with little change of the angle of view. The histograms represent primary colors, which are red, green and blue. When the colors are extracted, they are seperated and counted into red, green and blue histograms. The use of color (viewed and used as a vector) was proposed as an important mean of retrieve similarities. One of these examples is a research [6] using color histogram to measure the similarity between two images can be defined.

Fuzzy C-Means (FCM) [7,8] is one of the well-known unsupervised clustering techniques. It allows one piece of data belong to two or more clusters. The aim of FCM is to find cluster center (centroids) that minimize dissimilarity. Its strength over the famous k-Means algorithm [8] is that, given an input point, it yields the points membership value in each of the clusters. The drawback of clustering algorithms like FCM which are based on hill climbing heuristic, is prior knowledge of the number of clusters in the data is required. It was indicated in [7] that FCM have significant sensitivity to cluster center initialization.

Fixed threshold clustering used in [9] is applied with FCM here. Such a clustering is a segmentation of a hierarchical technique for clustering. A large cluster is divided into smaller clusters. A distance comparison between the mean of the cluster and an image is calculated. The result is the number of clusters and the cluster centers that are used to initialize FCM for further clustering process.

According to this paper, the algorithm of FCM initialized by fixed threshold clustering is proposed. The initialization is relevant to the number of clusters and the cluster centers as aforementioned. The case study regards to retrieve from the database the color JPEG images indexed by color histogram vectors. It is noticeable that after using the proposed clustering algorithms, all images have some degree of membership in each possible cluster. Then they are stored in the database for later search.

The paper is organized as follow. First, the introduction is described. In section 2, the algorithm of FCM initialized by fixed threshold clustering for image clustering is shown. Section 3 illustrates the search for images using a sample image one. Section 4 shows the experiment and results. The conclusion is drawn in the final section.

2 Fuzzy C-Mean initialized by Fixed Threshold Clustering for image clustering

Similar to [6], a color histogram was used to represent color compositions of an image. Its utilization as the useful features array can express the characteristic of the image. The computation procedures of the color histogram are shown as follows:

Step 1: A color space of three axes (red, green, and blue) is quantized into *n* bins for each axisn as shown in fig.1 [7]. Then, the histogram can be represented as an $n \times n \times n$ array.

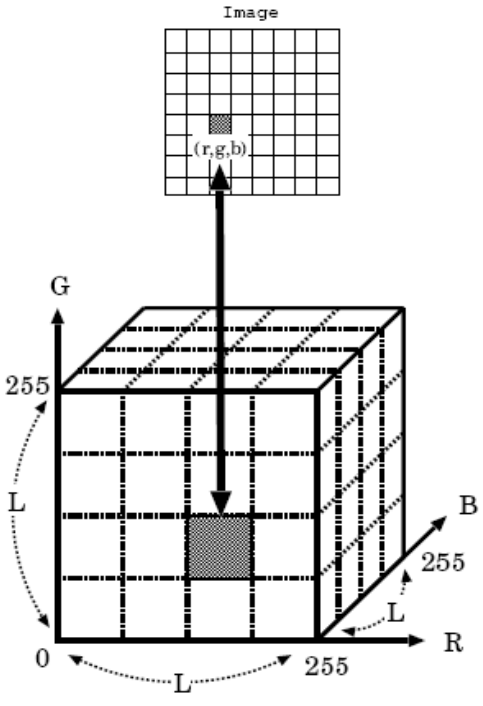

Fig.1 Calculation of the color histogram

Step 2: The colors in the image are mapped into a discrete color space *(r, g, b) (r, g, b = 0, 1, …, n-1)*. Then, the color histogram *F(r, g, b) (r, g, b = 0, 1, …, n-1)* of the colors in the target image are obtained.

Step 3: The probability distribution *P(r, g, b) (r, g, b = 0, 1, …, n-1)* is defined by normalizing the color histogram *F(r, g, b) (r, g, b = 0, 1,…, n-1)* as follows:

$$P(r,g,b) = \frac{F(r,g,b)}{\sum_{i=0,j=0,k=0}^{n-1} F(i,j,k)} \quad (1)$$

When there is a need to cluster images, color histogram could also be of help. By utilizing the color histograms, the similarity between two images can be defined as shown in step 4.

Step 4: The similarity between two images f_1 and f_2 is denoted by $S(f_1,f_2)$ and defined as follows:

$$S(f_1,f_2) = \sum_{r,g,b-0}^{n-1} min\left(P_1(r,g,b),P_2(r,g,b)\right) \quad (2)$$

where $P_1(r, g, b)$ and $P_2(r, g, b)$ express the color histograms of the images f_1 and f_2, respectively. Further, min $(P_1(r, g, b), P_2(r, g, b))$ represents the minimum value of $P_1(r, g, b)$ and $P_2(r, g, b)$.

Color histogram vectors could be used for representing or indexing the images. After doing this, the process of clustering images begins. Random initialization may lead to improper convergence regarding to FCM algorithm. The algorithm requires a proper initialization for good convergence. Here, fixed threshold clustering is applied to a color histogram vector with an expectation to produce a more-proper initialization for FCM. Such a clustering is a segmentation of a hierarchical technique for clustering. A large cluster is divided into smaller clusters. There is a distance comparison between the mean of the cluster and an image. If the distance value is lower than the threshold limit then the image would be labeled as a cluster name. If it is higher than the threshold then the image would be assigned as "*NoneMember*" and continue seeking for a suitable cluster in the next generation.

The algorithm of fixed threshold clustering is shown in fig.2:

For p=1 to p=n (n = number of image)

image$_p$ is named as *"NoneMember"*
j = 0
Loop Until all images are not named as *"NoneMember"*
Set cluster = C_j
For *p*=1 to *p*=*n*
image$_p$ is randomly selected as a *Mean_of_* C_j
Compute Distance (*Dist*) between *Mean_of_*C_j and *image*$_p$ using City Block method
If ((*Dist*) <= *threshold*)
Then
image$_p$ is named as C_j
Compute *New_mean* between *Mean_of_* C_j and *image*$_p$
Save *image*$_p$ as a membership of C_j
End If
End For
Save C_j and *Mean_of_*C_j
j++
End Loop

Fig.2 Fixed threshold clustering

*Mean_of_*C_j refers to the center of cluster C_j. Such a clustering results in a list of center values of image clusters and an identified image cluster of each image. Such identification leads to a recognition of the number of clusters. Then they are stored in a database.

Later, Fuzzy C-Means algorithm uses the cluster centers obtained from the previous clustering process as an initialization. The initialization consists of the number of clusters and the cluster centers. The general purpose of FCM is to minimize the objective function

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^m \left\| image_i - C_j \right\|^2 , 1 \le m < \infty \quad (3)$$

Let CL^k = the center vector at time $t = k$
$CL^k = [\, C_j \,]$
$m = 2$

u_{ij} is the degree of membership of *image*$_i$ in the cluster C_j

$U^k = [u_{ij}]^k$

$$\sum_{j=1}^{C} u_{ij} = 1, \text{ for } \forall\, i = 1,.., n \quad (4)$$

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\left\| image_i - C_j \right\|}{\left\| image_i - C_k \right\|} \right)^{\frac{2}{m-1}}} \quad (5)$$

$$C_j = \frac{\sum_{i=1}^{N} u_{ij}^m \cdot image_i}{\sum_{i=1}^{N} u_{ij}^m} \quad (6)$$

The FCM algorithm following the algorithm providing a proper initialization is given in fig.3:

$\varepsilon = 0.01$
Initialize CL^0 = [C_j] provided by fixed threshold clustering
U^0 = [0]
k = 1
Calculate U^k
Loop Until $| U^k - U^{k-1} | < \varepsilon$
Calculate CL^k
k++
Calculate U^k
End Loop
$U^* = U^k$

Fig.3 FCM algorithm following fixed threshold clustering

Then U^* is stored in a database. U^* could be represented by $[\, u_{ij} \,]^*$. Each element in U^* matrix indicates the degree of membership of the clusters for each image.

3 Searching Images

To search for images in the system, a vector $U_i^* = [u_j]_i^*$ is set to be the elements in row *i* in U^*. Therefore, U_i^* represents the degree of membership of *image*$_i$ for every possible cluster , C_j. Then, the vector of a sample image , $X = [x_j]$ is used to query for a set of in-condition images from the database. The query would be simply written as :

Select U_i^* *where* $|U_i^* - X| \leq \varepsilon$ *from the database*

Where $\varepsilon = 0.01$. If $|U_i^* - X| \leq \varepsilon$ is true, when the following predicate is true

$$\forall j \left(\sqrt{(u_1 - x_1)^2 + .. + (u_j + x_j)^2} \leq \varepsilon \right) \quad (7)$$

4 Experiments and results

4.1 Experiments

After, a variety of images has been extracted in to color histograms vectors. Using the FCM initialized by fixed threshold clustering method, all images are clustered into suitable groups and are stored in a database.

1850 color JPEG images are used for the experiments. 143 color JPEG images are utilized for testing the system. Each image contains 13 similar images. Twelve of them are exactly the same. Only one of them is similar but not the same as the other twelve. The images are 128×128 pixels in size and in many different classes, such as flowers, buildings, natural, etc.

4.2 Results of the Experiments for Image Retrieval

The experimental research is concerned with the accuracy of the image retrieval, shown in fig. 4, 5, 6, 7 as examples of the result. Fig.7 is shown as an example of a result that is not accuretely retrieved from our system. The larger image on the left is the query image, and the 12 images on the right are the image results. The comparison among the proposed method, color histogram using FCM with random initialization, and color histogram is performed. It is noticeable that all three algorithms proceed on the same input data. Such input data are color histogram vectors produced from the same set of color JPEG images. Table 1 shows the comparison in percent of accurately retrieved images and the time for extraction per 1846 among the proposed method, color histogram using FCM with random initialization, color histogram method.

Fig.4 Accurately retrieval result in our method

Fig.5 Accurately retrieval result in our method

Fig.6 Accurately retrieval result in our method

Fig.7 Inaccurately retrieval result in our method

TABLE 1

SHOW COMPARISON IN PERCENT OF ACCURATELY RETRIEVED IMAGES AND TIME FOR EXTRACTION.

Method	% accurate	Time for Extrac
Color histogram using FCM initialized by fixed threshold clustering	81.3017	00:03.09.8943
Color histogram using FCM with random initialization	69.059	00:03.13.5630
Color histogram	56.8387	00:03.09.8943

5 Conclusions

In this paper, the algorithm of FCM initialized by fixed threshold clustering is proposed. The initialization is relevant to the number of clusters and the cluster centers. The case study regards to retrieve from the database the color JPEG images indexed by color histogram vectors. There exists the comparison among the proposed method, color histogram using FCM with random initialization, color histogram method. The same set of color histogram vectors are used as input data. The result shows that although the accurate percentage of the proposed method is not very high, but it gives more accurate results than FCM with random initialization and color histogram clustering do. Since the color histogram may not be very high efficient color image representation, therefore, the future work may be finding the very high efficient technique for representing color image to digital information. Another interesting future work would be optimizing the FCM initialization.

References:

[1] Qge Marques and Borko Furht, *Content-based image and video retrieval,* U.S.A., Kluwer Academic Publishers, 2002

[2] H. J. Zhang et al, An integrated system for content-based video retrieval and browsing, *Pattern Recognition*, May, 1997

[3] W. Y. Ma and H. J. Zhang, Content-based image indexing and retrieval, in *Handbook of Multimedia Computing*, Borko Furht, ed. CRC Press, 1998

[4] M.J. Swain and D.H. Ballard, Color indexing, *International Journal of Computer Vision*, Vol.7, No.1, 1991, pp. 11-32.

[5] Chuping Liu, Julien Lamoureux. Yuxin Wang and Yunan Xiang, *Histogram algorithm for image indexing,* http://www.ee.ualberta.ca.html [Accessed 2004 Sep 20].

[6] T. Ohara, T. Ogawa, M. Haseyama and H. Kitajima, A Similar Image Clustering Method Including Automatic Selection of Number of Clusters, *IWAIT Conference Program,* 2006

[7] F. Hoppner, F. Klawonn, R. Kruse and T. Runkler, *Fuzzy Cluster Analysis*, Wiley, 1999

[8] Bezdek JC, Partition Structures: A tutorial, In: *The Analysis of fuzzy Information,* J.C. Bezdek, Ed. Boca Raton, FL: CRC Press, 1987, pp. 81-108.

[9] P. Phokharatkul, S. Chaisriya, S. Somkuarnpanit, S. Phiboon and C. Kimpan, Developing the Color Temperature Histogram Method for Improving the Content-Based Image Retrieval, *International Academy of Sciences,*Vol.8, 2005, pp. 270-275.