Improving Image Retrieval by Fuzzy C-Means Initialized by Fixed Threshold Clustering: case studies relating to a color temperature histogram and a color histogram

DOUNGPORN NIYOMUA¹, SIRIPORN SUPRATID², CHOM KIMPAN³
Faculty of Information Technology
Rangsit University Muang-Ake, Paholyotin Road, Patumtani, 12000 THAILAND

Abstract: Fuzzy C-Means (FCM) algorithm is one of the well-known unsupervised clustering techniques. Such an algorithm can be used for unsupervised image clustering. The different initializations cause different evolutions of the algorithm. Random initializations may lead to improper convergence. This paper proposes FCM algorithm initialized by fixed threshold clustering. The purpose of the algorithm is to retrieve from the database the color JPEG images. Two case studies regard to index or represent the color images by either using color temperature histogram or color histogram vectors. The clustering process produces from such an image index the information, which is a degree of membership for each image. This information would be stored in a database. This paper shows that for both two cases, FCM algorithm initialized by fixed threshold clustering gives more accurate results than FCM with random initialization does.

Key-Words:- Fuzzy-C Means, 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 or image representation 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]. 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. One way to use the color composition to represent or index an image is to utilize color temperature histogram. The color temperature is a term used to describe certain properties of the light sources [7]. It can represent 46 - color temperature range from 1667º - 100,000º K and also can represent black, gray images. It can separate different colors that have the same color temperature. Another one way is to use color histogram vector. Such a vector is obtained by discretizing the image colors and counting the number of times each discrete color occurs in the image. The advantages of color 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. However, histograms represent only 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 also proposed as an important mean of retrieve similarities. One of these some research such as [6] using color histogram to define the similarity between two images however, It is noticeable that color temperature histogram takes the advantage of color temperature and color histogram.

In order to retrieve similar images from an image database for a given query image, many clustering techniques are widely used. Fuzzy C-Means (FCM) [8, 9] was originally introduced by Jim Bezdeck [a] as an improvement on earlier clustering method. It is one of the well-known unsupervised clustering techniques. It provides a method that shows how to group data points that populate some multidimensional space into a specific number of different clusters. In order word, 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 [9] 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 [8] that FCM have significant sensitivity to cluster center initialization.

Fixed threshold clustering used in [10] 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. Two case studies are with regard to index or represent the color images by either using color temperature histogram or color histogram vectors. Then the clustering process produces from such an image index the information, which is a degree of membership for each image. This information would be stored in a 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 with in two case studies. 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-Means initialized by Fixed Threshold Clustering for image clustering

According to this paper, all images are to be indexed or represented by using either one of these two techniques illustrated in 2.1 and 2.2

2.1 Color histogram

Similar to [6], a color histogram could be 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 axis in as shown in fig.1 [6]. Then, the histogram can be represented as an $n \times n \times n$ array.

![Fig.1 Calculation of the color histogram](#)

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

2.2 Color temperature histogram

Color temperature is used to described a certain property of the light sources [7]. It is calculated from Planck’s formula and it based on the relationship between the temperature of a black body radiator and the color of its emitted light when it was increased higher temperatures [11]. The color temperature is used to show the color appearance of the light by compare with color of a black body radiator. When the color of the light matches the black-body radiator then the light has the same color
temperature value as the black-body radiator. However, many radiators are not exactly equal to any of the color of a black-body radiator. So, a correlated color temperature concept is proposed for color temperature calculation of these radiators. The correlated color temperature was calculated by Robertson’s algorithm [7] and used the perceptually homogeneous UV chromaticity diagram (CIE 1960) [12]. In this research, the ‘color temperature’ term is practically used instead of the ‘correlated color temperature’.

The color temperature histogram is a combination of the color temperature method and histogram method for descriptive color in an image. A color temperature histogram consists of 46 bins. Each bin represents an interval of color temperature values. The color temperature values are calculated and derived from approximated color regions [13].

The color temperature histogram is created from the following steps:

1) Segment an image into blocks because an image has various color temperatures. A block has a size of 4 x 4 pixels. This size gives higher accurate retrieved results and lowers the computation cost as shown in Table 1 in the experiments and results section

2) Calculate the color temperature of each segment using color temperature estimation algorithm [12]. This method cannot be used with a black color image because R, G, B value is 0. So \( \beta \), a value out of 0-600 and parameter = 3 are defined for a segment that is black. For a segment with equal R, G and B value, one of the parameters 0=black, 1=gray and 2=white is added to separate the degree of gray, white and black.

3) All colors are represented by a color temperature value. They are calculated from the same isotemperature line. Then a parameter, \( \chi \) that indicate the color temperature located over or lower than Planckian locus line [11]. A parameter \( \chi \) is 1=over a Planckian locus line, 0=lower than a Planckian locus line.

4) Create the color temperature histogram by counting the segments that have the same color temperature value and parameter \( \chi \) then put it in the bin.

2.3 Clustering Process

When all images have been changed into color temperature histogram vectors or color histogram vectors, 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:

\[
\text{threshold} = 0.05 \\
\text{For} p=1 \text{ to } p=n \ (n = \text{number of image}) \\
\text{image}_p \text{ is named as “NoneMember”} \\
\text{Set cluster} = C_j \\
\text{Image}_i \text{ is randomly selected as a} \\
\text{Mean_of}_C_j \\
\text{For} p=1 \text{ to } p=n \\
\text{Compute Distance (Dist)} \\
\text{between Mean_of}_C_j \text{ and} \\
\text{image}_p \text{ using City Block method} \\
\text{If ( (Dist) } \leq \text{ threshold )} \\
\text{Then} \\
\text{image}_p \text{ is named as } C_j \\
\text{Compute New_mean between} \\
\text{Mean_of}_C_j \text{and image}_p \\
\text{Save image}_p \text{ as a membership of } C_j \text{ End If} \\
\text{End For} \\
\text{Save } C_j \text{ and Mean_of}_C_j \\
f++ \\
\text{End Loop}
\]

Fig 2 Fixed threshold clustering

\[\text{Mean_of}_C_j \text{ refers to the center of cluster } C_j. \text{ Such a clustering results in a list of center values of image clusters and an identified image cluster of each image. Such identification leads to a recognition of the number of clusters.}\]

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

\[ J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^m \| \text{image}_i - C_j \|_2^2, I \leq m < \infty \] (2)

Let \( CL_k = \) the center vector at time \( t = k \)

\[ CL_k = [ C_j ] \]

\( m = 2 \)

\( u_{ij} \) is the degree of membership of \( \text{image}_i \) in the cluster \( C_j \)

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

\[ \sum_{j=1}^{C} u_{ij} = 1, \text{ for } i = 1, \ldots, n \] (3)

\[ u_{ij} = \frac{1}{\sum_{k=1}^{C} \left( \frac{\| \text{image}_i - C_k \|_2}{\| \text{image}_i - C_j \|_2} \right)^{m-1}} \] (4)

\[ C_j = \frac{\frac{N}{\sum_{i=1}^{N} u_{ij}^m} \cdot \text{image}_i}{\sum_{i=1}^{N} u_{ij}^m} \] (5)

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

\( \varepsilon = 0.03 \)

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 \( U_i^* = [u_{ij}]^* \) is set to be the elements in row \( i \) in \( U^* \).

Therefore, \( U_i^* \) represents the degree of membership of \( \text{image}_i \) for every possible cluster \( C_j \). Then, the vector of a sample image \( X = [x_j] \) contains the degree of membership of the clusters for that image. It is used to query for a set of in-condition images from the database. The query would be simply written as:

\[ \text{Select } U_i^* \text{ where } |U_i^* - X| \leq \varepsilon \text{ from the database} \]

Where \( \varepsilon = 0.03 \). If \( |U_i^* - X| \leq \varepsilon \) is true, when the following predicate is true:

\[ \forall j \left( (|u_{ij} - x_i| + \ldots + |u_{ij} - x_j|) \leq \varepsilon \right) \] (6)

4 Experiments and results

4.1 Experiments

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

1859 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, cartoons, animals, etc.

4.2 Results of the Experiments for Image Retrieval

The experimental research is concerned with the accuracy of the image retrieval. Fig. 4, 5, 6 and 7 are shown as examples of the results when using color temperature histogram. The larger image on the left is the query image, and the 12 images on the right are the image results. Fig.4, 5 and 6 are three examples of accurate retrieval results fig.4, 5 and 6 shows the advantages of color histogram obtained from color histogram with respect to the invariance
for rotation of reviewing axis, as seen in picture 5.3, 5.4, 5.7.

Fig. 7 is shown as an example of a result that is not accurately retrieved from our system. The error may occur from two sources: a) the efficiency of color image transformation. b) the fixed threshold retrieved in fig. 2 and ε in section 3 may be too high to indicate the different. The comparison between two case studies using FCM with random initialization, and FCM initialized by fixed threshold clustering is performed. All the cases proceed on the same input data. Such input data are color histogram vectors or color temperature histogram vectors produced from the same set of color JPEG images. Table 2 shows the comparison in percent of accurately retrieved images and the time for extraction per 1859.

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

<p>| TABLE 1 |
| SUITABLE SIZE OF SEGMENT IN THE IMPLEMENTATION |</p>
<table>
<thead>
<tr>
<th>Size</th>
<th>Incorrect Retrieve</th>
<th>Time for Extract (for 1859 images)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1x1</td>
<td>0.3543</td>
<td>02:32:23.5687</td>
</tr>
<tr>
<td>2x2</td>
<td>0.2253</td>
<td>00:09:12.8243</td>
</tr>
<tr>
<td>3x3</td>
<td>0.0945</td>
<td>00:03:04.6871</td>
</tr>
<tr>
<td>4x4</td>
<td>1.4954</td>
<td>00:01:45.8765</td>
</tr>
</tbody>
</table>
TABLE 2
SHOW COMPARISON IN PERCENT OF ACCURATELY RETRIEVED IMAGES AND TIME FOR EXTRACTION.

<table>
<thead>
<tr>
<th>Method</th>
<th>% accurate</th>
<th>Time for Extract</th>
</tr>
</thead>
<tbody>
<tr>
<td>Color temperature histogram using FCM</td>
<td>94.174</td>
<td>00:02:06.6530</td>
</tr>
<tr>
<td>initialized by fixed threshold clustering</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Color temperature histogram using FCM</td>
<td>87.63</td>
<td>00:02.34.9820</td>
</tr>
<tr>
<td>with random initialization</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Color histogram using FCM initialized by</td>
<td>81.3017</td>
<td>00:03.09.8943</td>
</tr>
<tr>
<td>fixed threshold clustering</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Color histogram using FCM with random</td>
<td>69.059</td>
<td>00:03.13.5630</td>
</tr>
<tr>
<td>initialization</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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. Two case studies are with regard to index or represent the color JPEG images by either using color temperature histogram or color histogram vectors. Then the clustering process produces from such an image index the information, which is a degree of membership for each image. This information would be stored in a database. There exists the comparison between two case studies using FCM with random initialization, and FCM initialized by fixed threshold clustering. The same set of color histogram vectors and color temperature vectors are used as input data. The result shows the accurate percentage of the algorithms. For both two cases, FCM algorithm initialized by fixed threshold clustering gives more accurate results than FCM with random initialization does. The future work may be finding the higher efficient technique for representing color image to digital information. There may be other more distance measurements than a simple city block. Another interesting future work would be finding the better way to optimize the FCM initialization.

References: