Comparison of Three Techniques for Clustering Color Histogram Based Image

Petcharat Pattanasethanon
Faculty of Accountancy and Management
Mahasarakham University
Mahasarakham, Thailand
e-mail: april_it2005@hotmail.com

Singthong Pattanasethanon
Faculty of Engineering
Mahasarakham University
Mahasarakham, Thailand
e-mail: sing191@hotmail.com, singthong.p@msu.ac.th

Abstract—**In this paper, we proposed an assessment of three techniques in color image clustering that is efficient enough to be applied to databases namely, Hierarchy, K-Means and Fuzzy C-Means clustering techniques. Color image histograms in the RGB space and CIEL*a*b space is processed in three clustering algorithm. The details of the three clustering techniques are compared for its accuracy. Furthermore, the computation algorithm is developed into a GUI application by MATLAB for its convenient use. In this experiment, 100 color images is classified into 10 clusters. The results that we obtain from each cluster are based on the RMSE (Root Mean Square Error), Entropy, and Purity. The result indicates that Fuzzy C Means method as a clustering technique shows the most clustering efficient, proven by its RMSE which obtains the least difference in error. The mean of the results in terms of Entropy and Purity is also at the minimum. However, clustering techniques such as K-Means and Hierarch shows lower clustering quality as it is illustrated in the results.**

Keywords-hierarchy; k-means; fuzzy c-means; clustering; colour spaces

I. INTRODUCTION

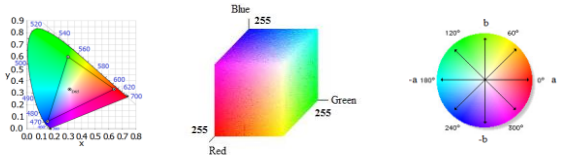
Recently, the problem of the color images clustering is far more complexed than early days. It is almost impossible for human to do all of the work without the assistance of machines. Clustering data is effect to many applications in image processing such as improving speed for image retrieval [1], increasing precision of the object image recognition, or reducing the complexity of ontology system processing [2]. Furthermore, there are distinctively several clustering techniques which are proposed such as supervised, semi-supervised, and unsupervised techniques [3]. However, only few literatures address the comparison of clustering techniques performance.

In this research, we attempt to compare unsupervised clustering efficiency in color space RGB and CIEL*a*b. The clustering methods compose of three techniques namely K-Means (KM) clustering, Hierarchical clustering, and Fuzzy C Means (FCM) clustering. The focus of this paper is also to compare the advantages of three clustering algorithms; hierarchical clustering, K-means clustering, and Fuzzy C-means clustering, respectively. This distinguished objective is addressed for color images, on the basis of color histograms of RGB and CIEL*a*b in the color space [4]. Our paper evaluates the methods of error detection based on RMSE (Root Mean Square Error). The efficiency of data clustering in this case can be analyzed by entropy and purity measures [5]. This paper is organized accordingly to each section. Our second section includes relationship of color space and histogram manipulation. The third section discusses about the three clustering algorithms on its principles: Hierarchy, KM, and FCM. The fourth section is about the clustering efficiency of each method. The fifth section is the experiments and comparison. Finally, all sections are concluded.

II. COLOR SPACES AND HISTOGRAM MANIPULATION

A. *Relationship of RGB color space and CL*a*b*

In digital color image, we consider color image in RGB space or CIEL*a*b space [6]. Due to the fact that RGB color space and CIEL*a*b color space are part of the color space in order to widely used since 1931 [7], we therefore, chose this method to test the clustering performance in terms of its efficiency. Color images with different data in color space will be transformed into CIEL*a*b. Each visible color has nonnegative coordinates X, Y, Z. which consist in trichromatic diagram as shown in figure 1.(a)



(a) (b) (c)

Figure1. Colour spaces (a) Trichromatic diagram with standard D65, (b) 3 dimensions of chromatic diagram, and (c) L*a*b in 2 dimension

The RGB transformation algorithm to L*A*B in digital color images is as described below.

- Set the reference values as in X, Y, Z according to the standard D65 (D65 is Daylight at 65005 K) ; X=95.047, Y=100, Z=108.883
- Normalize X, Y, Z the transfer to x, y, z
- Calculate; l, *a, *b* ;

$$l^* = (116 \cdot y) - 16 \quad (1)$$

$$a^* = 500 \cdot (x-y) \quad (2)$$

$$b^* = 200 \cdot (y - z) \quad (3)$$

In this research, we used *srgb2lab* command in MATLAB to transform into color space model.

B. Color histogram manipulation

The color histogram always used to be the characteristic of color image. It is constructed by counting the number of pixels of each color. In this work, the color histogram manipulation refers to the probability mass function of the image intensities. Formally, the color histogram is defined by

$$h_{A,B,C}(a,b,c) = N \cdot Prob(A = a, B = b, C = c) \quad (4)$$

Where, *A*, *B* and *C* represent the three color channels (R, G, B) and *N* is the number of pixels in the image. Computationally, the color histogram is formed by discretizing the colors within an image and counting the number of pixels of each color.

Since the typical computer represents color images with up to 256 colors, this process generally requires substantiation of the color space. The color histogram can be thought of as a set of vectors. For color images, the color histograms are composed of 4-D vectors. One of the easiest is to view separately the histograms of the color channels. This type of visualization does illustrate some of the salient features of the color histogram [8].

III. Clustering Image Techniques

A. K Means Clustering

As it is mentioned in the introduction that k-means clustering method deals with minimizing average squares Euclidean distance [9]. The k- means algorithm of our works is first to perform a partition step which divides 'n' objects in the data set of 'k' categories. Given the partition into k groups and substitute each group with its average squared Euclidean distance as follows.

$$D_k = | x_i^{(n)} - c_k |^2 \quad (5)$$

Where D_k is the minimum distance of the centroid of the object
c_k is the centroid of group k[th], generated by random
k is the cluster number
n is the number of images in the group; 1,2,3,…n
$x_i^{(n)}$ is the image vector i in group n

Convergence of the centroid in each group to settle, the new centroid (C_{new}) is calculated as follow,

$$C_{new} = \frac{\vec{x} + C_{old(x)}}{2} \quad (6)$$

Where, $C_{old(x)}$ is the old centroid value; $\vec{x}$ is the image vector

In this case, the centroids or means of the categories is able to determine. The idea is that the clusters should not overlap and we have no labeled training set in clustering for which we know which data should be in the same cluster. A measure of how well the centroids represent the member of their clusters is the residual sum of square (RSS) or uncertain radius, the squared distance of each vector from its centroid summed over all vectors:

$$RSS_k = \sum_{\vec{x} \in \omega_k} |\vec{x} - \vec{\mu}(\omega_k)|^2 \quad (7)$$

Where $\vec{\mu}(\omega_k)$ is the center or centroids of the document of cluster ω_k

RSS is an objective function in K-means and the key is to minimize it. A measure of how well the centroids represent their data is subjected to the fact that minimizing RSS is equivalent to minimizing the average squared distance for which 'n' is fixed. The member in the data set is determined by the shortest average squared Euclidean distance from the centroid. The shortest path will discriminate the data as the step consecutively repeat its iteration and stop as the centroid is constant and stable. The process of k-means algorithm in this paper is as shown in the figure 2.

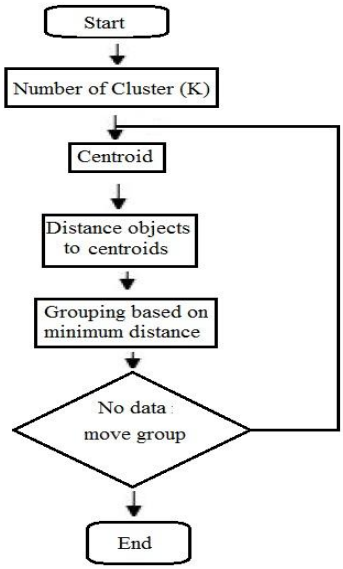


Figure2. Flow chart of K Means algorithm

In this process, the first step is to arbitrarily set the 'k' into perspective number of groups. For instance, in this paper we set to 10 clusters, k=10. Then follow by initializing the centroid for each cluster and find the shortest path for Average Square Euclidean distance in order to eliminate the RSS from the objective function. In this clustering section, we used *Kmeans* command in MATLAB to conduct the clustering process.

B. Hierarchy Clustering

A concept hierarchy consists of two parts: a node set and link set [10]. This paper extends the distance hierarchy structure with link weight. Each link has a weight representing a distance. It is also a better mechanism to facilitate the representation and computation of the distance between categorical values. We can consider hierarchy structure as follows. Suppose, A point X in a distance hierarchy consists of two parts, an anchor and a positive real value offset, denoted

as *X(N,d)*,that is, anchor*(X)*= N and offset *(X)*= d. The anchor becomes leaf node and the offset represents distance from the root of hierarchy to the point. A point X is an ancestor of Y if X is in the part from Y to the root of the hierarchy. Let X(Nx,dx) and Y(Ny,dy) be two points, the distance between X and Y can be defined as [11]

$$|X - Y| = d_X + d_Y - 2d_{LCP(X,Y)} \quad (8)$$

Where LCP(X,Y) is the least common point of X and Y in the distance hierarchy, $d_{LCP(X,Y)}$ is the distance between the least common point and the root.

Hierarchical clustering technique can be easily expressed in binary tree graph or dendrogram. It represents the relationship of the clusters, sub clusters, and level of clusters. In our work, the hierarchy clustering command in MATLAB is also conducted in the clustering processes. Those are *pdist*, *linkage* and *cluster* function respectively.

C. Fuzzy C Means Clustering

The Fuzzy C Means (FCM) clustering is the combination of fuzzy algorithm, C Means clustering, and thresholding algorithm [14]. The goal of clustering analysis is to divide a given set of data or objects into a cluster, which represents subsets or a group [12]. The partition should have two properties. Those are homogeneity inside clusters and heterogeneity between the clusters.

1) Clustering

In our work, a partition, P is a set of disjoint subsets of E and the element Ps of P is called *cluster* and the centers of the clusters are called *centroids* or prototypes. The membership functions do not reflect the actual data distribution in the input and output spaces. To build membership function from the data available, a clustering technique may be used to partition the data, and then produce membership functions from the resulting clustering.

2) C Means

Many techniques have been developed for data clustering. In this paper c-means clustering is used for data grouping or classification when the number of the clusters is known. It consists of the following steps:

Step 1: Choose the number of clusters-K

Step 2: Set initial centers of cluster $C_1, C_2, \ldots, C_k$;

Step 3: Classify each vecter $X_1 = [x_{11}, x_{12}, \ldots x_{1n}]^T$ into the Closest center C_i by Euclidean distance

measure: $\| x_i - c_i \| = \min \| x_i - c_i \|$

Step 4: Recomputed the estimates for the cluster enter C_i; let $C_i = [C_{i1}, C_{i2}, \ldots C_{in}]^T$; C_{im} be computed by:

$$C_{im} = \frac{\sum X_{li} \ \epsilon cluster \ (i^{Xlim})}{N_i} \quad (9)$$

Where N_i is the Number of vectors in the i-th cluster

Step 5: If more of the cluster center (C_i = 1,2,…,k) in step 4 stop; otherwise go to step 3

In practice, the criterion function used for fuzzy C- means clustering is;

$$J(v) = \sum_{i=1}^{c} \sum_{k=1}^{n} U_{ik}^{n} |x_k - v_i|^2 \quad (10)$$

Where $X_1 \ldots X_n$ - 'n' data sample vector,

$V_1 \ldots V_c$ - 'c' denotes cluster centers (centroids);

$U = U_{ik}^{cxm}$ matrix, where U_{ik} is the i-th membership value of the *k-th* input sample *xk*, and

The membership values satisfy the following conditions:

$0 \leq u_{ik} \leq 1; \quad i = 1, \ldots, c; k = 1, \ldots, n;$

$\sum_{i=1}^{c} u_{ik} = 1; \quad k = 1, \ldots, n;$

$0 < \sum_{k=1}^{n} u_{ik} < 1; \quad i = 1, \ldots, c;$

$m \in [1, \infty)$ is an exponent weight factor

3) Thresholding

Here fuzzy c means clustering [13] is used based on thresholding. In normal fuzzy c means clustering the segment part cannot be seen clearly. For this reasons, thresholding is applied to extract the region color of input part.

For our work for FCM clustering, the Fuzzy Logic Toolbox command line function, fcm, assigns every data point a membership grade for each cluster. By iteratively updating the cluster centers and the membership grades for each data point. FCM iteratively moves the cluster centers to the right location within a data set.

IV. Clustering Efficiency

Clustering efficiency is always used to indicate accuracy of the image cluster. The three main components to define the quality of our clustering methods are RSME, purity, and entropy. These three components display several aspects of the data and centroid as well as the clustering quality [14].

A. Root Means Square Error

RMSE is the solution of the square root from mean square error which can be expressed as the equation below [15].

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(Y_t - \hat{Y}_t)^2} \quad (11)$$

Where Y_t is the pixel value of the data in the cluster and $\hat{Y}_t$ is the mean value of the data in each cluster.

B. Purity

In each and every clustering algorithm the measurement of cluster quality is required and the well-known measurements are purity and entropy. Purity measurement can be expressed as the equation below [16].

$$p(c_j) = \frac{1}{|c_j|_{k=1,\ldots,c}} \max \ |C_{j,k}| \quad (12)$$

Where $|c_j|$ is the amount of members in category '*j*'. And $c_{j,k}$ is the member date of cluster 'k'. However, each category may consist of members from several categories. Purity is under the range of [1/c,1]. As the value as approaches 1, it is proven that clustering quality is higher.

C. *Entropy*

The concept of entropy was used to measure structural complexity. Several invariants such as the number of vertices, the vertex degree sequence, and extended degree sequences (i.e., second neighbor, third neighbor etc.) have been used in the construction of entropy-based measures [17]. Entropy is chosen arbitrarily from a pile of data to measure the relationship between the members.

$$\bar{I}(c_j) = -\sum_{i=1}^{k} P_i log(P_i) = -\sum_{i=1}^{k} \frac{c_{j,k}}{c_j} \log \frac{c_{j,k}}{c_j} \quad (13)$$

We can wide spread the probabilities out of the initial scope because entropy determines the partition of each category under one group. This made entropy more precise than purity. On the other hand, when entropy is normalized in the range of [0,1] the value approaches zero performs better clustering quality.

V. Methodology

The procedure of this experiment is as shown in figure 3. The query image of 100 colored images in total is executed by MATLAB version 7.10. GUI is applied in order for the results to be displayed. The steps are as described below.

1. Input Color image RGB color space then convert RGB color space to CIEl*a*b* color space.
2. Compute for histograms of each image in order to proceed to the clustering step.
3. Compare the color histograms by the mean of the data. Then categorize the data using both hierarchical, k-means and Fuzzy C Means clustering method.
4. Solve for RMSE, entropy, and purity of each cluster to define the clustering quality of the data.

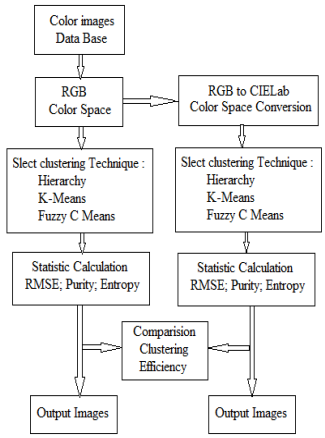


Figure 3. diagram of color image clustering

A. *Constructing GUI*

The displaying methods on the GUI consist of the top left which we fill in the number of 'k' groups and choose the clustering methods; hierarchy or k-means or fuzzy c means. Reset the program to start over. The second part of the GUI on the left is where the centroid of color and the data member of each category is displayed. The left most button is where the value of each data and the output of each cluster is illustrated from the first to the tenth cluster. If the output cluster is greater than 10 'Next' button will display the rest of the data. The last part of our GUI is where the output images of 10 columns and rows are displayed.

VI. Experiments

The image data in the test consist of 100 images of roses in different colors. Some examples of the testing data are as shown in figure 4. The experiment has separated the data into 10 clusters using the three techniques mentioned above; k-means, hierarchy, and fuzzy c means. We also test the color both from RGB and CIEL*a*b color space, respectively. The results demonstrate the comparison of six cases as shown in figure 5-10. In the first case, figure 5 and 6, is where we compare k-means algorithm with the difference of two color space models.

Figure 4. some color images (1-100) for testing

The experiment has separated the data into 10 clusters using the three techniques mentioned above; k-means, hierarchy, and fuzzy c means. We also test the color both from RGB and CIELab color space, respectively. The results demonstrate the comparison of six cases as shown in figure 5-10. In the first case, figure 5 and 6, is where we compare k-means algorithm with the difference of two color space models.

Figure5. RGB color image clustering using k mean method; k = 10

Figure 6. CIEL*a*b color image clustering using k mean method; k = 10

The second case, figure 7 and 8, is the comparison of two color space models but using hierarchy clustering algorithm as the clustering method. In this case, we also set out 'k' value at 10.

Figure 7 RGB color image clustering using hierarchy method; k = 10

Figure 8 CIEL*a*b color image clustering using hierarchy method; k = 10

Figure 9 RGB color image clustering using fuzzy c mean method; k = 10

Fig.10 CIEL*a*b color image clustering using fuzzy c mean method; k = 10

The third case is described in figure 9 and 10. It basically compares both color space models but the algorithm for clustering is Fuzzy C Means, instead. In our experiment for FCM clustering method, we have done the same as k-means and Hierarchy clustering method, by setting the 'k' value as 10 for both input image in color space RGB and CIEl*a*b.

A. Clustering Results

The six cases of clustering algorithms can be described as the results of RGB color space and CIELab color space in table I. The evaluation is based on 5 indices; cluster members, Centroid mean, centroid error (RMSE), entropy, and purity.

TABLE I. NUMBER OF IMAGES IN EACH CLUSTER COMPARISON

Method	FCM		K - Means		Hierarchy	
Cluster No.	RGB	L*a*b	RGB	L*a*b	RGB	L*a*b
1	15	14	6	7	3	1
2	9	10	8	4	21	87
3	9	8	12	11	1	1
4	11	12	15	7	1	3
5	13	3	3	3	55	1
6	11	11	13	22	5	2
7	4	9	4	3	1	1
8	5	12	13	19	6	1
9	11	9	9	8	5	2
10	12	12	17	16	2	1

TABLE II. AVERAGE OF FIVE INDICES FOR CLUSTERING IN SIX CASES

Method	FCM		K - Means		Hierarchy	
Means	RGB	L*a*b	RGB	L*a*b	RGB	L*a*b
Purity	0.747	0.868	0.707	0.842	0.827	0.904
Entropy	0.117	0.109	0.134	0.157	0.147	0.134
RMSE %	3.058	6.96	3.375	7.98	1.692	2.085
Centroid	84.61	77.75	131	149.56	78.22	90.38

1) Results Analysis

Table I shows color images number of 10 clusters with-in RGB and CIEL*a*b color space by clustering with k means,

hierarchy and fuzzy c means, respectively. From the results, FCM and K Mean algorithm can cluster its member both in RGB color space model and CIEL*a*b color space with neighboring members. However, the hierarchical clustering algorithms generate profusely distinctive difference.

From table II, FCM determined the most accurate. That is to say it is about 3.058% and 6.96% tolerance for RMSE. The entropy and purity also proves its clustering performance which is between 0.117 and 0.109 and 0.747 and 0.868 respectively.

The performance of k-means clustering shows accuracy with the RMSE of 3.375% and 7.98%. The entropy is between 0.134-0.157 and the purity is between 0.707 and 0.842.

Hierarchical clustering cannot be considered in this case, since there is no sufficient reference that we can compare. Therefore,it is obvious that the algorithm is insufficient.

VII. Conclusion

This paper proposed an assessment evaluation to determine the sufficient most well known clustering technique such as k-means, hierarchy, and fuzzy C means algorithm. The experiments are executed in MATLAB and a GUI is constructed to compute and display the results statistically. The research focus is on the clustering algorithms by referencing from color space models such as RGB and CIEL*a*b. The implemented algorithms are k-means clustering, hierarchical clustering, and fuzzy C means clustering, where the task is to cluster the data into 10 clusters.

From the results we can conclude that the FCM algorithm is most sufficient and optimized, since it obtains an optimizing data member from the set centroid and euclidean distance. Simultaneously, the entropy and purity are consistent.

K-means clustering algorithm is also a sufficient algorithm for clustering as well however, it shows a lower performance than those of FCM. The values might vary due to the initial values that are chosen arbitrarily.

Our last algorithm is hierarchical clustering algorithm which is proven from its result that it is not sufficient enough to cluster our data. Especially, ambiguous data with perplexingly visible semantic gap. This is because the selection condition of this method is based on a binary tree algorithm.

To sum up, this research hopefully provides better understandings to its principle of each represented clustering algorithm and would furthermore develop a better searching and sorting algorithm for image retrieving models.

Acknowledgment

The author would like to thank Assoc. Prof. Dr. John morris who support all facility for this work.

References

[1] Bo Gun Park, Kyoung Mu Lee, and Sang Uk Lee, " Color-Based Image Retrieval Using Perceptually Modified Hausdorff Distance", EURASIP Journal on Image and Video Processing, Vol. 2008

[2] Kristof Van Learhoven, "Combining the Self Organizing Map and K-Means Clustering for On-Line Classification of Sensor Data", Lecture Note in Computer Science, 2001, Volume 2130, Artificial Neural Network-ICANN 2001, Page 464-469.

[3] Jacob Goldberger, Shiri Gordon, and Hayit Greenspan, "Unsupervised Image Set Clustering Using an Information Theoretic Framework", IEEE Trans. Image Processing, Vol. 15. No.2, Feb, 2006.

[4] Yong-mao Wang, Zheng-guang Xu, "Image Retrieval Using the Color Approximation Histogram Based on Rought Set Theory", ICIECS 2009. International Conference on Information Engineering and Computer Science, 2009, pp. 1-4

[5] Graham Cormode, Andrew McGregor, "Approximation Algorithms for Clustering Uncertrain Data", PODS'08, Proceedings of the twenty seventh ACM SIGMOD-SIGACT-SIGART symposium on Principle of database systems, June 9-12, 2008 Vancouver, BC, Canada, ACM New York, NY, USA 2008.

[6] R.W.G.Hunt, "Measuring Color", Fountain Press England, 1988

[7] G.Hoffmann, "Color Order Systems RGB/HLS/HSB, http://www.fho-emden.de/Hoffman/hlscone03052001.pdf

[8] BikeshKr. Singh and bidyut Mazumdar, "Content Retrieval From X-RAY Image Using Color & Texture Features", International Journal of Electronics Engineering, 2(1), 2010, pp 25-28

[9] Chia Yu Yen, Krzysztof J. Cios, "Image recognition system based on novel measures of image similarity and cluster validity", Neurocomputing, Vol. 72, 2008, pp 401-412

[10] Chung Chian Hsu, Yan Ping Huang, "Incremental clustering of mixed data based on distance hierarchy" Expert System with Applications, Vol. 35, 2008, pp. 1177-1185.

[11] G.Padmavathi, Muthukumar, "Image segmentation using fuzzy c means clustering method with thresholding for underwater images", Int. J. Advanced Networking and Applications, Vol. 02, Issue 2, pp. 514-518, 2010

[12] Rumiana Krasterv, "Bulgarian Hand-Printed Character Recognition Using Fuzzy C-Means Clustering", Bulgarian Academy of sciences problems of engineering cybernetics and rototics, 53, 2002,pp. 112-117

[13] Y. Ye, Z. Ye, J. Luo, P. Bhattacharya, H. Majlesein, R. Smith, "On Linear and Nonlinear Processing of Underwater, Ground, Aerial and Satellite Images", IEEE International Conference on Systems, Man and Cybernetics, pp.3364-8, Oct.10-12, 2005.

[14] V.S.V.S. Murthy, E. Vamsidhar J., J.N.V.R. Swarup Kumar, P. Sankara Rao, "Content Based Image Retrieval using Hierarchy and K- Means Clustering Techniques", International Journal of Engineering Science and Technology, Vol.,2(3), 2010, pp. 209-212

[15] Ming-Ghao Chiang, Chun Wei Tsai, Chu Sing Yang, "A time efficient pattern reduction algorithm for k-mean clustering", Information Science: and International Journal, Vol.181, Issue 14 (July, 2011), pp. 716-731.

[16] Gustavo B. Borba, Humberto R. Gamba, Oge Marques and Liam M. Mayron, "An Unsupervised Method for Clustering Image Based on Their Salient Regions of Interest", Proceedings of the 14th annual ACM international conference on Multimedia, 2006

[17] N. Rashevsky, "Life information theory and topology", Bulletin of Mathematical Biophysics, Vol. 17 No.3, 1955, pp. 229–235