ISSN:0975-9646



# A Generic Frame Work for Image Data Clustering Via Weighted Clustering Ensemble

A.Lakshmi Lavanya , RamaSree Sreepada
*CSE Dept*
*Adithya Engineering College*
*Surampalem*

**Abstract: This paper has a further exploration and study of visual feature extraction. Image retrieval based on multi-feature fusion is achieved by using normalized Euclidean distance classifier. According to the HSV (Hue, Saturation, Value) color space, the work of color feature extraction is finished, the process is as follows: quantifying the color space in non-equal intervals, constructing one dimension feature vector and representing the color feature by cumulative histogram. Similarly, the work of texture feature extraction is obtained by using gray-level co-occurrence matrix (GLCM) or color co-occurrence matrix (CCM).using color feature ,and texture feature based algorithms clustering temporal data process is very complex. Combining the low-level visual features and high-level concepts, the proposed approach fully explores the similarities among images in database, using such clustering algorithm and optimizes the relevance results from traditional image retrieval system by firstly clustering the similar images in the images database to improve the efficiency of images retrieval system ,to reduce the complexity to cluster image data ,we propose a dynamic algorithm, weighted ensemble learning approach to image data clustering which combines image data .**

## 1. INTRODUCTION

The use of images in human communication is hardly new – our cave-dwelling ancestors painted pictures on the walls of their caves, and the use of maps and building plans to convey information almost certainly dates back to pre-Roman times. But the twentieth century has witnessed unparalleled growth in the number, availability and importance of images in all walks of life. Images now play a crucial role in fields as diverse as medicine, journalism, advertising, design, education and entertainment Firstly, we discuss the color and texture features separately. On this basis, a new method using integrated features is provided, and experiment is done on the real images, satisfactory result is achieved, verify the superiority of integrated feature than the single feature. Retrievals are useful only in the limited domain. The content and metadata based system gives images using an effective image retrieval technique. Many other image retrieval systems use global features like color, shape and texture. But the prior results say there are too many false positives while using those global features to search for similar images. Hence we give the new view of image retrieval system using both content and metadata.

## 2. MOTIVATION

After examining the issues involved in managing visual information in some depth, the participants concluded that images were indeed likely to play an increasingly important role in electronically- mediated communication. However, significant research advances, involving collaboration between a numbers of disciplines, would be needed before image providers could take full advantage of the opportunities offered. They identified a number of critical areas where research was needed, including data representation, feature extractions and indexing, image query matching and user interfacing. One of the main problems they highlighted was the difficulty of locating a desired image in a large and varied collection. While it is perfectly feasible to identify a desired image from a small collection simply by browsing, more effective techniques are needed with collections containing thousands of items. Journalists requesting photographs of a particular type of event, designers looking for materials with a particular color or texture, and engineers looking for drawings of a particular type of part, all need some form of access by image content. The existence – and continuing use – of detailed classification schemes such as ICONCLASS [Gordon, 1990] for art images, and the Opitz code [Opitz et al, 1969] for machined parts, reinforces this message.

## 3. MODEL DESCRIPTION:

This paper has a further exploration and study of visual feature extraction. Image retrieval based on multi-feature fusion is achieved by using normalized Euclidean distance classifier.

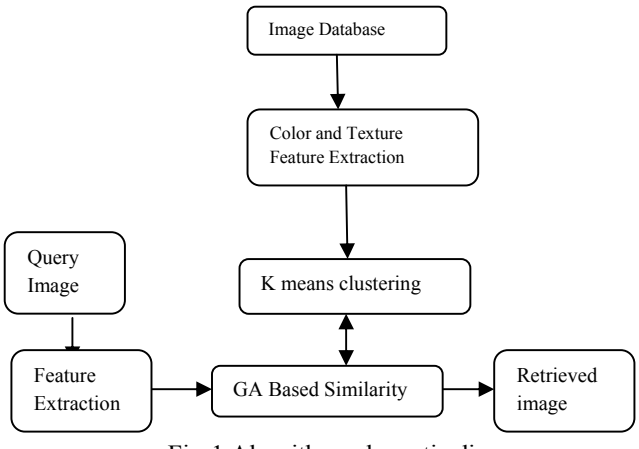


Fig-1 Algorithm schematic diagram

According to the HSV (Hue, Saturation, Value) color space, the work of color feature extraction is finished, the process is as follows: quantifying the color space in non-equal intervals, constructing one dimension feature vector and representing

the color feature by cumulative histogram. Similarly, the work of texture feature extraction is obtained by using gray-level co-occurrence matrix (GLCM) or color co-occurrence matrix (CCM).using color feature ,and texture feature based algorithms clustering temporal data process is very complex. To reduce the complexity of temporal data ,we propose a dynamic algorithm i.e. Genetic algorithm, weighted ensemble learning approach to temporal data clustering which combines temporal data As illustrated this model consists of four steps: quantization ,color and texture feature extraction , assigning weights ,generating clusters

**3.1 Feature extraction of HSV color**

HSV color space is widely used in computer graphics, visualization in scientific computing and other fields [5]. In this space, hue is used to distinguish colors, saturation is the percentage of white light added to a pure color and value refers to the perceived light intensity. The advantage of HSV color space is that it is closer to human conceptual understanding of colors and has the ability to separate chromatic and achromatic components.

**3.2Non-interval quantization**

Because of a large range of each component, if directly calculate the characteristics for retrieval, then computation will be very difficult to ensure rapid retrieval. It is essential to quantify HSV space component to reduce computation and improve efficiency. At the same time, because the human eye to distinguish colors is limited, do not need to calculate all segments. Unequal interval quantization according the human color perception has been applied on *H, S, and V* components. Based on the color model of substantial analysis, we divide color into eight parts. Saturation and intensity is divided into three parts separately in accordance with the human eyes to distinguish. In accordance with the different colors and subjective color perception quantification, quantified hue (H), saturation(S) and value (V) are showed as equation (3). In accordance with the quantization level above, the H, S, V three-dimensional feature vector for different values of with different weight to form one-dimensional feature vector named G [7]:

$$G = QSQV\,H + QV\,S + V \qquad (1)$$

Where *QS* is quantified series of S, *QV* is quantified series of V. Here we set$QS = QV = 3$, then

$$G = 9H + 3S + V \qquad (2).$$

$$H = \begin{cases} 0 \ if \ \ h \in [316,20] \\ 1 \ if \ \ h \in [21,40] \\ 2 \ if \ \ h \in [41,75] \\ 3 \ if \ \ h \in [76,155] \\ 4 \ if \ \ h \in [156,190] \\ 5 \ if \ \ h \in [191,270] \\ 6 \ if \ \ h \in [271,295] \\ 7 \ if \ \ h \in [296,315] \end{cases} \qquad S = \begin{cases} 0 \ if \ \ s \in [0,0.2) \\ 1 \ if \ \ s \in [0.2,0.7) \\ 2 \ if \ \ s \in [0.7,1) \end{cases}$$

$$V = \begin{cases} 0 \ if \ \ v \in [0,0.2) \\ 1 \ if \ \ v \in [0.2,0.7) \\ 2 \ if \ \ v \in [0.7,1) \end{cases}$$

In this way, three-component vector of HSV form one-dimensional vector, which quantize the whole color space for the 72 kinds of main colors. So we can handle 72 bins of one-dimensional histogram. This quantification can be effective in reducing the images by the effects of light intensity, but also reducing the computational time and complexity

**3.3Characteristics of color cumulative histogram**

Color histogram is derived by first quantize colors in the image into a number of bins in a specific color space, and counting the number of image pixels in each bin. One of the weaknesses of color histogram is that when the characteristics of images should not take over all the values, the statistical histogram will appear in a number of zero values. The emergence of these zero values would make similarity measure does not accurately reflect the color difference between images and statistical histogram method to quantify more sensitive parameters. Therefore, this paper represents the one-dimensional vector G by constructing a cumulative histogram of the color characteristics of image after using non-interval HSV quantization for G.

## 4. TEXTURE FEATURE EXTRACTION

**4.1 Texture feature extraction based on GLCM**

GLCM creates a matrix with the directions and distances between pixels, and then extracts meaningful statistics from the matrix as texture features. GLCM texture features commonly used are shown in the following:

GLCM is composed of the probability value, it is defined by $P(i,j|d,\theta)$ which expresses the probability of the couple pixels at $\theta$ direction and *d* interval. When $\theta$ and *d* is determined, $P(i,j|d,\theta)$ is showed by *Pi*, *j* . Distinctly GLCM is a symmetry matrix; its level is determined by the image gray-level. Elements in the matrix are computed by the equation showed as follow:

$$P(i,j|d,\theta) = \frac{P(i,j|d,\theta)}{\sum_i \sum_j P(i,j|d,\theta)} \qquad (4)$$

GLCM expresses the texture feature according the correlation of the couple pixels gray-level at different positions. It quantification ally describes the texture feature. In this paper, four features is selected, include energy, contrast, entropy, inverse difference.

$$\text{Energy} = E = \sum_x \sum_y p(x,y)^2 \qquad (5)$$

It is a gray-scale image texture measure of homogeneity changing, reflecting the distribution of image gray-scale uniformity of

Weight texture

$$\text{Contrast I} = \sum \sum (x-y)^2 p(x,y) \qquad (6)$$

Contrast is the main diagonal near the moment of inertia, which measure the value of the matrix is distributed and images of local changes in number, reflecting the image clarity and texture of shadow depth. Contrast is large means texture is deeper.

$$\text{Entropy} = \sum_x \sum_y S\ p(x,y) \log p(x,y) \qquad (7)$$

Entropy measures image texture randomness, when the space co-occurrence matrix for all values is equal, it achieved the minimum value; on the other hand, if the value of co-occurrence matrix is very uneven, its value is greater. Therefore, the maximum entropy implied by the image gray distribution is random.

Inverse difference= $\sum_x \sum_y \frac{1}{1+(x-y)^2} p(x,y)$ (8)

It measures local changes in image texture number. Its value in large is illustrated that image texture between the different regions of the lack of change and partial very evenly.

Here $p(x, y)$ is the gray-level value at the coordinate $(x, y)$.

**4.2 Feature extraction based on CCM**

Assuming color image is divided into $N \times N$ image sub-block, for anyone image sub-block $T(1 \le i \le N, 1 \le j \le N)$ Using the main color image extraction algorithm to calculate the main color $C(i, j)$. For any two 4-connected image sub-block$T(i,j)$and $T_{(k,j)}(|i-k|=1 and j=1; or\ |j-l|=1 and i=k)$ if its corresponds to the main color and in the HSV space to meet the following condition (1) *Cj* And *Ci* belong to the same color of magnitude, that is, its HSV components $hi = hj$ , $si = s\,j$ , $vi = v\,j$ ; (2) *Cj* And *Ci* don't belong to the same color of magnitude, but satisfy $si* 3 + vi = s\,j * 3 + v\,j$ , and $hi - hj = 1$; or satisfy $hi = hj$ , $si = s\,j$ and $vi$ , $v\,j \in \{0,1\}$ . We can say image sub-block $T(i, j)$ and $T(k, l)$ are color connected. According to the concept of color-connected regions, we can make each sub-block of the entire image into a unique color of connected set $S = \{Ri\}$ $(1 \le i \le M)$ in accordance with guidelines 4-connected. The set $S$ corresponds to the color-connected region.

For each color-connected region$\{Ri\}(1 \le i \le M)$ , the color components R, G in RGB color space and H in HSV color space are respectively extracted the CCM at the direction $\delta = 1;\theta = 0, 45, 90, 135$ . The same operation is done with i (intensity of the image). The statistic features extracted from CCM are as follows:

Energy $E = \sum_{i=1}^{D} \sum_{j=1}^{D} [m(i,j])^2$

Contrast $I = \sum_i^D \sum_j^D m(i-j)^2 * m(i,j)$

Entropy $S = \sum_i^D \sum_j^D m(i,j)^2 * \log[m(i,j)]$

Where， if $m(i, j) = 0$ ， $\log[m(i, j)] = 0$

$H = \sum_i^D \sum_j^D \frac{m(i,j)}{1+(i-j)^2}$

Through this method, we can get a 16 dimensional texture feature for component R, G, H and I, each component correspond to four statistic values E, I, S and H.

$F$=[$F$R,$F$G,$F$H,$F$I]=[$f$RE,$f$RI,$f$RS,$f$RI,…,$f$IE,$f$II,$f$IS,$f$IH]

4.3 K-MEANS CLUSTERING

Clustering algorithm has been widely used in computer vision such as image segmentation and database organization. The purpose of clustering is to group images whose feature vectors are similar by similarity judgment standard.

In the paper, we apply k-means algorithm to analysis images similarities in the database.

Let X= {xi}, i=1…n be the set of n d-dimensional points to be clustered into a set of k clusters= {ck, k=1….k}; k-means algorithm finds a partition such that the squared error between the empirical mean of a cluster and the points in the cluster is minimized. Let uk be the mean of cluster ck. The squared error between uk and the points in cluster ck is defined as

$$J(c_k) = \sum_{x_i \in c_k} || x_i - u_k ||^2 \quad (9)$$

The goal of K-means is to minimize the sum of the squared error over all K clusters,

$$J(c_k) = \sum_{k-1}^{k} \sum_{x_i \in c_k} || x_i - u_k ||^2 \quad (10)$$

Minimizing this objective function is known to be an NP-hard problem (even for k =2). Thus K-means, which is a greedy algorithm, can only converge to a local minimum, even though recent study has shown with a large probability k-means could converge to the global optimum when clusters are well separated (Meila, 2006).

K-means starts with an initial partition with k clusters and assign patterns to clusters so as to reduce the squared error. Since the squared error always decrease with an increase in the number of clusters k (with J(C) = 0 when k= n), it can be minimized only for a fixed number of clusters. The main steps of k-means algorithm are as follows:

(1) Select an initial partition with K=k clusters; repeat steps (2) and (3) until cluster membership stabilizes.

(2) Generate a new partition by assigning each pattern to its closest cluster center.

(3) Compute new cluster centers.

**4.4 Implementation of Genetic Algorithm**

To measure the degree of match between two feature sets *P* and *Q,* fitness function is constructed by the partial bidirectional Hausdroff distance. The output of GA is the feature set which has the highest value of fitness function.

***A. Chromosome Code***

The initial population is generated randomly. However, the generating scope of chromosomes is not arbitrary but limited within the image size. Each chromosome's detailed construction method is seriating of the feature sets and coding them into binary codes. Here binary coding is used because the image resolution is always limited. The feature sets can be represented by integers in the scope of the image resolution

***B. Fitness Function***

Clearly, if the match degree between *P* and *Q* can be measured, it is equivalent to evaluating the fitness of the chromosome [12]. Considering the partial bidirectional Hausdorff distance between feature sets *P* and *Q*, the smaller the distance is the match degree between *P* and *Q* is larger.

So, the fitness function can be selected as the inverse of the partial bidirectional Hausdorff distance

$$\textbf{fitness}=\frac{1}{a+H_{LK}(p,q)} \qquad (11)$$

Where $a$ is a positive constant. In order to avoid the denominator is a zero, the partial directional Hausdorff distance is added $a$.

### *C. Genetic Operators*

A pair of chromosome is randomly chosen from the population and is used as the parents to reproduce the offspring. The selection principle is the more chromosome number of its new offspring to the next generation, the bigger fitting function value *Fi* with the larger probability *pi.* The nature choice phenomenon of the biosphere is simulated by

$$P_S=\frac{F_i}{\sum_{j=1}^{l} F_j} \qquad (12)$$

Crossing operator is obtained after resetting the parent's facts which are got the selection processing. Crossing processing is performed according to a certain probability, which is called crossing probability *pc.* The crossing effect is to produce better chromosome after the combination of the generating materials of the parents. Here, single point crossover operator is adopted. Mutation operates on each binary bit of a chromosome in another predefined probability *pm;* the mutation operator is realized by reversing the value of the current binary bit, i.e. 0-1, 1-0.

### *D. Genetic Algorithm Based Feature Match*

For two features sets *P* and *Q*, determine population size *N*, crossing probability *pc*, mutation probability *pm*, fractions *fL* and *fK* of the partial bidirectional Hausdorff distance and the maximum iterative steps *Gmax*

*Step* 1: Randomly generate *N* feature sets in the test image and then convert them into chromosomes for initial generation.

*Step* 2: Evaluate the fitness of each chromosomes in current population and then create a new population by repeating following steps until the new population is complete.

1) Select two parent chromosomes from a population according to their fitness (the better fitness, the bigger chance to be selected).

2) With a crossover probability cross over the parents to form new offspring (children). If no crossover was performed, offspring is the exact copy of parents.

3) With a mutation probability mutate new offspring at each locus (position in chromosome).

4) Place new offspring in the new population.

*Step* 3: Use new generated population for a further run of the algorithm.

*Step* 4: If the end condition is satisfied, ***stop***, and return the best solution *T*best in current population, where *T*best be the feature sets determined by the best chromosome, and feature sets *T*best (*P*) and *Q* according to the simple nearest neighbor rule. If the maximum iterative step *Gmax* is not reached, go to *Step* 2.

### Experiment Results

In order to show performance in image retrieval system, we design a series of test on the clustering performance. Fig.2a shows the query image and color feature extraction. Based on the color features similar images are displayed on the screen, the color features are extracted based on histogram of input image. Fig 2b shows the histogram of input image Fig 3 shows the color+grarscale feature extraction of image retrieval, fig 4 shows the color+ccm image retrieval Fig-5 shows the GA based image retrieval k-means clustering algorithm in the images retrieval can throw away some images that are visually irrelevant to the query image for reducing images retrieval space. A leaf image is displayed in the firstly retrieval results in While through doing k-means clustering analysis for image retrieval,

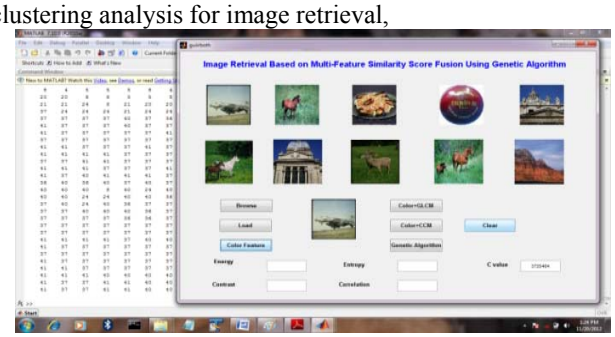


**Fig-2a**

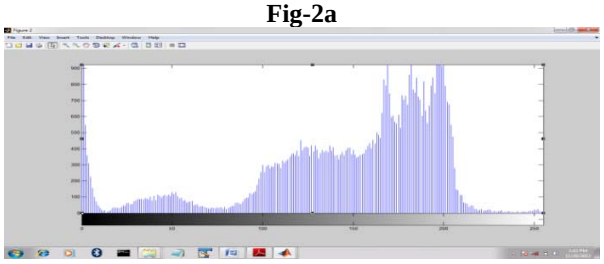

**Fig-2b**

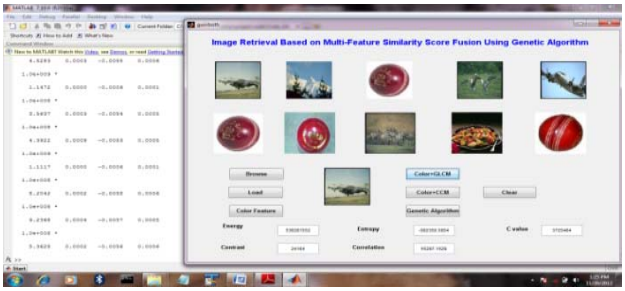


**Fig-3**


**Fig-4**

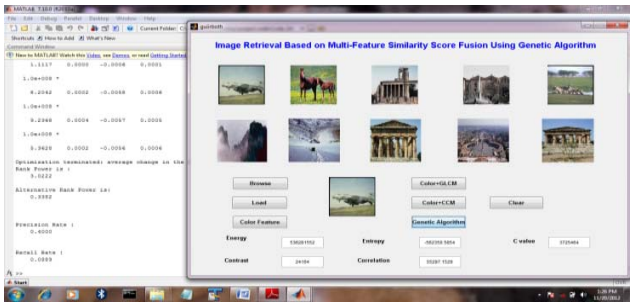


**Fig-5**

## CONCLUSION

In this paper we considered feature based clustering algorithm to cluster image data. This paper has a further exploration and study of visual feature extraction. Image retrieval based on multi-feature fusion is achieved by using normalized Euclidean distance classifier. According to the HSV (Hue, Saturation, and Value) color space, the work of color feature extraction is finished. Experiment results show that the efficiency and effectiveness of k-means algorithm in analyzing image clustering, which also can improve the efficiency of image retrieving and evidently promote retrieval precision. This k-means algorithm independent on the feature extraction algorithm is used as a post-processing step in retrieval. The improvements in selecting neighborhood vertices of the retrieval results from tradition image retrieval system in image feature space could also improve the recall rate.

## REFERENCES

[1]Rui, Y., Huang, T.S., Mehrotra, S. [Sharad],"Retrieval with relevance feedback in MARS", In Proc of the IEEE Int'l Conf. on Image Processing, New York, pp. 815-818, 1997.

[2] H. T. Shen, B. C. Ooi, K. L. Tan, "Giving meanings to www images" Proceedings of ACM Multimedia, pp. 39-48, 2000.

[3] B S Manjunath, W Y Ma, "Texture feature for browsing and retrieval of image data", IEEE Transaction on PAMI, Vol 18, No. 8, pp.837-842 1996.

[4] Y. Rui, C. Alfred, T. S. Huang, "Modified descriptor for shape representation, a practical approach",In: Proc of First Int's workshop on Image Database and Multimedia Search, 1996.

[5] Cao LiHua, Liu Wei, and Li GuoHui, "Research and Implementation of an Image Retrieval Algorithm Based on Multiple Dominant Colors", Journal of Computer Research & Development, Vol 36, No. 1, pp.96-100, 1999.

[6] Temporal Data Clustering via Weighted Clustering Ensemble with Different Representations Yun Yang and Ke Chen, Senior Member, IEEE Transactions on knowledge and Data Engineering, vol. 23, no. 2, February 2011

[7] E. Keogh and S. Kasetty, "On the Need for Time Series Data Mining Benchmarks: A Survey and Empirical Study," Knowledge and Data Discovery, vol. 6, pp. 102-111, 2002.