International Journal of Computer Applications Technology and Research
Volume 6–Issue 8, 405-409, 2017, ISSN:-2319–8656

Image Indexing Using Color Histogram and K-Means Clustering for Optimization CBIR

Juli Rejito
Department of Computer Science
Padjadjaran University
Bandung, Indonesia

Deni Setiana
Department of Computer Science
Padjadjaran University
Bandung, Indonesia

Rudi Rosadi
Department of Computer Science
Padjadjaran University
Bandung, Indonesia

Abstract: Retrieving visually similar images from image database needs high speed and accuracy. The researchers are investigating various text and content based image retrieval techniques to match the image features accurately. In this paper, a content-based image retrieval system (CBIR), which computes colour similarity among images, is presented. CBIR is a set of techniques for retrieving semantically relevant images from an image database based on automatically derived image features. The colour is one important visual elements of an image. This document gives a brief description of a system developed for retrieving images similar to a query image from a large set of distinct images with histogram colour feature based on image index. Result from the histogram colour feature extraction, then using K-Means clustering to produce the image index. Image index used to compare to the histogram colour element of a query image and thus, the image database is sorted in decreasing order of similarity. The results obtained by the proposed system apparently confirm that partitioning of image objects helps in optimization retrieving of similar images from the database. The proposed CBIR method is compared with our previously existed methodologies and found better in the retrieval accuracy. The retrieval accuracy is comparatively good than previous works offered in CBIR system.

Keywords: *CBIR, Image Features, Colour Histogram, K-Means clustering*

1. INTRODUCTION

In CBIR, the feature vector is extracted from images. Query feature vector is matched with the stored feature vector on one to one basis, which resulted in slow down the processing time. To improve the speed of executing and better result, many researchers are paying attention to uses the clustering algorithm for CBIR. Clustering is a process of separating a data set into groups in such a way that the object in one group is more similar to those objects in the other group [1].

Chin-Chin Lai et.al. [2] have proposed an interactive genetic algorithm (IGA) to reduce the gap between the retrieval results and the users' expectation called semantic gap. They have used HSV color space that corresponds to the human way of perceiving the colors and separate the luminance component from chrominance ones. They have also used texture features like the entropy based on the grey level co-occurrence matrix and the edge histogram. They compared this method with others approaches and achieved better results.

Kannan A, et al [3] have proposed Clustering and Image Mining Technique for fast retrieval of Images. The primary objective of the image mining is to remove the data loss and to extract the meaningful information to the human expected needs. The images are clustered based on RGB Components, Texture values and Fuzzy C mean algorithm.

Chakravarti and Meng [4] have published a paper on Color Histogram Based Image Retrieval. They have used color histogram technique to retrieve the images. This method allows retrieval of images that have been transformed regarding their size as well as translated through rotations and flips.

Kumar, et.al [5] have published on Content Based Image Retrieval using Color Histogram. They have used Color Histogram technique to retrieve the similar images. To speed up the recovery, they have used the proposed grid-based indexing to obtain the nearest neigh hours of the query image, and accurate images are retrieved. Indexing can be performed in vector space to improve retrieval speed. Mainly, they have implemented CBIR using color histogram technique and refined with the help of grid method to improve the image retrieval performance.

CBIR with clustering algorithm is an alternative that is expected to improve the performance and accuracy of searches in the query image. K-means is the core clustering algorithm, but in this case, k-means is very sensitive for first grouping and delicate to outliers and noise [1,6]. Also, to form the clusters, K-mean depends on initial condition, which causes the algorithm to give a suboptimal solution. As compare to k-means ant colony algorithm are more prominent for initializing the cluster. Due to the optimal global nature of particle swarm optimization algorithm give the optimum solution for clustering. These most prominent features of ant colony algorithm and particle swarm optimization are used for clustering.

The implementation of a query optimization proposed in this paper was related to CBIR in image databases aimed at obtaining the image in the database with a high image content similarity level during an image searching process. In this proposal, proposed by carrying out a color extraction process of each image database, a method of clustering with the k-means algorithm, and results of clustering of each image database were used to a filtering process based on query images.

2. LITERATURE REVIEW

2.1 CBIR

CBIR is a method that is used to look at image features like (color, shape, texture) to find a query image from the database. The difficulties of CBIR lie in reducing the differences of contents based feature and the semantic based

functions. This problem in giving useful retrieval images and channelize the researchers to use (CBIR) system, to take global color and texture features to achieve, the right recovery, where others used local color and texture features [6]. The method in [7] presented the holistic representation of spatial envelope with a very low dimensionality for making the incident image.

The method in [8] proposed a modern approach for image classification with the open field design and the concept of over-completeness methodology to achieve an excellent result. As reported in [8], this method produced the best classification performance with much lower feature spatiality compared to that of the former schemes in image classification task. For similarity search [9] the user needs to enter keywords along with the query image that might appear in the text of patents. Higher average precision and recall rates compared to the traditional Dominant Color method were obtained successfully [10]. The texture and color attributes are computed in a way that model the Human Vision System (HSV) [11].

2.2 Color Histogram Feature Extraction

Obtained by extraction of the color histogram feature extraction of image pixels for each color component R, G, and B then calculated the frequency of each color index from 0 to 255 and raised in the form of histogram value for each color component, and is written as a vector shown in the equation as follows:

$$H = \{H[0], H[1], H[2], H[3], \dots, H[i], \dots, H[n]\} \quad (1)$$

Where i is the color in the color histogram storage and H [i] indicates the number of pixels of color i in the image, and n is the number of colors used in the storage of the color histogram.

Results histogram value for each color component (x, y, z) of the image is sought (H_q) and record images (H_i) then calculate resemblance to calculate the distance to the known color histogram. Histogram Intersection Technique (HIT) [12], using the formula in the equation as follows:

$$S(H_q, H_i) = \frac{\sum_{x \in X, y \in Y, z \in Z} \min\left(H_q(x,y,z), H_i(x,y,z)\right)}{\sum_{x \in X, y \in Y, z \in Z} H_q(x,y,z)} \quad (2)$$

Possess the formula distance values tend to have small differences so that the formula developed into the equation as follows:

$$S(H_q, H_i) = \frac{\sum_{x \in X, y \in Y, z \in Z} \min\left(H_q(x,y,z), H_i(x,y,z)\right)}{\min\left[\sum_{x \in X, y \in Y, z \in Z} H_q(x,y,z), \sum_{x \in X, y \in Y, z \in Z} H_i(x,y,z)\right]} \quad (3)$$

2.3 K-Means Algorithm

The k-means algorithm is effective in producing good clustering results for many applications [13]. The reasons for the popularity of k-means are ease and simplicity of implementation, scalability, speed of convergence and adaptability to sparse data [14]. K-means clustering is a partitioning clustering technique in which clusters are formed with the help of centroids. From these centroids, groups can vary from one another in different iterations. Also, data elements can differ from one cluster to another, as groups are based on the random numbers known as centroids [15]. The k-means algorithm is the most extensively studied clustering algorithm and is effective in producing good results. The k-means algorithm is computationally expensive and requires time proportional to the product of the number of data items, the number of clusters and the number of iterations. K-means is formally described by Algorithm 1.

Algorithm 1: Basic K-means algorithm.
1: Select K points as initial centroids.
2: **repeat**
3: Form K clusters by assigning each point to its closest centroid.
4: Recomputed the centroid of each cluster.
5: **until** Centroids do not change.

2.4 Measurement Image Quality

Measurements with this model are widely used because of the ease of calculation, have a physical understanding and mathematically easy to use for optimization purposes, although not very useful in matching visual quality. This measurement consists of several formulas that MAE (Maximum Absolute Error), MSE (Mean Square Error), RMSE (Root Mean Square Error), SNR (Signal to Noise Ratio), and PSNR (Peak Signal to Noise Ratio) [16].

2.4.1 Maximum Absolute Error (MAE)

MAE is the highest value of the absolute value difference between the input image $f(x,y)$ and the output image $g(x,y)$. *MAE* calculation is mathematically written in equation form:

$$MAE = max|f(x,y) - g(x,y)| \quad (4)$$

where $f(x,y)$ is the value of the initial / original image intensity in the position *(x, y)* and $g(x,y)$ is the value of the image's intensity in the position (x, y).

2.4.2 Mean Square Error (MSE)

MSE is the average value of squared error between the input image $f(x,y)$ to the output image $g(x,y)$, where both of the images have the same values. Good MSE values are where the value near zero (*MSE* ≈ 0) [17]. Mathematically *MSE* value calculation using the equation

$$MSE = \frac{1}{MN} \sum_{x=1}^{M} \sum_{y=1}^{N} [f(x,y) - g(x,y)]^2 \quad (5)$$

where M, N are the width and height of the image, $f(x,y)$ is the initial image intensity use values/position (x, y) and $g(x,y)$ native to the image of the intensity value at the position (x, y).

2.4.3 Peak Signal to Noise Ratio (PSNR)

PSNR is a value of the ratio between the maximum image reconstruction results with square root value of *MSE* or equal to the value of *RMSE* [17]. For 8-bit image pixel, the maximum value is 255. The criteria of image quality will be better if the result of the greater *PSNR* values and mathematically generated from *MSE* value shown in the equation:

$$PSNR = 10\ log_{10}\left[\frac{255^2}{MSE}\right] dB \quad (3)$$

3. PROPOSED APPROACH

To solve the above problem, the necessary stages of completion as shown in Figure 1.

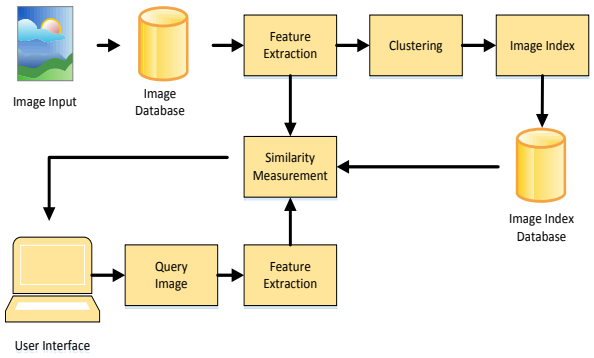


Figure 1. Architecture Systems

From the figure 1 can be explained the sequence of the process is

1. Put the whole picture of WANG database to a database of images. Using equation 1 did the color feature extraction process to obtain the value vector R, G, B based on a base image and calculate the difference in distance using the formula of HIT in equation 2 and 3.
2. Result values $S(H_q, H_i)$ was used to classify the image by using the k-means algorithm, Save the cluster for each image as the image index.
3. To perform image searches done by inserting the image of the user interface applications that have been created in the form of a query image. By using the image index was then determined the position of the cluster image and serve as a filter record by his cluster group.
4. Using equation 4 calculating the similarity between the query image and the image record.

4. RESULTS AND DISCUSSION

The database image used in proposed method is WANG The database, and the Delphi program has been implemented. The WANG database contains 1,000 images in JPEG format. The size of each image is 256x386 and 384x256. It consists of 10 classes such as (Africa, beach, monuments, buses, dinosaurs, elephants, flowers, horses, mountains and food) each class contains 100 images. Figure 2 shows the sample of WANG database.

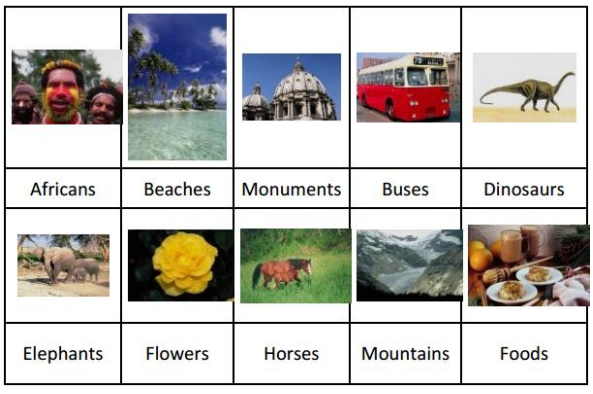


Figure 2. Example Image from each of the 10 Classes of WANG Database

The retrieval effectiveness of the proposed method is measured by using 30 different query images from each class. It is tested for 300 query images. In the beginning, the image size is changed to [128,128] to get a similar size for each image. The color features are extracted according to HSV histogram values. Euclidean distance is used to measure the similarity between query image and database images. Top 100 images are retrieved depending on the minimum distance. Next stage, the shape features are extracted from 100 images that resulted from the first stage and similarity measurements is performed between shape features of query image and the 100 images. Top 50 images are retrieved depending on the minimum distance. The final stage in the proposed method is extracted the first order features from 50 images and compared with the query image. The nearest ten images of the query image are retrieved from the image database.

4.1 The Result of K-Means

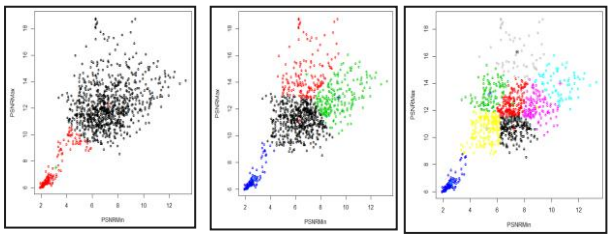
The implementation of the clustering was applied in several cluster groups; namely, 2 clusters, 4 clusters, 8 clusters, and the amounts of iteration of each cluster and the values of minimum PSNR and maximum PSNR for each cluster were shown in Table 1 and Figure 3.

Table 1. Clustering of 1000 WANG Database Records in 2, 4, 8 clusters

Cluster		Centroid PSNR (dB)		Record Count
Cluster	N	Minimum	Maximum	
2	1	7.238144	12.298757	838
2	2	3.217335	7.589639	162
4	1	6.318496	11.190640	529
4	2	6.894847	14.625369	155
4	3	2.638533	6.818268	119
4	4	9.449756	12.881852	197
8	1	10.119414	13.778683	110
8	2	8.000599	11.858669	209
8	3	5.688476	11.963560	188
8	4	6.923792	10.427970	147
8	5	7.305264	17.609415	24
8	6	4.749914	9.994102	95
8	7	6.782889	13.985441	121
8	8	2.518710	6.560355	106

Table 1 shows that the record is processed in WANG database of 1,000 records with the results of the process for the second cluster to cluster-1 has a value of 7.238144 dB PSNR minimum centroid and centroid PSNR maximum of 12.298757 dB with a record number of 838, while the cluster 2 has a value of 3.217335 dB PSNR minimum centroid and centroid PSNR maximum amounted to 7.589639 dB with an unprecedented number of 162. Likewise, for the formation of 4 and 8, as shown in Table 1.

Figure 3 shows the graph plots each record based on the minimum and maximum PSNR value on the formation of 2, 4, and 8 clusters using the K-Means algorithm database. Each cluster in the plot with a different color, for example in the formation of two clusters to cluster groups to plot-1 is shown in black, while for cluster 2nd plot shown in red. It is also shown at 4 and 8 clusters.



2 Clusters 4 Clusters 8 Clusters

Figure 3. The plot for K-Means Clustering in 2,4, and 8 clusters for WANG Database.

4.2 The Result of Precision

The precision and recall are used to measure the performance of retrieval. The recall is used to measure the system's ability to retrieve all the images that are relevant, while precision is used to measure the system's ability to extract only the images that are relevant. The equation of the recall and precision are illustrated in the following:

$$precision = \frac{Number\ of\ relevant\ images\ retrieved}{the\ total\ number\ of\ images\ retrieved} = \frac{N}{N+B} \quad (6)$$

$$recall = \frac{Number\ of\ relevant\ images\ retrieved}{the\ total\ number\ of\ relevant\ retrieved} = \frac{N}{N+C} \quad (7)$$

Where the number of retrieved images is represented by N and the irrelevant images are represented by B, while the C accounts for the number of relevant images not retrieved. The similarity measurement used in this work is Euclidean distance. Table 1 shows the results of precession when using color, shape and texture features. The best results are obtained in using cascade color, form and texture features.

Table 1 shows the results of precession value group by categories obtained by my proposed method (histogram color feature with 5 clusters) with other retrieval systems: Histogram color, GLCM (Gray-level co-occurrence matrix) texture, histogram color + GLCM texture, and histogram color + GLCM texture with sub-block. The best results are obtained in using histogram color feature with 5 clusters.

Table 2. The result of precession value group by categories obtained by proposed method with other retrieval systems

No	Categories	Histo gram Color	GLCM Texture	Hist.Clr - GLCM Texture	Hist.Clr+ GLCM Texture +sub-block	Histogram Color with 5 Clusters
1	Africa	0.36	0.21	0.34	0.41	0.92
2	Beaches	0.27	0.35	0.21	0.32	0.42
3	Building	0.38	0.50	0.24	0.37	0.38
4	Bus	0.45	0.22	0.51	0.66	0.52
5	Dinosaur	0.26	0.29	0.39	0.43	1.00
6	Elephant	0.30	0.24	0.26	0.39	0.68
7	Flower	0.65	0.73	0.81	0.87	0.92
8	Horses	0.19	0.25	0.28	0.35	0.97
9	Mountain	0.15	0.18	0.20	0.34	0.25
10	Food	0.24	0.29	0.25	0.31	0.78
Average		**0.33**	**0.33**	**0.35**	**0.45**	**0.68**

In Table 5 shows that the average value of precision in a row are 0.33, 0.33, 0.35, 0.45 and 0.68. Value 0.68 is the greatest value of the proposed proposal.

Figure 4 shows the precession obtained by the different methods. It indicates that the histogram color feature with 5 clusters gave the best result in image retrieval.

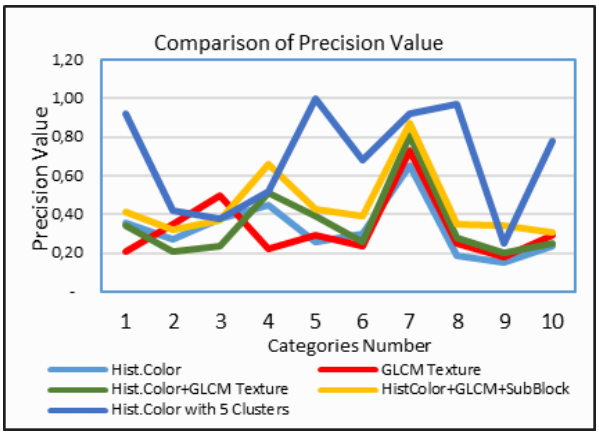


Figure 4. The Result of Precision Value Group by categories obtained by proposed method with other recovery systems

5. CONCLUSIONS

Cluster initialization process would be the main key in an information retrieval process of a query using PSNR minimum and PSNR Maximum. For each record in WANG database used as a base of record order, and then the distance between clusters was determined by computing total record divided by the number of clusters and ended by establishing each cluster that was taken from ordered records based on changes in their distances. The results in WANG database by using color histogram that was taken randomly by an amount of 1,000 showed that highest level of accuracy in 5 clusters shown on the precision value of 0.68.

6. REFERENCES

[1] Rejito J., Wardoyo R., Hartati S., Harjoko, 2012, A.,"Optimization CBIR using K-Means Clustering for Image Database," International Journal of Computer Science and Information Technologies, Vol. 3 (4), 4789-4793.

[2] Chin-Chin Lai, and Ying - Chuan Chen, ‖ A, 2011, "User-Oriented Image Retrieval System Based on Interactive Genetic Algorithm," IEEE transactions on instrumentation and measurement, vol. 60, no. 10,

[3] Kannan, A, Mohan, V., Anbazhagan, N., 2010, "Image Clustering and Retrieval using Image Mining Techniques" IEEE Conference.

[4] Chakravarti, R, and Meng, X, 2009. "A Study of Color Histogram Based Image Retrieval," Sixth International Conference on Informational Technology.

[5] Kumar A.R, and Saravanan, D., 2013 "Content Based Image Retrieval Using Color Histogram," International Journal of Computer Science and Information Technologies, Vol. 4 (2), 242–245.

[6] Datta R., Joshi D., Li J., and Wang J. Z., 2008, "Image Retrieval," ACM Comput. Surv., Vol. 40, no. 2, Apr. pp. 1–60.

[7] Oliva A., and Torralba A, 2001, “Modeling the shape of the scene: “A holistic representation of the spatial envelope,” Int. J. Comput. Vis., vol. 42, no. 3, pp. 145 – 175.

[8] Jia Y., Huang C., and Darrell T., 2012, “Beyond spatial pyramids: “Receptive field learning for pooled image features,” in Proc. IEEE Conf. Comput. Vis.Pattern Recognit. (CVPR), pp. 3370–3377.

[9] Tiwari A. and Bansal V., 2004, "PATSEEK: Content Based Image Retrieval System for Patent Database," Proceedings of international conference on

electronic business, pp. 1167-1171.

[10] Krishnan N., Banu M.S., and Callins C., 2007, "Content Based Image Retrieval Using Dominant Color Identification Based on Foreground Objects," International Conference on Computational Intelligence and Multimedia Applications, Vol. 3, pp. 190-194.

[11] Ahmed H.A., Gayar N.E., Onsi H., 2008, “A New Approach in Content-Based Image Retrieval Using Fuzzy Logic” INFOS.

[12] Smith, J. R., 1997, “Integrated Spatial and Feature Image Systems: Retrieval, Analysis, and compression." Ph.D. thesis, Columbia University, New York, NY.

[13] Nikman T.A, Bakbak A., 2009, "An efficient hybrid approach based on PSO, ACO, and K-Means for Cluster Analysis," Elsevier.

[14] Ghosh A., Parikh J., Sangar V. and Haritsa J., 2002, "Query Clustering for Plan Selection, Tech Report," DSL/SERC, Indian Institute of Science.

[15] Kumar, M. Varun, M. Chaitanya V., and Madhavan, M., 2012. "Segmenting the Banking Market Strategy by Clustering." International Journal of Computer Applications 45.

[16] Ahmed H.A., Gayar N.E., Onsi H., 2008, “A New Approach in Content-Based Image Retrieval Using Fuzzy Logic” INFOS.

[17] Datta R., Joshi D., Li J., and Wang J. Z., 2008, "Image Retrieval," ACM Comput. Surv., Vol. 40, no. 2, pp. 1–60.