A Content-Based Image Retrieval Method By Exploiting Cluster Shapes

Hanan Al-Jubouri
Computer Engineering Department
Mustansiriyah University
Ali Taleb Street, Baghdad, Iraq
hananaljubouri @uomustansiriyah.edu.iq

Hongbo Du
Applied Computing Department
Buckingham University
MK18 1EG, Buckingham, UK
hongbo.du @buckingham.ac.uk

Abstract *Content-Based Image Retrieval (CBIR) is an automatic process of retrieving images that are the most similar to a query image based on their visual content such as colour and texture features. However, CBIR faces the technical challenge known as the semantic gap between high level conceptual meaning and the low-level image based features. This paper presents a new method that addresses the semantic gap issue by exploiting cluster shapes. The method first extracts local colours and textures using Discrete Cosine Transform (DCT) coefficients. The Expectation-Maximization Gaussian Mixture Model (EM/GMM) clustering algorithm is then applied to the local feature vectors to obtain clusters of various shapes. To compare dissimilarity between two images, the method uses a dissimilarity measure based on the principle of Kullback-Leibler divergence to compare pair-wise dissimilarity of cluster shapes. The paper further investigates two respective scenarios when the number of clusters is fixed and adaptively determined according to cluster quality. Experiments are conducted on publicly available WANG and Caltech6 databases. The results demonstrate that the proposed retrieval mechanism based on cluster shapes increases the image discrimination, and when the number of clusters is fixed to a large number, the precision of image retrieval is better than that when the relatively small number of clusters is adaptively determined.*

***Index Terms*—Content-Based Image Retrieval, Semantic gap, EM/GMM clustering algorithm, Discrete Cosine Transform (DCT), Kullback-Leibler divergence**

I. INTRODUCTION

Early systems of image retrieval use a textual annotation such as keywords or phrases to index images in a database. A user searches images by entering the textual annotation, and the system ranks images in a list based on the degree of match to the annotation. However, such an approach suffers from certain constraints. For example, it is infeasible to annotate images in a large database manually. Text annotations may not be always available at the time of image capture for various reasons. Even when a descriptive text for the image can be obtained, subjective interpretations of the image content may lead to inconsistencies in the annotation. Consequently, Content-Based Image Retrieval (CBIR) using visual content to index images automatically is still attracting the attentions of researchers from different fields [1].

As a result of extensive research in CBIR [2, 3, 4] in the last two decades, several CBIR systems, such as QBIC [5], VisualSEEK [6], BlobWorld [7], and Google Similar Image Search [8] have been produced. In addition, different approaches are developed to reduce or narrow the semantic gap between high level conceptual meaning and the low-level content-based features. Clustering [9], Region of Interest (ROI) [10], Relevance Feedback (RF) [11], Bag of Visual Words (BOVW) [12] and Browsing [13] are the main existing approaches each of which is an active research area by its own right. All the approaches involve using feature extraction and similarity measures to retrieve the most similar images in a ranked list. In the clustering approach, an algorithm for cluster detection is deployed to group feature vectors into clusters based on similarity function. Algorithmic considerations, data, and cluster characteristics are factors that affect the effectiveness of the process. In fact, different categories of clustering algorithms such as Prototype-based (e.g. *k*-means), model-based (e.g. EM/GMM), density-based (e.g. the mean shift), and graph-based (e.g. the normalized Laplacian spectral clustering) have been used in practice [14]. In the ROI approach, an interested area is firstly defined, and features are then extracted to index images. Involving the end user in specifying the ROI during a retrieval session is one major limitation. The principle of the FR approach is to continuously refine the retrieved images interactively between the end user and the CBIR system in determining positive or negative (relevant/irrelevant) images, which may cause encumbrance to the user. The idea of BOVW was borrowed from the field of information retrieval where documents are represented by bags of vocabulary/words (BOW). In CBIR, images are divided into patches from where visual features are extracted and then quantized by using a clustering algorithm, and the resulting clusters correspond to vocabularies and their centroids to words. Many clusters may be needed for good retrieval results, an issue of concern of this approach regarding retrieval efficiency. While most of the approaches mentioned above make a query by issuing an example image, an alternative browsing approach uses effective tools to navigate through many images and select the ones of the interest. The challenge here is how to visualize the whole or part of the image collections and how to provide an effective and efficient mechanism to navigate through many database images.

A more recent work investigated a new direction in using deep learning to reduce the semantic gap issue for CBIR [15]. Convolution neural networks were used to directly learn feature representations from image contents. The method was empirically tested on several large image databases such as ImageNet, Pubfig83LFW, Caltech256, Oxford, etc. and the test results have showed better levels of retrieval accuracy over the conventional methods. Some existing works are concerned more with the retrieval efficiency, and explored the use of various efficient tree structures such as B-tree, R^+-tree, KD^+-tree, etc. [16, 17, 18] to speed up the search. However, such solutions suffer from issues regarding storage requirement and difficulties in handling high-dimensional data. Hashing methods deal with those downsides by storing compact binary codes that represent the original data and use Hamming distance to locate efficiently similar neighbors [19].

This paper presents a new method that addresses the semantic gap issue by exploiting *cluster shapes*. The method first extracts Discrete Cosine Transform (DCT) coefficients in local areas of the image as the basic features representing local colour and texture. The Expectation-Maximization Gaussian Mixture Model (EM/GMM) clustering algorithm is then applied to the local feature vectors to obtain clusters of various shapes and sizes based on the local features. At the image retrieval stage, the proposed method uses a proximity measure based on the Kullback-Leibler divergence principle that measures dissimilarity between cluster shapes of two images. The paper further investigates two respective scenarios of clustering: (a) when the number of clusters is fixed to a specific value, and (b) when the number of clusters is adaptively determined according to cluster quality. Experiments are conducted on publicly available WANG and Caltech6 databases. The results show that cluster shape based retrieval further increases the image discrimination with higher level of precision than similar works in the past [14] where only the centroids of the clusters were considered. The paper discovers that the precision of image retrieval improves as the number of clusters increases. The level of precision of a large fixed number of clusters tends to be better than that when the number of clusters is adaptively determined according to cluster quality alone.

The rest of this paper is organized as follows. Section 2 reviews the basics that will be utilized in the proposed method. Section 3 presents the details of the proposed method and explains the rationale behind it. Section 4 evaluates the effectiveness of the proposed method. Section 5 discusses the effects of the number of clusters to the retrieval precision. Section 6 concludes the paper and outlines future works at the next stage of research.

II. RELATED WORK

A. Discrete Cosine Transform (DCT)

DCT is one of many transformation methods. At the heart of DCT is the following operation that is executed iteratively on the pixel intensity values of an 8 x 8 block window on the image:

$$C(u,v) = \frac{1}{4}k(u)k(v)\sum_{i=0}^{7}\sum_{j=0}^{7} f(i,j)\cos\left(\frac{(2i+1)u\pi}{16}\right)\cos\left(\frac{(2j+1)v\pi}{16}\right) \quad (1)$$

$$k(u),k(v) = \begin{cases} 1/\sqrt{2} & \textit{if } u \textit{ and } v = 0 \\ 1 & \textit{otherwise} \end{cases}$$

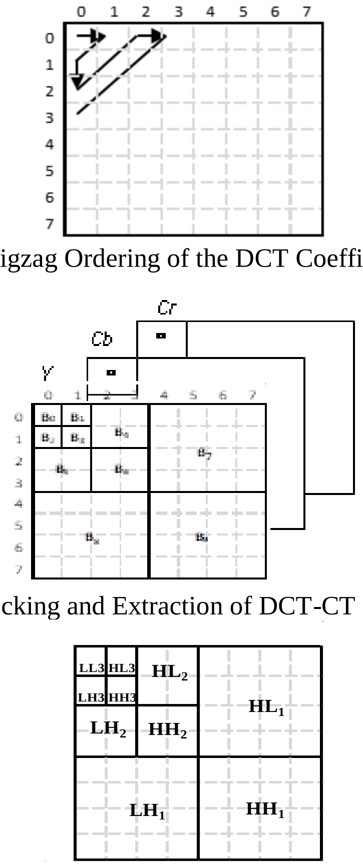
where 0 ≤ u, v ≤ 7 and *f*(i, j) is the pixel intensity value at location i, j. C(0, 0) is known as a low frequency DC and the remaining as high frequency ACs. The DC coefficient captures the average colour intensity of the block whereas the AC coefficients represent colour intensity variations, i.e. the textures of the block.

For colour images, DCT is applied to each separate colour channel. DCT has been used to extract low level image content features in the frequency domain [20, 21, 22] from the *YCbCr* colour space. Although DCT features have also been extracted from other colour spaces such as *RGB*, *YCgCb*, *YUV*, *YIQ*, *XYZ*, and *LUV*, it has been established that the DCT features from the *YCbCr* space is most effective [23].

DCT coefficients can be exploited in different orders. A commonly used order is a zigzag manner as depicted in Fig.1 (a), where the coefficients are arranged from low to high frequencies. However, a feature vector with all DCT coefficients, either taken in any order has deficiency of robustness due to the vector's high dimensionality [24], and potential vulnerability for *over-fitting*, i.e. the feature vector has too much specific details of a local block. The work presented in [26] divides the 8x8 block in the Y channel into sub-blocks as depicted in Fig.1(b). Then, the standard deviations of the coefficients in B_4, B_5, B_6, B_7, B_8 and B_9 sub-blocks are calculated. A 12 dimensional feature vector, i.e. ($C_Y(0,0)/8$, $C_{Cb}(0,0)/8$, $C_{Cr}(0,0)/8$, $C_Y(0,1)$, $C_Y(1,0)$, $C_Y(1,1)$, $std(B_Y4)$, $std(B_Y5)$, …, $std(B_Y9)$), is then constructed. The feature vector captures colour information in the three channels (i.e. average colour intensity of the block for Y, Cb and Cr channels) as well as textural information (i.e. intensity variations) in the Y channel, the channel preserving luminance changes, i.e. the textural patterns.

There is a degree of similarity between the DCT-CT feature extraction and feature extaction from wavelet domain. The DCT coefficients in a 8x8 block are ordered similar to multi-resolution decomposition of discrete wavelet transform in three level sub-bands [33], where B_0, B_1, B_2, and B_3 correspond to level 3 frequency sub-bands LL_3, HL_3, LH_3, and HH_3, the coefficients in the sub-blocks B_4, B_5, and B_6 correspond to level 2 sub-bands HL_2, LH_2, and HH_2, and the coefficients in the sub-blocks B_7, B_8, and B_9 correspond to level 1 sub-bands HL_1, LH_1, and HH_1, representing multi-resolution textural information in high frequency bands (see Fig.1 (c)).

The appeal of the DCT-CT feature for CBIR is its relatively low dimensionality and hence its robustness in representing both colour and texture information in a local area of the image. The work presented in [14] showed that the performance of the feature is better than the DCT coefficients in a traditional zigzag order and DWT itself. It is therefore also adopted in this paper.



(a) Zigzag Ordering of the DCT Coefficients

(b) Blocking and Extraction of DCT-CT Feature

(c) Comparison with DWT Feature Space

Fig 1. DCT-CT Feature Extraction from a 8 x 8 block

B. EM/GMM Clustering Algorithm

Clustering is a process of grouping data objects into homogeneous clusters according to their similarities. The desirable result is a high degree of intra-cluster similarity and a high degree of inter-cluster differences [27]. Different categories of clustering algorithms have been developed over the last five decades [28], and have been used for CBIR. For instance, the *k*-means and EM/GMM methods have been exploited in segmenting an image into shapes [7, 26] and organizing images containing similar shapes into an indexing hierarchy of groups [29]. Our earlier work [30, 31] showed that the EM/GMM method (*model-based*) is more effective in representing image content by grouping local feature vectors into clusters than the *k*-means method (*partition-based*) and the Mean Shift method (*density-based*). This paper consequently uses the EM/GMM algorithm to group the DCT-CT local features.

The EM/GMM algorithm works by finding the best fit GMM for a given data set [27]. The algorithm consists of two primary steps. In the first expectation step, the probability that each data object is drawn from each of the *k* distributions is calculated according to the estimated parameters for *k* distributions previously (randomly chosen initially). For a mixture of Gaussians, $\theta = \{\theta_1, \theta_2, \ldots, \theta_K\}$, the probability that a data object x belongs to a Gaussian is expressed as:

$$p(x|\theta) = \sum_{k=1}^{K} a_k\, p(x_n|\theta_k),\ \sum_{k=1}^{K} a_k = 1 \qquad (2)$$

where $p(x_k|\theta_k)$ is often taken as the probability density function for the Gaussian distribution:

$$p(x_n|\theta_k) = \frac{1}{(2\pi)^{d/2}} |R_k|^{-1/2} \exp\left\{-\frac{1}{2}(x_n-\mu_k)^t R_k^{-1}(x_n-\mu_k)\right\} (3)$$

where x_n is data object n, μ_k is mean vector for k distribution, and R_k is covariance matrix for k distribution. Assuming that each data object is drawn independently, the probability of obtaining the whole data set is therefore:

$$p(X|\theta) = \prod_{n=1}^{N} \sum_{k=1}^{K} a_k\, p(x_n|\theta_k) \qquad (4)$$

The logarithm of the function above is known as the log likelihood function.

In the second step of the EM algorithm, the probabilities from the expectation step are used to derive new estimates for the parameters of the *k* distributions such that the value of the log likelihood function is increased. The process continues until the log likelihood function has reached its maximum value, indicating that the data set is the most likely result modeled by the final GMM.

Although the basic EM algorithm assumes that *k* is known, attempts have been made to optimize the order of GMM automatically. Among them is the CLUST algorithm by Bouman based on Rissanen's Minimum Description Length (MDL) principle [32]. Starting with a large value for *k* and terminating when *k* = 1, the algorithm iteratively derives the best fit GMM to the data set using the EM method and calculates the Rissanen's MDL measurement. The algorithm then finds the optimal *k* value associated with the minimum MDL measurement.

This paper will focus on not only the centers but also the shapes of clusters that are respectively represented by the cluster centroids and covariance matrices. Fig. 2 shows ellipsoid shapes of two clusters, i.e. two multivariate Gaussians. The figure clearly illustrates the difference of the two clusters in terms of their ellipsoid shapes. The paper will investigate whether this representation makes any differences in measuring the dissimilarity of images and hence the results of image retrieval. In other words, when we compute the dissimilarity between two images, we measure both the distances between the centers of ellipsoid-shaped clusters and the difference in shapes of the clusters.

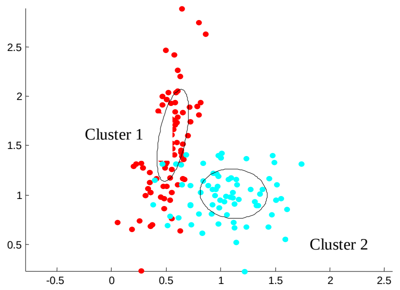


Fig 2. Gaussian Mixture Model with two clusters.

III. The Proposed Method

The process of the proposed method consists of four stages: image pre-processing, features extraction, clustering, and similarity measurement, as outlined in Fig.3. The pre-processing stage involves a single operation that converts a given image from the *RGB* space into *YCbCr* space. Once the conversion is complete, the feature extraction stage starts. It first divides the image into 8 x 8 blocks, and then applies the Discrete Cosine Transform to each block on *Y* channel and the *Cb* and *Cr* channels. Note that only $C_{Cb}(0,0)$, and $C_{Cr}(0,0)$ are calculated. The DCT coefficients from these channels are taken as the local feature vector in the way as explained in Section 2.1. All local feature vectors for the image are then collected and passed to the clustering stage where the EM/GMM clustering algorithm is applied to the collection of the extracted feature vectors to obtain clusters of ellipsoid shapes. Each cluster is then represented by the cluster centroid and its covariance matrix.

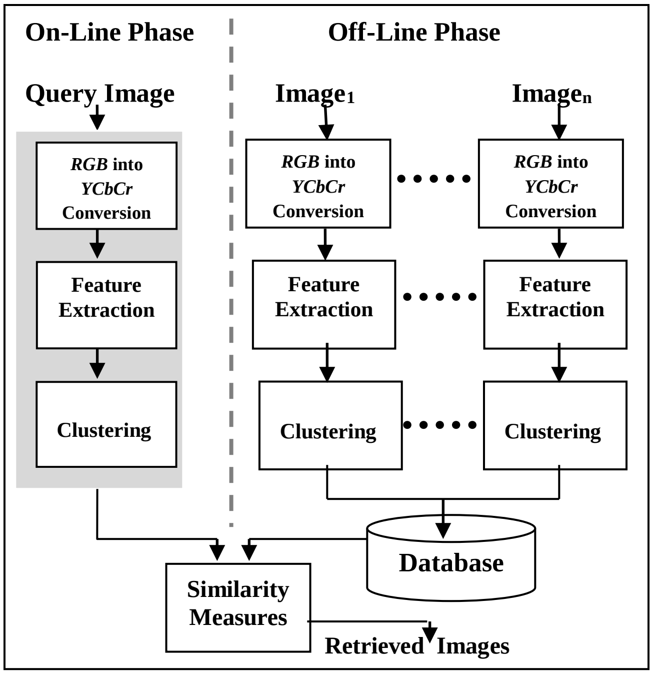


Fig 3. Framework of image retrieval.

The CBIR deals with various natural world images that vary in terms of the complexity and number of distinct objects, textures, and colours. In one of our earlier works [30], we proposed a dissimilarity measure called Aggregate Distance (AgD). This dissimilarity measure is explained again as follows. Given a query image *Q* and database image *B*, let $c^Q = \{c_1^Q, \dots, c_n^Q\}$ represent the centroids of the clusters in *Q*, and $c^B = \{c_1^B, \dots, c_m^B\}$ the centroids of the clusters in *B*. The difference between c^Q and c^B spans a dissimilarity matrix and hence:

$$D(Q,B) \propto \{d(c_i^Q, c_j^B)\}_{i,j} \in R^{n\times m}$$

where d() can be a distance function. Also in [30], it was established that the City-block distance function (D_{L1}) performed better than its other counterparts (e.g. Euclidean distance). Hence, the dissimilarity between images Q and B is defined as:

$$AgD(Q,B) = \sum_{i=1}^{n} \min(\mathrm{d}(\mathrm{c}_i^{\mathrm{Q}}, \mathrm{c}_j^{\mathrm{B}}))\ for\ j = \{1, \dots, m\}\ (5)$$

Table 1 illustrates an example of the dissimilarity matrix between the query image *Q* of three clusters, and the database image *B* of four clusters. The dissimilarity between images *Q* and *B* is calculated as the sum of the minimum distance in each row, i.e. 0.184 + 0.1056 + 0.2851 = 0.5747.

Table 1 An example dissimilarity matrix

		Database Image B			
	Clusters	1	2	3	4
Query Image Q	1	0.3523	0.6039	**0.184**	0.9823
	2	**0.1056**	0.5572	0.3428	0.4642
	3	0.9831	**0.2851**	0.3746	0.7602

Pair-wise distance functions such as D_{L1} has its limitation: it only considers the distance between two data points. In the clustering approach for CBIR, this means the proximity between cluster centroids. The dissimilarity between centroids of clusters may not be enough to distinguish between two images because different shaped clusters may have the same mean vectors. We therefore apply the Kullback-Leibler divergence [33] instead of D_{L1} to measure the dissimilarity between two clusters within the framework of the AgD dissimilarity measure. This dissimilarity measure calculates both the dissimilarity over the centroids and the shapes of the two clusters. Given a cluster W in image Q and a cluster Z in image B, the Kullback-Leibler divergence (*KLD*) is calculated as follows.

$$D_{KLD}(W,Z) = \frac{1}{2} trace\{(\Sigma_W^{-1} + \Sigma_Z^{-1})(\mu_W - \mu_Z)(\mu_W - \mu_Z)^T + \Sigma_W \Sigma_Z^{-1} + \Sigma_Z \Sigma_W^{-1} - 2d\} \quad (6)$$

where *trace* is sum of diagonal of resulted *d* x *d* matrix, *μ* is the mean vector, *Σ* is a covariance matrix, and *d* refers to the dimensionality of the local features (i.e. 12 in this paper). Using Dissimilarity Kullback-Leibler divergence (D_{KLD}) to represent pair-wise cluster dissimilarity within the dissimilarity matrix, the *AgD* measure then calculates the amount of dissimilarity between images *Q* and *B* by including the cluster shape dissimilarity in its measure of dissimilarity between images. Such a measure should increase the capability of discriminating two images.

IV. Evaluation of Proposed Method

A. Experiment Data and Evaluation Protocol

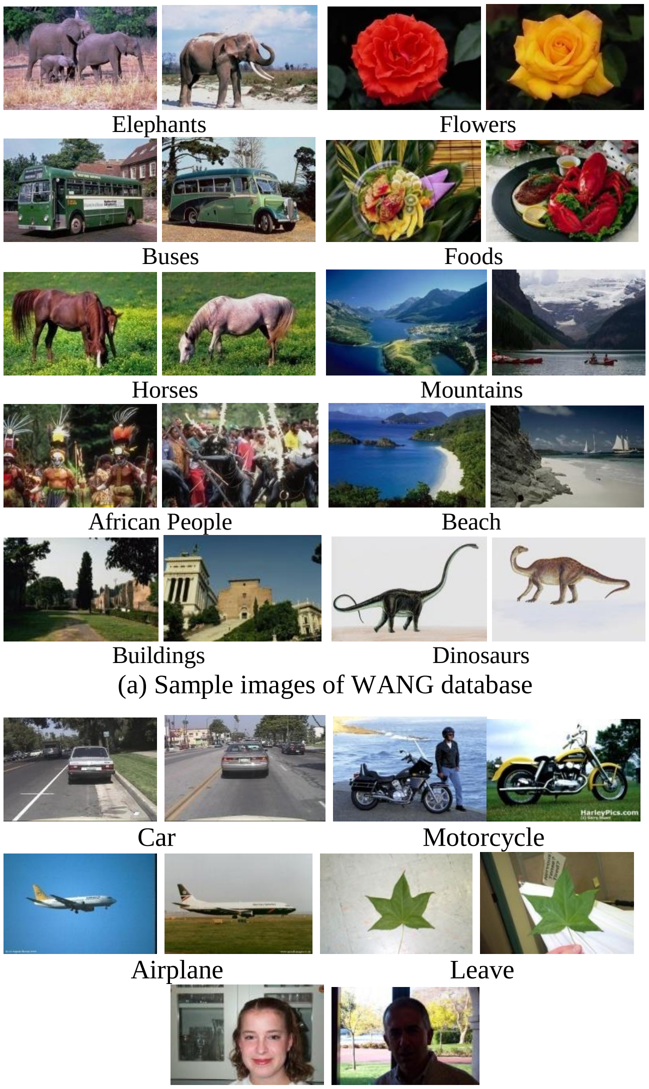
Several benchmark image databases have been used to evaluate the performance of methods by CBIR researchers. Among those benchmark databases, WANG database and Caltech 6 database are frequently used. The WANG database comprises 1000 images of sizes 256x384 or 384x256. The images are divided into 10 semantic classes such as Elephants, Flowers, Buses, Foods, Horses, Mountains, African people, Beach, Buildings, and Dinosaurs. Each class includes 100 images [34]. Fig. 4(A) shows a variety of samples from the WANG database. The Caltech 6 database includes six classes: Cars (527 images of size 360x240), Motorcycles (828 images of variables size), Airplanes (1076 images of variables size), Faces (452 images of size 896x592), Leaves (188 images of size 896x592), and Background (550 images of size 896x592) [35]. We excluded the Background class of images because they are greyscale images which differ from the rest. To use the Caltech 6 database in a similar manner as for the WANG database, 100 images of each class are randomly selected. Fig. 4(B) shows sample images from the Caltech6 database.

Many different performance metrics such as Precision Recall-graph (PR-graph), $Rank_1$, $\widetilde{Rank}$, $P(20)$, $P(50)$, $P(N_R)$, $R_P(0.5)$ and $R(100)$ have been used for CBIR [36]. In this paper, we will use two performance metrics: precision rate and ranked positions of the retrieved images. Precision rate is defined as follows:

$$Precision(C) = \frac{N_{RIC}}{R_{CID}} \qquad (7)$$

where N_{RIC} refers to the number of correct images of class *C* in the result list and R_{CID} represents the total number of images in the result list returned from the database.

Our experiments follow the same evaluation procedure. The iterative process starts by taking one image of a class as the query image, and the rest of the images in the whole database as the stored images. We then calculate the rate of precision for each query image by examining the returned list of the top 10 most similar images from the stored image collection. Once every image of a specific class has been used as the query image, and precision rate for the image is obtained, we then take the average of the precision rates for all images of the class. Once the entire database is searched, we then take the mean of the average precision rate (known as the Mean Average of Precision (MAP)) across all classes to reflect the general performance of image retrieval by the proposed method.



(a) Sample images of WANG database

(b) Sample images of Caltech6 database

Fig 4. Sample images from two databases.

B. Experiments, Results and Analysis

Two experiments have been conducted. In the first experiment, each image is represented by the centroids of the clusters from the EM/GMM clustering algorithm where the number of clusters is adaptively determined. Image dissimilarity is calculated using the D_{L1} distance function inside the AgD dissimilarity measure. In the second experiment, each image is indexed by the centroids and the shapes of the clusters from the EM/GMM clustering algorithm with the adaptively determined number of clusters. Dissimilarity between images is calculated using the D_{KLD} function inside the AgD dissimilarity measure. The performance results of the two different measurements are then compared. Table 2 presents the average precision rates for each class and the overall MAP.

Table 2 Performance metrics for *AgD* with D_{L1} and D_{KLD} when the number of clusters is adaptively determined

Image Class	**Adopted Function in AgD**	
	D_{L1}	**D_{KLD}**
Elephants	0.67	0.59
Flowers	0.88	0.93
Buses	0.80	0.81
Foods	0.52	0.57
Horses	0.88	0.90
Mountains	0.54	0.48
People	0.48	0.72
Beach	0.55	0.47
Buildings	0.44	0.59
Dinosaurs	0.95	0.98
MAP	**0.67**	**0.70**

(a) WANG database

Image Class	**Adopted Function in AgD**	
	D_{L1}	**D_{KLD}**
Car	0.99	0.997
Motorcycle	0.67	0.72
Airplanes	0.74	0.85
Faces	0.88	0.995
Leaves	0.91	0.98
MAP	**0.84**	**0.91**

(b) Caltech6 database

Figures in the table clearly indicate that for the images from the WANG database, there is improvement in precision for the majority of the classes (7 out of 10). The amount of improvement varies from a mere 1% for the Bus class to as high as 24% for the People class. For the images from the Caltech6 database, the performance raised about 1% to 12% for all five classes. Hence, the cluster-shape based retrieval performs better than the centroid-based retrieval.

We further use the *t*-test [37] to evaluate the significance of the performance differences. This statistical method is widely used. The static *t* value is calculated as:

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n} - \frac{s_y^2}{m}}} \qquad (8)$$

where $\bar{x}$ and $\bar{y}$ are the sample precision rates, s_x and s_y are the sample standard deviations, and n and m are the sample sizes. For the WANG database and Caltech6 database, we have a sample for each class where the size of the sample equals to 100 elements (i.e. precision values). The hypotheses are stated as follows. The null hypothesis (H_0) is that $\bar{x}$ - $\bar{y}$ = 0. The alternative hypothesis (H_A) is that $\bar{x}$ - $\bar{y} \neq 0$. The *t*-test was conducted using MATLAB upon two samples of precision values under the two circumstances: obtained using centroids alone, and obtained using centroids with cluster shapes to represent images. A returned value H = 0 refers to the acceptance of the null hypothesis, and a returned value H = 1 refers to a rejection of the null hypothesis. Fig 5 shows the hypothesis values returned against image classes in WANG and Caltech6 databases. It is clear that performance differences are significant for 7 out of the total 15 classes.

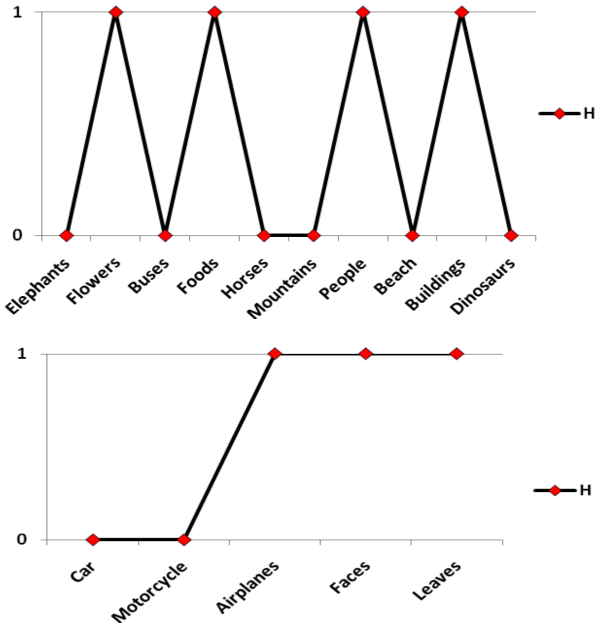
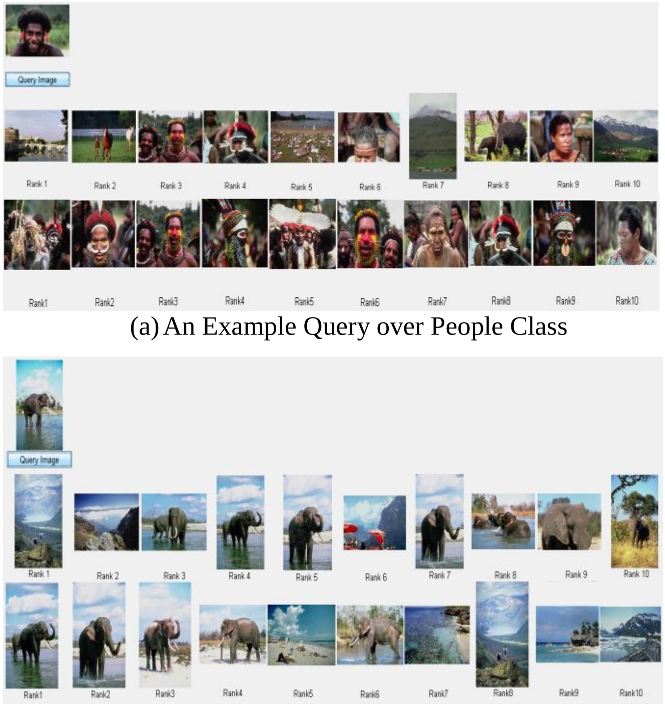


Fig 5. Hypothesis values on WANG and Caltech6 classes.

P-value of the test is the probability of observing a test. Small values of *p* cast doubt on the validity of the null hypothesis. Results of the *t*-test for Flowers, Foods, People and Buildings, classes show respective significance levels at $p = 0.018$, $p \approx 0$, $p \approx 0$, and $p = 0.0000944$. Meanwhile, the significance differences with Airplanes, Faces, and Leaves classes in the Caltech6 database are respectively $p = 0.001$, $p \approx 0$, and $p = 0.0000263$.

To further the understanding, the ranked list of 10 returned images of the classes with most different results are also closely examined. Fig. 6(a) shows an example query from the People class, and Fig. 6(b) presents an example query from the Elephants class. In both figures, the top 10 retrieval result images, when D_{L1} (first row) and D_{KLD} (second row) are respectively used, are shown.



(a) An Example Query over People Class

(b) An Example Query over Elephant Class

Fig 6. Top 10 retrieved images from using CLUST algorithm with D_{L1} and D_{KLD} distances.

In Fig. 6(a), the query image contains the face and shoulders of a person in the foreground and grass and trees in the background. The first row contains only 4 images from the class and 6 irrelevant images of other classes. However, the irrelevant images also contain grass and trees which are similar to the background of the query image, and some objects of a similar colour and texture to the body of the person in the foreground. Meanwhile, the retrieved list in the second row contains 10 relevant images that include the face and/or shoulders of the person in the query image, and images of people with similar background to that of the query image. In other words, ellipsoid-shapes of clusters add value in addition to the centroids to represent images, and therefore image discrimination is increased. In other words, cluster shapes can further distinguish relevant from the irrelevant ones. However, the similarity of cluster shapes in images of different classes may result in inclusion of images of irrelevant classes too. In Fig. 6 (b), for the elephant query image, the cluster shapes bring irrelevant images from Beach and Mountains classes in ranked positions 5, 7, 8, 9, and 10 in the second row including clouds and water. At the same time, using cluster centroids only helps to pick relevant images of elephants and images of Mountain class that have segments of colour and texture similar to the query image such as sky and mountains in the first row. However, it is interesting to note that using cluster shapes has resulted in ranking the relevant images higher in the ranked list than the irrelevant ones (i.e. the first 4 top ranked images are of relevant class).

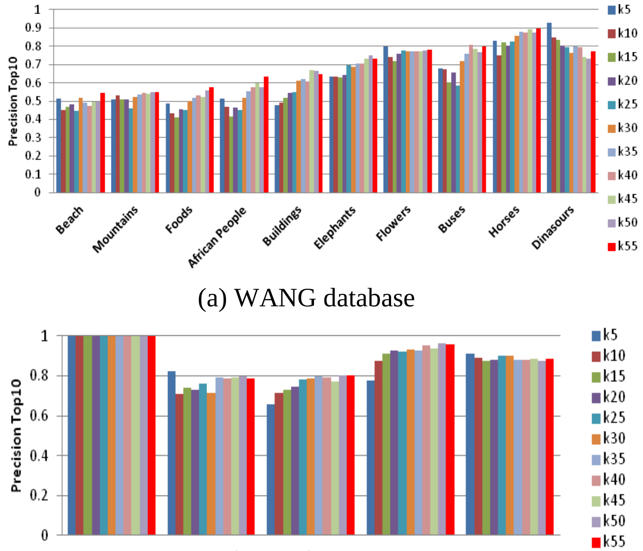
V. Effect of Number of Clusters

In the experiments described in Section 4, the EM/GMM algorithm adaptively determines the number of clusters by using the MDL as a cluster quality measure. In other words, the number of clusters is decided in an *unsupervised* way. The MDL principle is in favor of fewer clusters. From test results, it seems that the benefits of cluster shape very much relate to the image content. It is interesting therefore to study (a) the effect when the number of clusters is set manually as an external parameter particularly when the number of clusters increases, and (b) the effect of the number of clusters on the use of cluster shapes in the image retrieval process.

A. Effects of Number of Clusters Regardless Cluster Shapes

To address the first point of interest, the EM/GMM algorithm is tested using different fixed values of *K* while the AgD dissimilarity measure does not consider cluster shapes. The results are shown in Fig. 7.

From the figure, it can be seen that setting the number of clusters to a large value, the average precision rates for most of the classes across the two databases are better than the number of clusters adaptively determined using the MDL. As the number of clusters increases, the discriminating power increases too, reaching the best performance at *K*=55 for most classes. The results seem to indicate that many smaller ellipsoid-shaped clusters capture the image content better than a fewer larger and tightly organized clusters can do. The MDL principle adopted within the clustering algorithm appears oversimplifying the clusters, limiting the discriminating power of those clusters.



(a) WANG database

(b) Caltech6 database

Fig 7. Top 10 image retrieval using EM/GMM with different *K* cluster values.

Such findings seem coincident with the findings of the Bag of Visual Words approach where a much bigger number of clusters, between 500 and 10000, were used to achieve higher performance in image retrieval [38, 39]. Generating such large cluster numbers for CBIR is deemed inefficient and hence a drawback for the BOVW approach. In [40], a framework was presented for the BOVW method that reduces the size of clusters to 100 which still large. It must be noted that the increase of precision rate is not always monotonic as the number of clusters increases, and certainly not for all classes.

B. Effects of Number of Clusters by Considering Cluster Shapes

Furthermore, we have repeated the two tests mentioned in Section 4 with *K* is set to 55. The average precision rate of the top 10 retrieved images for each class together with the overall MAP is shown in Table 3. The results show that there are improvements in precision rates for all classes of both databases except Elephants, Mountains, and Beach from WANG because the percent of similarity objects between the last two classes is high.

Table 3 Performance metrics for *AgD* with D_{L1} and D_{KLD} when the number of clusters *K=55*

Image Classes	**Adopted Measure in AgD**	
	D_{L1}	**D_{KLD}**
Elephants	0.73	0.66
Flowers	0.78	0.88
Buses	0.79	0.90
Foods	0.57	0.82
Horses	0.89	0.91
Mountains	0.55	0.39
People	0.63	0.85
Beach	0.54	0.35
Buildings	0.64	0.75
Dinosaurs	0.77	0.97
MAP	**0.69**	**0.75**

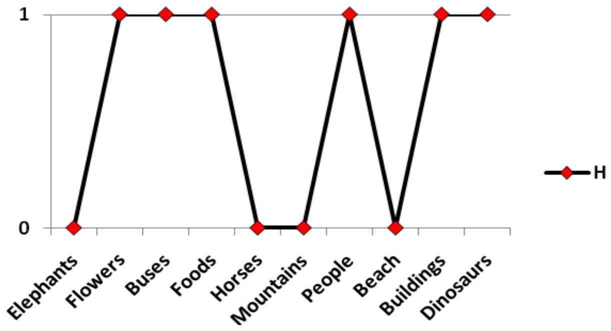
(a) WANG database

Image Classes	**Adopted Measure in AgD**	
	D_{L1}	**D_{KLD}**
Car	1	1
Motorcycle	0.79	0.99
Airplanes	0.80	0.82
Faces	0.95	1
Leaves	0.88	0.99
MAP	**0.88**	**0.96**

(b) Caltech6 database

Another *t*-test result shows that the significance levels with Flowers, Buses, Foods, People, Buildings, and Dinosaurs are respectively as p = 0.00307, 0.00020, 1.17E-10, 1.56E-08, 0.00348, and 1.85E-15. The significance levels with Motorcycles, Faces, and Leaves are at *p*-values of 1.18E-10, 6.35E-06, and 1.46E-06 respectively. Fig 8 shows the values of hypothesis values against WANG and Caltech6 classes. Hence, setting K = 55 improves the precision of image retrieval almost for all classes of the database. So considering shapes of 55 clusters has generally increased the discrimination between images of different classes.

In conclusion, the cluster shape does affect retrieval precision for most classes of images, contributing towards narrowing the semantic gap by setting the number of clusters to a large value instead of adaptively determining the number of clusters. For example, if an elephant query image is adaptively represented by 3 clusters for body, grass, and sky and 5 clusters for people image, body, grass, sky, colour of face, and clothes. Then the similarity measure (AgD) gives a chance to retrieve the people image as the most similar to the elephant query image. Meanwhile, fixed K=55 divides body, colour of face, and clothes into more clusters which are different to the query image and this will increase the discrimination between two images.



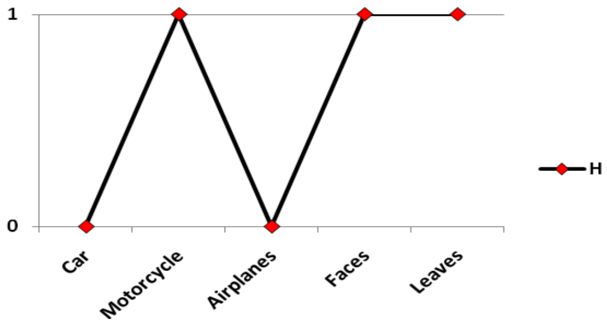


Fig 8. Hypothesis values against WANG and Caltech6 classes.

Table 4 compares the results of the proposed method using cluster shapes against other methods from the CBIR literature over the WANG database.

Table 4 Comparison of MAP of different methods using WANG database

Method	Top10
Proposed Method	0.75
[9]	0.58
[41]	0.73
[42]	0.75
[43]	0.74
[44]	0.757
[45]	0.6575

The MAP of the proposed method is higher than the method in [9] that used Comb SUM on outcomes of resulted signatures using *K-means* clustering method and global descriptors; the method in [41] that used different global and local features; and the method in [43] that used a hybrid approach of combining global and local features and applying Stationary Wavelet Transform on images. The proposed method also outperforms that reported in [45], Feng *et al.* presented a Global Correlation Descriptor (GCD) feature that represents colour and texture visual content to index images. In addition to Global Correlation Vector (GCV) and Directional Global Correlation Vector (DGCV) were proposed to integrate the advantages of histogram statistics and Structure Element Correlation (SEC) to capture colour and texture.

The result of the proposed method matches the result by the method in [42] that integrated colour features in *HSV* space (i.e. mean, standard deviation, and skewness) and texture feature (i.e. histogram of LBP with 8 neighbours and 1 radius over greyscale images). The proposed method has a less successful performance against that reported in [44]. In [44], the pre-processing step involves a colour space conversion from *RGB* to *CIE* $L^*a^*b^*$ that better represents human perception. Then Non-Subsampled Contourlet Transform (NSCT) which applied because it is fully shift-invariant, multi-scale, and multi-direction expansion with fast applicability. The best achievement is using 4 sub-bands 1, 2, 4, and 4 decompositions (1+2+4+4=11) and was done for luminance channel (L^*) to represent texture information and for chromatics channels a^* and b^* to represent colour information (i.e. 33 sub-bands for each image). Each sub-band summarized by its mean, standard deviation and energy.

Consequently, the length of the feature is 99D that represents the image in the database.

Comparing to those existing methods, our proposed method has the advantages of using more robust feature vectors of low dimensionality and a simpler algorithmic structure. The added cluster shape information within the measurement of image dissimilarity further enhanced the power of discriminating images of different kinds.

V. Conclusion and Future Work

This paper presented a new cluster-based method for CBIR that considers shapes of clusters when the image-based local colour and texture features are extracted and dissimilarity between images are measured. The experimental study results reported in the paper demonstrated that the ellipsoid shapes of the clusters formed from local DCT-CT features can help in further discriminating images of different classes of objects. This is particularly so when the number of clusters is set to a large value rather than adaptively determined using a MDL-based cluster quality measure. This seeming counter-intuitive result indicates that the localized colour and texture features are basic factors for influencing the results if content based image retrieval.

Future work includes a more extensive test on another larger database such as the Caltech256 database, a further exploration of spatial features besides local colour and textures, and applying different clustering algorithms such as density-based clustering to investigate the effect of arbitrary cluster shapes than the ellipsoid shape Gaussians.

References

[1] D. Feng, W. C. Siu, & H. J. Zhang (eds), “Multimedia informaton retrieval and management technolgical fundamentals and applications”, *Springer-Verlag Berlin Heidelberg*, 2003.

[2] R. C. Veltkamp and M. Tanase, “Content-based image retrieval systems: A survey”, *Netherlands: Technical Report UU-CS-2000-34, Dept. of Computing Science, Utrecht University*, October 2002.

[3] A. W Smeulders, M. Worring, S. Santini, A. Gupta and R. Jain, “Content-based image retrieval at the end of the early years”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12), pp. 1349-1380, 2000

[4] R. Datta, W. Ge, J. Li, and J. Z. Wang, “Toward Bridging the Annotation-Retrieval Gap in Image Search”, *IEEE Multimedia*, 14(3), pp. 24-35, 2007.

[5] W. Niblack, R. Barber, W. Equitz, M. Flickner, E.H. Glasman, D. Petkovic and P. Yanker, “The QBIC project: querying images by content using color, texture, and shape”, *Storage and Retrieval for Image and Video Databases, San Jose, CA, USA*, pp. 173-187, 1993.

[6] J. R. Smith and S. F. Chang, “Querying by color regions using VisualSEEk content-based visual query system”, *Intelligent Multimedia Information Retrieval, MIT Press, Cambridge, MA, USA*, pp. 23-41, 1997.

[7] C. Carson, M. Thomas, S. Belongie, J.M. Hellerstein and J. Malik, “Blobworld: A system for region-based image indexing and retrieval”, *Visual Information and Information Systems, Springer, Berkeley, USA*, pp. 509-517, 1999.

[8] http://googleblog.blogspot.co.uk/2009/10/similar-images-graduates-from-google.html. Accessed 3 March 2017

[9] J. Lokoč, D. Novák, M. Batko, and T. Skopal, “Visual image search: feature signatures or/and global descriptors”, *LNCS on Similarity Search and Applications, Springer-Verlag, Berlin, Heidelberg*, pp. 177-191, 2012.

[10] S. Sakji-Nsibi and A. Benazza-Benyahia “Region-based image retrieval using a joint scalable Bayesian segmentation and feature extraction”, *24th IEEE European Conference on Signal Processing (EUSIPCO)*, pp. 1272-1276, 2016.

[11] L. Duan, S. Dong, S. Cui and W. Ma, “Extreme Learning Machine with Gaussian Kernel Based Relevance Feedback Scheme for Image Retrieval”, *Proceedings in Adaptation, Learning and Optimization, Springer, Cham, Vol 1*. pp 397-408, 2016.

[12] E.G. Karakasis, A. Amanatiadis, A. Gasteratos and S.A. Chatzichristofis, “Image moment invariants as local features for content based image retrieval using the bag-of-visual-words model”, *Pattern Recognition Letters, 55,* pp.22-27, 2015.

[13] W. Plant and G. Schaefer, “Navigation and Browsing of Image Databases”, *International Conference of Soft Computing and Pattern Recognition, Malacca*, pp. 750-755, 2009.

[14] H. Al-Jubouri, “Multi-evidence fusion scheme for content-based image retrieval by clustering localised colour and texture features”, *Doctoral thesis, University of Buckingham,* 2015.

[15] J. Wan, D. Wang, S.C.H. Hoi, P. Wu, J. Zhu, Y. Zhang and J. Li, “Deep Learning for Content-Based Image Retrieval: A Comprehensive Study”, *Proceedings of the 22nd ACM International Conference on Multimedia: November,* pp. 157-166, 2014.

[16] G. Graefe, “A survey of B-tree locking techniques”, *ACM Trans Database Systems, vol. 35,* pp.16:1--16:26, 2010.

[17] TK. Sellis, N. Roussopoulos, and C. Faloutsos, “The R+-tree: a dynamic index for multi-dimensional objects”, *Proceedings of the 13th international conference on very large data bases*, pp. 507-518, 1987.

[18] L-Y. Wei, Y-T Hsu, W-C Peng, and W-C Lee, “Indexing spatial data in cloud data managements”, *Pervasive and Mobile Computing*, Vol. 15, pp. 48-61, 2014.

[19] J. Wang, W. Liu, S. Kumar and S-F Chang, “Learning to Hash for Indexing Big Data - A Survey” Proceedings of IEEE, vol. 104, no. 1, pp. 34-57, 2015.

[20] J. Lay and L. Guan, “Image retrieval based on energy histograms of the low frequency DCT coefficients”, *Proceedings of 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 3009-3012, 1999.

[21] G. Schaefer, “Content-based image retrieval: Advanced topics”, *Man-Machine Interactions 2. Springer Berlin Heidelberg*, pp. 31-37, 2011.

[22] W. M. Abd-Elhafiez and W. Gharibi, “Color Image Compression Algorithm Based on the DCT Blocks,” *arXiv preprint arXiv*:1208.3133, 2012.

[23] H. B. Kekre, S. D. Thepade, R. N. Chaturvedi, & S. Gupta, “Walsh, Sine, Haar & Cosine Transform with various color spaces for 'Color to Gray and Back’”, *Internetional Journal of Image Processing (IJIP)*, Vol.6, No.5, pp349-356, 2012.

[24]Y-L. Huang and R-F. Chang, “Texture features for DCT-coded image retrieval and classification”, *IEEE International Conference on Acoustics, Speech, and Signal Processing, Phoenix, AZ* , pp. 3013-3016, 1999.

[25] T. Westerveld, A.P. de Vries, A. van Ballegooij, F. de Jong, and D. Hiemstra, “A probabilistic multimedia retrieval model and its evaluation,” *EURASIP Journal on Applied Signal Processing*, Vol.2003, No.2, pp. 186-198, 2003.

[26] H. Nezamabadi-Pour and S. Saryazdi, “ Object-based image indexing and retrieval in DCT domain using clustering techniques”, *International Journal of Computer and Information Engineering*, Vol.1, No.3, pp. 98-101, 2007.

[27] H. Du, “Data mining techniques applications: an introduction,” *Cengage Learning EMEA*, 2010.

[28] A. Jain “Data clustering beyond K-means”, *Pattern Recognition Letters*, Vol 31, pp. 651-666, 2010.

[29] N. Vasconcelos, “Image Indexing with Mixture Hierarchies,” *IEEE Conference in Computer Vision and Pattern Recognition*, pp. 3-10, 2001.

[30] H. Al-Jubouri, H., Du, & H. Sellahewa, “Applying Gaussian Mixture Model on discrete cosine features for image segmentation and classification”, *Proceedings of 4th Computer Science and Electronic Engineering Conference (CEEC), Colchester, UK*, pp. 194-199, 2012.

[31] H. Al-Jubouri, H., Du, & H. Sellahewa, “Adaptive clustering based segmentation for image classification,” *5th Computer Science*

and Electronic Engineering Conference (CEEC), Colchester, UK, pp. 128-133, 2013.

[32] C. A. Bouman, “Cluster: unsupervised algorithm for modeling Gaussian mixtures”, 1997. https://engineering.purdue.edu/~bouman/software/cluster/manual.pdf, Accessed on 8 December, 2018

[33] T. A. Myrvoll and F. K. Soong, “On divergence based clustering of normal distributions and its application to HMM adaptation”, *Proceedings of 8th European Conference on Speech Communication and Technology, Geneva*, pp. 1517-1520, 2003.

[34] J. Z. Wang, “Integrated region-based image retrieval,” *Norwell, MA, USA: Kluwer Academic Publishers*, 2001.

[35] R. Fergus, P. Perona and A. Zisserman, “Object class recognition by unsupervised scale-invariant learning,”. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Oxford, UK*, pp. II-264, 2003.

[36] H. Müller, W. Müller, D. M. Squire, S. Marchand-Maillet and T. Pun, “Performance Evaluation in Content-based Image Retrieval: Overview and Proposals,” *Pattern Recognition Letters*, 22(5), pp. 593-601, 2001.

[37] A. Field, "Discovery statistics using SPSS", London, UK: SAGE Publications, 2006.

[38] C. Feng, and X. Wang, “Image retrieval system based on bag of view words model”, *Proceedings of 15th International Conference on Computer and Information Science (ICIS)*, pp. 1-4, 2016.

[39] R. Vieux, J. Benois-Pineau and J-P. Domenger, “Content Based Image Retrieval Using Bag-of-regions”, *Springer-Verlag Berlin Heidelberg*, pp. 507-517, 2012.

[40] Z. Lu, L. Wang and J.R. Wen, “Image classification by visual bag-of-words refinement and reduction”, *Neurocomputing*, Vol.173, pp.373-384, 2016.

[41] V. Karpagam and R. Rangarajan, “A Simple and Competent System for Content Based Retrieval of Images using Color Indexed Image Histogram Combined with Discrete Wavelet Decomposition”, *European Journal of Scientific Research*, 73(2), pp. 278-190, 2012.

[42] M. Salmi and B. Boucheham, “Content based image retrieval based on Cell Color Coherence Vector (Cell-CCV)”, *Proceedings of 4th International Symposium ISKO-Maghreb: Concepts and Tools for knowledge Management*, pp. 1-5, 2014.

[43] M. D. Chaudhary and A. B. Upadhyay, “Fusion of local and global features using Stationary Wavelet Transform for efficient Content Based Image Retrieval”, *IEEE Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, pp. 1-6, 2014.

[44] M.K. Kundu, M. Chowdhury, and S.R. Bulò, “A graph-based relevance feedback mechanism in content-based image retrieval”, Knowledge-Based Systems, Vol.*73*, pp.254-264, 2015.

[45]L. Feng, J. Wu, S. Liu and H. Zhang. “Global correlation descriptor: a novel image representation for image retrieval”, *Journal of Visual Communication and Image Representation*, Vol.*33*, pp.104-114, 2015