International Journal of Multimedia and Ubiquitous Engineering
Vol.11, No.2 (2016), pp.79-88
http://dx.doi.org/10.14257/ijmue.2016.11.2.10

Average Analysis Method in Selecting Haralick's Texture Features on Color Co-occurrence Matrix for Texture Based Image Retrieval

Abd Rasid Mamat[1], Mohd Khalid Awang[1] , Norkhairani Abdul Rawi[1] ,Mohd.Isa Awang[1] and Mohd Fadzil Abdul Kadir[1]

[1] Faculty of Informatics and Computing,
Universiti Sultan Zainal Abidin (UniSZA),
Terengganu, Malaysia
[1]arm@unisza.edu.my, [1]khalid@unisza.edu.my, [1]khairani@unisza.edu.my,
[1]isa@unisza.edu.my, [1]fadzil@unisza.edu.my

Abstract

Many textures based image retrieval researchers use global texture features for representing and retrieval of images from an image database. However, this leads to misrepresentation of local information leading to the inefficient image retrieval performance. This paper presents an approach to overcome the problem. The approach focuses on extracting local Haralick's texture feature based on a predetermined region using the color co-occurrence matrix method, the selection of the 'significant' Haralik's texture features and evaluation of the performance of the combination of the 'significant' features. The proposed method which is an Average Analysis and a well known method, Principal Component Analysis were applied to obtain 'significant' features. In order to compare the performance, a series of experiments were carried out for both methods, which is the proposed Average Analysis and the Principal Component Analysis. Experiments were performed on a 1000 selected images from the Coral image database which were divided into ten categories. Based on the experimental results, it is interesting to note that for the combination 'significant' features obtained from the proposed Average Analysis showed better retrieval performance compared to the Principal Component Analysis for almost all categories. This finding has an important implication in deciding the correct combination of 'significant' features for certain image properties. It has shown that the proposed method is able to produce less computational processing time due to a reduced amount of processing involved. The result is also compared to the previous researches and has shown an increase of an average precision from 8.5% to 26%.

Keywords: *Texture based image retrieval, Color Co-occurrence matrix, Haralick's texture features, Average Analysis, Principal Component Analysis*

1. Introduction

Due to the exponential growth of image data, there is a compelling need for innovative tools which can easily manage, retrieve and visualize images from a large multimedia databases. Accordingly, CBIR field has grown with the emergence of many advanced techniques. The main goal of the CBIR is to find images which are similar to the query image visually without using any textual descriptions of the image. Low level features such as color and texture features were widely used in CBIR [1] and these features may be extracted from either global or local regions from the images. This study, the proposed method is discussed in two parts.The first part is based on grey level techniques which are

ISSN: 1975-0080 IJMUE
Copyright ⓒ 2016 SERSC

adapted from the color information to produce a Color Co-occurrence Matrix (CCM), meanwhile the second part focused on the feature selection.

The rest of the paper is organized as follows: Section two of this paper presents the related work, section three presents our proposed methodology that improves on the existing CBIR mechanisms and algorithms, section four presents the experimental results, section five presents the conclusion and the final section six, presents the future work of this paper.

2. Related Work

Color histogram, which is the first order statistical measure, usually represents color features that were extracted from the global image [2]. This method ignored the spatial distribution and local variation in the color image. Local spatial variation of pixel intensity commonly were used to capture texture information, this is an established method known as gray level Co Occurrence Matrix (GLCM) [3]. The GLCM method initially used grey image, but for a decade the study of texture information was extended for color images [4]. A researcher [6], proposed the method based on the existing gray level that are adapted to take the color information from color images. Another researcher [11] introduced CCM as statistical features which both measure the color distribution in an image and consider the spatial interactions between the color pixel for color space. Haralicks features are used as texture features and extracted from CCM. In his research, he compares the performances of texture classification from different color space. Meanwhile, another researcher [12] developed a new retrieval system called CTDCIRS (Content based image Retrieval System on Dominant Color and Texture Features). In this system, the authors were using the dynamic dominant color (DDC) Motif co-occurrence Matrix (MCM- similar to Color co-occurrence matrix) and the difference between pixels of scan pattern (DBPSP). Many researchers who conducted the study of texture features using color images to replace the grayscale image and found the results to be disappointing, thus become the impetus for this study.

The second part of this study is the selection of features from a group of features. The authors in [13] used Principal Component Analysis (PCA) for a dimension reduced for feature vector, namely color bin histogram of the image and the results show more robust and computationally efficient. Meanwhile, according to authors in [14], they used PCA to extract the principal components of the feature values and these features are Radiance Histogram (RH) and Multispectral Co-occurrence Matrix (MCM). This method was performed and tested on a set of LANDSAT multispectral images from variant sceneries and the results show superior performance using this approach. Due to this, the strength of the PCA is to trigger it used in this research and then compare the performance of the proposed method.

3. Methodology

In this section, the methodology of the proposed method is presented. A part of this methodology is summarized as shown in Figure 1. Firstly, all images are converted into the CIE Lab color space [5] and were divided into a sunblock or sub region color image [6]. The sublbocks are then separated into their color band as L, a and b. In all of the experiments, this work will be used only for a and b band [7] and combination of (a and b) band. The following were generated Color Co-occurrence Matrix (CCM) based on pixel neighborhood and direction for each block. Then, eleven Haralick features were extracted from CCM. For the feature selection we applied to the method, they are namely as PCA and proposed method, Average Analysis (AA). The purpose of this process is to obtain significant features expected to improve the performance. Based on the average Precision and Recall (PR), the performance is compared.

 Copyright ⓒ 2016 SERSC

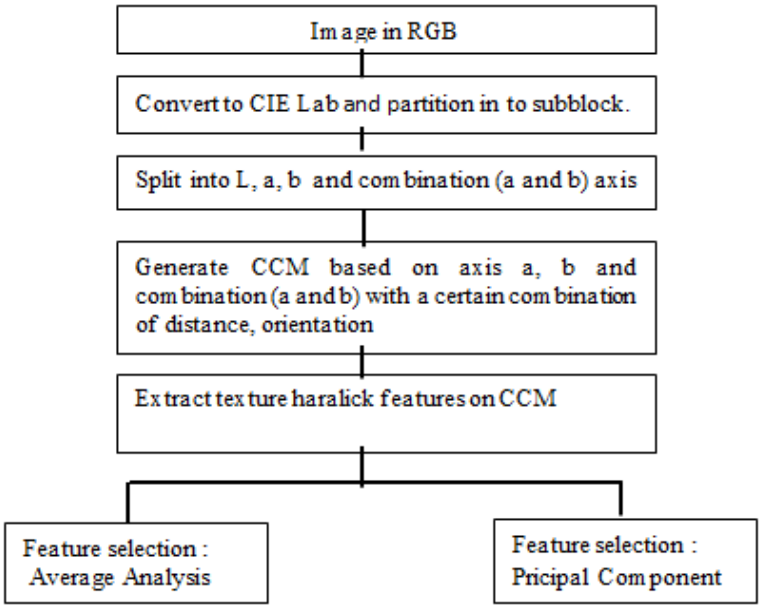

Figure 1. A Part of Methodology for the Proposed Method

3.1 Generated Color Co-Occurrence Matrix (CCM)

The GLCM technique for the texture analysis has been defined for grayscale images. For this research, the extension of this technique is used for color images, where this technique and their statistical features are computed for each band of the color space [9]. Based on this extension, the CCM are defined for each color space. The detail of the process to compute CCM is such as in [20]. Thus an image will produce 204 CCM. In general, the proposed method is different from the methods that have been made by other researchers in terms of partition image, a combination of bands and the direction that are used to extract the features. Another difference involves the selection of features using AA and PCA.

3.2 Extraction the Haralick Textures on CCM

Eleven Haralick features were extracted from CCM. These features are as follows:

i. Contrast (*f2*) = $\sum_i \sum_j (i-j)^2 p(i,j)$ (1)

ii. Energy (*f2*) =$\sum_i \sum_j p(i,j)^2$ (2)

iii. Entropy (*f3*)= $-\sum \sum p(i,j) \log p(i,j)$ (3)

iv. Homogeneity (*f4*)= $\sum_i \sum_j \frac{P_d(i,j)}{1+|i-j|}$ (4)

v. Sum of squares: variance (*f5*) =$\sum_i \sum_j (i-\mu)^2 p(i,j)$ (5)

vi. Sum of average *(f6)*=$\sum_{i=2}^{2Ng} i p_{x+y}(i)$. (6)

vii. Sum variance (*f7*) =$\sum_{i=2}^{2Ng} (i-f)^2 p_{x+y}(i)$ (7)

where $f = \sum_{i=2}^{2Ng} i p_{x+y}(i)$.

viii. Sum entropy*(f8)* =- $\sum_{i=2}^{2Ng} P_{x+y}(i) \log\{P_{x+y}(i)\}$ (8)

ix. Difference entropy *(f9)*= $-\sum_{i=0}^{Ng-1} P_{x-y}(i) \log\{P_{x-y}(i)\}$ (9)

x. Information measure correlation 1 *(f10)*=$\frac{HXY-HXY1}{max\{HX,HY\}}$ (10)

xi. Information measure correlation 2 $(f11)=(1-\exp[-2.0\,(HXY2-HXY)]^{1/2}$ (11)

where $HXY = -\sum_i \sum_j p(i,j)\log(p(i,j))$,

HX and HY are entropies of p_x and p_y, and

$HXY1 = -\sum_i \sum_j p(i,j)\log\{(p_x(i)p_y(j)\}$

$HXY2 = -\sum_i \sum_j p_x(i)\,p_y(j)\log\{(p_x(i)p_y(j)\}$

For each image *I*, it contains 204 CCM blocks and eleven Haralick features were extracted. Consequently, there are 2244 vector features produced from an image.

3.3 Feature Selection

In general, the result of the feature extraction process from the image database will obtain many useful features. To decide which features and combinations have a greater ' contribution' on the efficiency of the image retrieval is an important subject needed to be investigated. Towards achieving that goal, two methods have been used, namely PCA and AA. Finally, the performance comparisons between the methods were analyzed.

3.4 Principal Component Analysis

PCA is an established method and widely used for dimension reduction [15]. According to [17] the few steps to compute PCA are as follows:

- Subtract the mean

Subtract the mean from each of the data dimensions. Eq. (12) is used to calculate mean:

$$\text{Mean } X, (\bar{X}) = \frac{\sum_{i=1}^{n} X_i}{n} \quad (12)$$

where *n* is number of element of *X*.

- Calculate the Covariance Matrix

Covariance matrix is always measured between two dimensions of the data and to Calculate the covariance is very similar with calculated variance. Eq. (13) is used To calculate the covariance.

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{n-1} \quad (13)$$

Where *n* is number of data, $\bar{X}$ is the mean of dimension one and $\bar{Y}$ is the mean of Dimension two. To compute $\bar{X}$ and $\bar{Y}$ used eq. (14) and (15).

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \quad (14)$$

Where *n* is number of element of *X* and

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} \quad (15)$$

Where *n* is number of element of *Y*.

- Calculate the eigenvectors and eigenvalues of the covariance matrix.

Since the covariance matrix is square, the value of eigenvector and eigenvalues Will be calculated and the result is sorted in the descending order. This value gives The components in order significance. For example, eq. (16) shows the matrix Which contains the eigenvector and eigenvalue. In this case θ is eigen value and

 Copyright ⓒ 2016 SERSC

matrix $\begin{bmatrix} i \\ j \end{bmatrix}$ is eigen vector.

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \times \begin{bmatrix} e \\ f \end{bmatrix} = \begin{bmatrix} g \\ h \end{bmatrix} = \theta \times \begin{bmatrix} i \\ j \end{bmatrix} \quad (16)$$

In conclusion, the highest eigenvalue is represented by component one and followed by component two until the lowest value is represented by the last component. According to [18] there are three criteria, namely Eigenvalue, Scree Test and Proportion of variance accounted that might be used in making this decision to determine the meaningful component. For example, Figure 2 is showing the plot of Scree Test. In [19], the Scree Test means that the eigenvalues are plotted against the components.

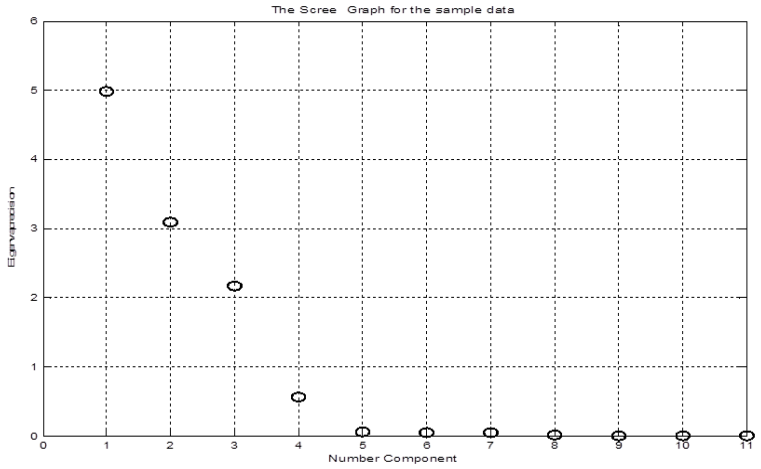

Figure 2. The ScreeTest Graph

From Figure 2, the mean of break occurs between component 3 and 4. This shows that, the component 1, 2 and 3 are categorized as meaningful components or 'significant' features and retained for further analysis.

3.5 Average Analysis

This method is based on analysis on the value PR for every feature of each category. The steps to compute this method are as follows:

- Compute and plot Precision and Recall (PR) graph for each feature of query image using (21) and (22).
- Compute and plot the Average of PR (APR) graph**.** Equation (17) is used to calculate of the APR.

$$\text{APR} = \frac{\sum_{n=1}^{10} PR_i}{n} \quad (17)$$

Where n =10 and i is number of features, i=1,2,3,..11.

- Analysis APR and compute new APR (newAPR). Apply (18) to calculate the value of the newAPR.

$$\text{newAPR} = \frac{\sum_{u}^{11} APR_u}{u} \quad (18)$$

Where u, is number of APR features. For example, Figure 3 shows the newAPR for category eight.

Copyright ⓒ 2016 SERSC

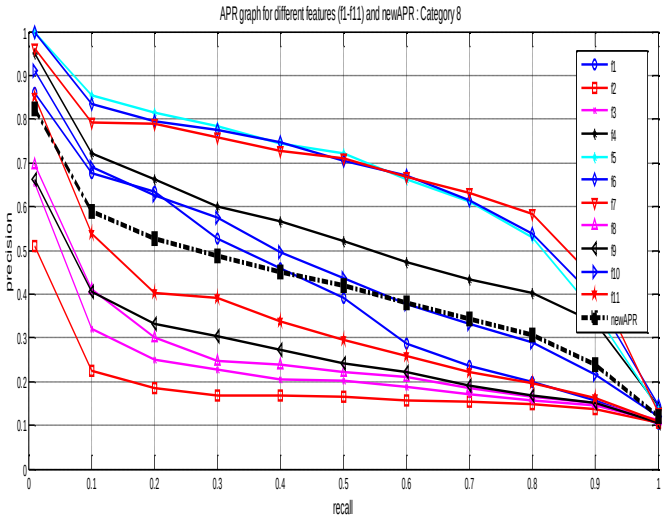

Figure 3. APR Graph of Different Features (*f1-f11*) and NewAPR for Category 8

Compare the APR and newAPR.

Compare the value of APR and newAPR graph. Label newAPR graph as *w*, and APR for each feature is *t*. If *w* is greater than *t*, it shows that this feature is 'better' or more 'significant' from other features and vice versa. This statement can be summarized as equation (19).

'Significant' APR of features:

$$= \begin{cases} \text{Significant (features) APR} & \text{if } t \geq w \\ \text{others} & \text{if } t < w \end{cases} \quad (19)$$

After obtaining the 'significant' features, the experiments were conduct again to find the APR of the combination and example the result shown in Figure 5(a) and 5(b).

3.6 Similarity Measure and Performance Evaluation

To obtain the similarity measures, the system computes the feature of query image and index image in the database image feature. It is responsible for finding the distance between them which will be interpreted in terms of relevancy with each other [21-23]. The distance between query image *X* and index images in database *Y*, were calculated using Euclidean distance as expressed in equation 20 [8].

$$\text{dist}(X,Y) = \left(\sum_j (x_{u,d,o,f} - y_{u,d,o,f})^2\right)^{1/2} \quad (20)$$

Where j as the number of blocks of an image, meanwhile u is color of bands, d is distance, o is orientations and f is a feature.

To compare the retrieval performance between different experiments, precision and recall were used. Precision (P) is defined as the ratio of the number of relevant images retrieved (N_r) to the number of total of the images retrieved K, whilst Recall (R) is defined as the number of retrieved relevant images N_r, over the total number of relevant images available in the database N_t.

$$\text{Recall} = \frac{N_r}{N_t} \quad (21)$$

 Copyright ⓒ 2016 SERSC

$$\text{Precision}=\frac{N_r}{K} \quad (22)$$

4. Experiment Setup

In this work the image database used was developed by Wang and his colleagues from the Pennsylvania State University and [10] is available at http://wang.ist.psu/edu. This image database contains 1000 of color images and was already categorized into 10 categories. The classification has been made based on manual perceptive view such as a category 1 is African people and village, category 2 is Beach, category 3 is Building, category 4 is Buses, category 5 is Dinosaur, category 6 is Elephants, Category 7 is Flowers, Category 8 is Horses, category 9 is Mountains and Glaciers and final category 10 is Food. Figure 4 shows an example of the images. 10% of is used as the query image.

Figure 4. Example of an Image

5. Results and Discussion

The results of the experiments are listed in Figure 5(a) –(b), Table 1 and Table 2. Figure 5(a)-(b) shows an example of the result from a combination 'significant' features obtained through AA (proposed method) and PCA methods. Table 1 represents the summary of Figure 5(a)-(b) based on cumulative of APR and finally Table 2 shows the performance of the proposed method compared with other researcher[16].

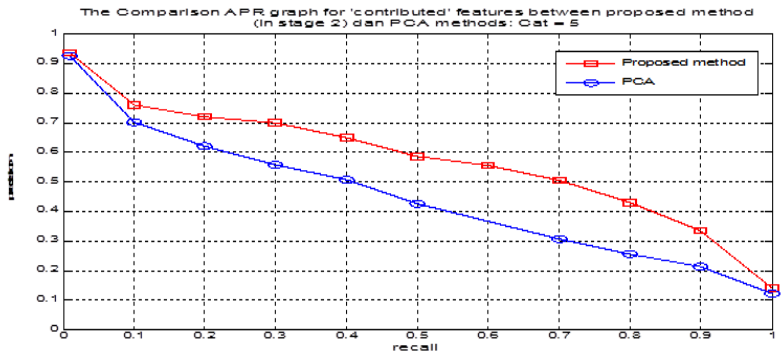

Figure 5 (a). Performance Comparison the AA and PCA for Cat.5

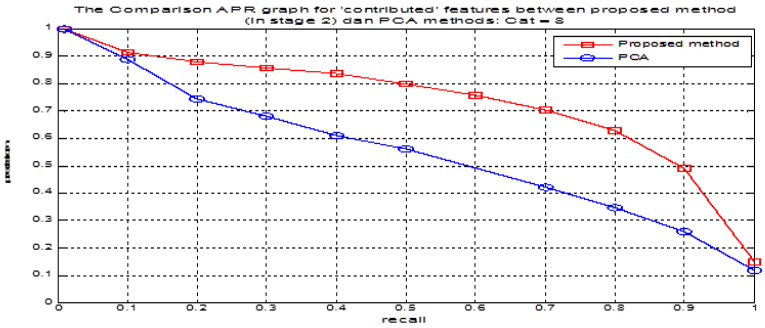

Figure 5 (b). Performance Comparison Values of AA and PCA for Cat.8

Table 1. Performance Comparison between AA and PCA

Category	AA	PCA	Accuracy(%) of retrieval
1	5.0055	4.0830	(+) 18.4297
2	2.5010	1.8713	(+) 25.1779
3	3.0413	3.0847	(-) 1.4069
4	4.4102	4.0539	(+) 8.0789
5	6.3154	5.0040	(+) 20.7651
6	2.7640	2.7964	(-) 1.1798
7	5.4040	5.0860	(+) 5.8845
8	7.0166	6.1215	(+) 12.7569
9	3.9519	2.6753	(+) 32.3034
10	3.9420	3.8020	(+) 3.5515

According to Table 1 the performance of the proposed method is better for all categories except categories 3 and 6. The highest performance is category 9 (32.3043%), followed by 2 (25.1779%), category 5 (20.7651%), category 1 (18.4297%), category 4 (8.0798%), Category 7 (5.8845%) and category 10 (3.5515 %). This shows that the proposed method is more accurate in determining the 'significant' features and combinations of these features is the main features possessed by the images in that category compared with the other features.

Table 2. Performance Comparison between AA and [16]

Category	**Average Precision using :**					
	Proposed by [16]					**Proposed Method**
	GTF	**CMW**	**CMR**	**GTF + CMW**	**GTF + CMR**	
1	0.37	0.75	0.75	0.74	0.74	0.81
2	0.27	0.46	0.38	0.38	0.38	0.51
3	0.33	0.25	0.35	0.30	0.36	0.55
4	0.35	0.67	0.78	0.60	0.77	0.81
5	0.99	0.74	0.83	0.96	0.95	0.83
6	0.39	0.60	0.45	0.58	0.44	0.64
7	0.75	0.42	0.61	0.71	0.69	0.66
8	0.27	0.55	0.70	0.47	0.67	0.96
9	0.24	0.67	0.62	0.72	0.69	0.56
10	0.20	0.43	0.43	0.36	0.41	0.63
Average Precision (%)	**43.5**	**55.4**	**59.0**	**58.2**	**61.0**	**69.5**

Copyright ⓒ 2016 SERSC

Thus the results produced by the AA methodare compareded with the work done by author in [16]. In [16], the database image and the number of queries are used is same for each category, but differs only in terms of the methodologies.The methodology is applied based on Gabor texture features (GTF), color moment based on the whole image (CMW), color moment from dividing the image into three (3) equal non overlapping horizontal regions (CMR), CMW + GTF and finally CMR + GTF. Performance evaluation based on the results of the top 10 images for each query is shown in Table 2. Based on Table 2, the performance of the AA is better for category 1, 2, 3, 4, 6, 8 and 10 except category 5 and category 9. In category 5 the proposed method is better only for CMW and CMR, meanwhile less than from GTW, GTF+CMW and GTF + CMR for category 7. Meanwhile, for category 9, the performance of the proposed method is better than the GTF only.

6. Conclusion and Future Work

In this paper, we have presented the local Haralick texture features for representing and retrieving images from an image database based on the predetermined region by using CCM method. From the experimental results, the performance of combination 'significant' features determined by the AA is better than PCA and increase the accuracy of retrieval compared to the previous method. In the future, we plan to extend our experiment using different feature selection techniques.

Acknowledgements

This project has been registered with the Center of Research and Innovation Management (CRIM) as External Project under a code of UniSZA/2015/PPL (018). We would like to thank CRIM and the management of Faculty of Informatics and Computing (FIK) for their support toward this project.

References

[1] V. S. Sural and A. K. Majundar, "An Integrated and Intensity Co-occurrence Matrix," Pattern Recognition Letters, vol. 28, **(2007)**, pp. 974-983.
[2] M. J. Swain and D. H. Ballard, "Color indexing", Internat. J. Comput, vol. 7, **(1991)**, pp. 11–32.
[3] M. Haralick, "Texture Image Classification", IEEE Transactions On Systems, Man and Cybenertics, vol. 3, no. 6, **(1973)**, pp. 610-621.
[4] A. Porebski, N. Vandenbroucke and L. Macaire, "Neighborhood and Haralick feature extraction for color texture analysis", In Proceeding of the 4th European Conference on Colour in Graphics, Image and Vision (CGIV'08), Terrassa, Spain, **(2008)**, pp. 316-321.
[5] X. Li, S. C. Chen, M. L. Shyu and B. Furth, "An Effective Content-based Visual Image Retrieval System", Computer Software and Application Conference (COMPSAC), Oxford, England, **(2002)**, pp. 914-919.
[6] V. Arvis, C. Debain, M. Berducat and A. Benassi, "Generalization of the Cooccurence Matrix For Colour Image", Application to Colour Texture Classification. Image Analysis and Streology, vol. 23, **(2004)**, pp. 63-72.
[7] A. Drimbarean and P. F. Whelan, "Experiments in Color Texture Analysis", Pattern Recognition Letter, vol. 22, **(2001)**, pp. 7-1161.
[8] Y. Liu, D. Zhang, G. Lu and W. Y. Ma, "A Survey of CBIR with High-level Semantics", Pattern Recognition, vol. 40, **(2007)**, pp. 262-282.
[9] A. Moulay, X. Maldague and W. B. Larbi, "A New Color-Texture Approach for Industrial Products Inspection", Journal Of Multimedia, no. 3, **(2008)**.
[10] J. Li and J. Z. Wang, "Automatic linguistic indexing of pictures by a statistical modeling approach", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 9, **(2003)**, pp. 1075-1088.
[11] C. Palm, "Color Texture Classification by integrative co-occurrence matrices", Pattern Recognition, vol. 37, **(2004)**, pp. 965-975.
[12] M. B. Rao, B. P. Rao and A. Govardhan, "CTDCIRS: Content based Image Retrieval System based on Dominant Color and Texture Features", International Journal of Computer Applications. vol. 18, no. 6, **(2011)**.

Copyright ⓒ 2016 SERSC

[13] T. Bhattacharjee, B. Banerjee and N. Chowdhury, "An Interactive Content Based Image Retrieval Technique and Evaluation of its Performance in High Dimensional and Low Dimensional Space", International Journal of Image Processing (IJIP), vol. 4, no. 4, **(2010)**, pp. 329.
[14] H. Shahbazi, M. Soryani and P. Kabiri, "Content Based Multispectral Image Retrieval Using Principal Component Analysis", CIVR, **(2008)**.
[15] D. F. Morrison, "Multivariate statistical methods", New York: McGraw-Hill, **(1967)**.
[16] S. Mangijao and K. Hemachandran, "Content-Based Image Retrieval using Color Moment and GaborTexture Feature", IJCSI International Journal of Computer Science Issues, vol. 9, Issue 5, no. 1, **(2012)**.
[17] S. I. Lidsay, "A tutorial on Principal Components Analysis", University of Otago, New Zealand, **(2002)**.
[18] O. Rourke, N. L. Hatcher and E. J. Stepanski, "A step-by-step approach to using SAS for univariate and multivariate statistics", Second Edition. SAS Institute, Inc., Cary, North Carolina, USA, **(2005)**.
[19] R. B. Cattell, "Handbook of multivariate experimental psychology", Chicago: Rand McNally, **(1966)**.
[20] A. R. Mamat, M. Muhammad, M. I. Awang, N. A. Rawi, M. F. A. Kadir and M. K. Awang, "Selecting Haralick's Texture Features on Color Co-occurrence Matrix for Image Retrieval using Average Analysis for Image Retrieval", The Third International Conference on Informatics & Applications (ICIA2014), **(2014)**.
[21] S. Simone and R. Jain, "Similarity Measures", IEEE Transactions On Pattern Analysis And Machine Intelligence, vol. 21, no. 9, **(1999)**.
[22] G. Qian, S. Sural and G. S. Pramanik, "Similarity between Euclidean and cosine angle distance fornearest neighbor queries", In Proceedings of the 2004 ACM symposium on Applied computing, ACM, **(2004)**, pp 1232-1237.
[23] D. Zhang, and G. Lu, "Evaluation of similarity measurement for image retrieval", In Neural Networks and Signal Processing. Proceedings of the 2003 International Conference on, IEEE, vol. 2, **(2003)**, pp. 928-931.

 Copyright ⓒ 2016 SERSC