Prevalent Color Extraction and Indexing

K.K.Thyagharajan [#1] and R.I.Minu [*2]

[#] Dean(Academic), RMD Engineering College,India
[1]kkthyagharajan@yahoo.com
[*]Research Scholar, Dept. of Computer Science,Anna University,India
[2]r_i_minu@yahoo.co.in

Abstract:- **Colors in an image provides tremendous amount of information. Using this color information images can be segmented, analyzed, labeled and indexed. In content based image retrieval system, color is one of the basic primitive features used. In Prevalent Color Extraction and indexing, the most extensive color on an image is identified and it is used for indexing. For implementation, Asteroideae flower family image dataset is used. It consist of more than 16,000 species, among them nearly 100 species are considered and indexed by dominating colors. To extract the most appealable color from the user defined images, the overall color of an image has to be quantized. Spatially, quantizing the color of an image to extract the prevalent color is the major objective of this paper. A combination of K-Mean and Expectation Minimization clustering algorithm called hidden-value learned K-mean clustering quantization algorithm is used to avoid the over clustering behavior of K-Mean algorithm. The experimental result shows the marginal differences between these algorithms.**

Keyword: **Color Quantization, K-Mean, EM Algorithm, Asteroideae, RGB, HSV;**

I. INTRODUCTION

An image can be represented either through global features or by local features. Extensively most of the image retrieval techniques such as CBIR [1][3] uses local features which are said to be the content of an image. These local Features are color, shape and texture of the images which were used to understand and identify an image.In Content based Image Retrieval systems, Color [1][2] is one of the essential low-level feature content used to index an image. A color is a parameter, which depends upon the frequency of light [12]. In digital image processing the colors are represented as a mathematical co-ordinates called as Color Models [14][20]. The commonly used color models are RGB, HSV, HSI, CMY and YCbCr models. Each model has its own characteristic, in this work of most prevalent color extraction the RGB and HSV mathematical models were used.

Dominant color extraction is one of the most profound research areas. B.S. Manjunath et al [4], [5] provide an effective way of determining maximum of 8 dominant color from a local image, which has been used as MPEG 7's [5] Dominant Color Descriptor (DCD) . In DCD the images are segmented into sub-region, the colors in these areas are quantized and color histograms were generated from these color bins of all the sub region the dominant colors were identified and were labeled uniquely. For those labeled colors the percentage, color variances and the spatial coherency are determined and the similarity between the colors of the pixels is identified using Euclidean distance. Using this DCD as one of the primitive low level feature [6][8][9][23] Image retrieval system were designed.

In this work, to extract the most prevalent color from the image: First the image is quantized using hidden-value learned K-mean clustering algorithm (EMK), the quantized image is then converted into HSV color model and the pre-domination color from the image is extracted and that image is then indexed as per there extracted color. These steps are all explained with corresponding result analysis in the forthcoming sections.

II. COLOR IMAGE QUANTIZATION

The images would be of raw data. This image has to be pre-processed before performing any mathematical operation on it. To identify the most domination color from the image, the colors on the image has to be quantized to limited set of colors. Colors in the image can be quantized either by Scalar Quantization methods or Vector Quantization methods [13]. There are many color quantization technique [18], some of the standard techniques used for statistical analyze are illustrated below:

A. Scalar Quantization

In Scalar or Uniform quantization method just the RGB color values are sliced to fixed ranges mostly 64 and all the color pixel values in RGB are quantized between [0 – 64] instead of [0-255] . Here the spatial color distribution was not considered while using this uniform quantization methods, so it losses most of the essential information regarding the colors.

B. Median Cut Quantization

In Vector or Non-Uniform quantization methods, the color values are quantized depending upon the color distribution. The quantization methods which uses unsupervised learning technique for adaptive quantization are Median-cut[15], OcTree[16] and K-Mean clustering quantization algorithm.

In Median-cut algorithm, the number of different color in the image ($C= c_1, c_2, ..c_m$) and the total number of pixels for a particular color (nc_1) are all determined. If K be the total number of quantized colors to be found, then a box kind of splitting algorithm with respect to the median value of ($nc_1, nc_2,..$) pixel is done iteratively until the total number of C is equal to K. Thus in this algorithm the quantization depends upon the total number of particular color pixel. For this kind of quantization the time complexity is very high.

C. OcTree Quantization

OcTree is a tree kind of data structure, which is used to store the data as per their spatial location. This algorithm [16] is used for color quantization. This algorithm uses the concept of tree data structure, where each node has exactly eight children. Here the nodes are considered as color thus this algorithm identifies 8 quantized colors for the given image. The 8 node of the tree is constructed by considering 4 MSB bit from Red, 2 MSB bit from Green and 2 MSB bit from Blue, so totally 8 bit which were used for OcTree partitioning.

D. K-Mean Quantization (KM)

K-Mean is one of the unsupervised learning techniques used for color quantization effectively [10]. In K-Mean clustering algorithm the color of an image get quantized into K different color which will be always less than 256 thus K is the number of cluster it makes. Initially in this algorithm, K numbers of random centroid pixels are chosen (which is the major drawback of this algorithm). With that centroid pixel, the Euclidean distance between other pixels are determined and grouped as cluster until the clusters got converge between the pixel values.

Algorithm1: K-Mean Color Quantization algorithm
Input: Raw RGB space image
Output: K Color quantized image

1. Begin
2. Let the size of the image be (I x J)
3. Let Q(i,j) be the pixel at the (i,j) position where i represents the x-coordinate of the image and j represents the y-coordinate of the image'
4. Let R(i,j) be the Red color pixel value at that position
5. Let G(i,j) be the Green color pixel value at that position
6. Let B(i,j) be the Blue color pixel value at that position
7. Let the RGB pixel value at the (i,j) position can be represented as Q[i,j,L], where L = 0 for R pixel, L = 1 for G pixel and L = 2 for B pixel.
8. Initially randomly generate 16 different color value [RGB] as the cluster mean value $M_{K=1 \text{ to } 16}$
9. Repeat until identified each cluster got convergence
10. for k : = 0 to 15
11. for i : = 0 to I
12. for j : = 0 to J
13. for L: = 0 to 2
14. if $C(K)= \left|C(i,j,L)- Q(i,j,L)\right|^2 \geq$ Cluster_Threshold of 255
15. Remove that pixel from that cluster or include otherwise
16. end
17. end
18. end
19. end
20. end
21. Find the updated centroid mean of the newly created cluster as
22. for k : = 0 to 15
23. for i : = 0 to I
24. for j : = 0 to J
25. for L: = 0 to 2
26. $M_K = \frac{Q(i,j,L)}{\text{Number of pixel in that k cluster}}$
27. end
28. end
29. end

30. end
31. End

E. Hidden-value learned K-mean Quantization (EMK)

The degradation of the K-Mean algorithm is mainly due to the initial set up. The final color quantized cluster can be re-defined by employing Expectation Maximization learning algorithm [19] to the created clusters.

The two main steps in K-Mean clustering algorithm is to allocate the K centroid value, then to update the value in each iteration until the values got convergence. Using EM algorithm, in E-Step the maximum likelihood expectation of the observed parameter is determined and in M-Step the maximization of the likelihood parameter is done iteratively until the values got converge. An EM algorithm required observed parameter X, Probability model M and the hidden variable Z.

In our work the observed parameter X would be the resized image pixel values, hidden variable Z be the centroid RGB color pixel value of the 16 clusters identified using K Means algorithm and there is a need to design a Probability model M .The Probability model would be the maximum likelihood of the cluster identified. The hidden-value learned K-mean clustering quantization algorithm is discussed below.

Algorithm 2: Hidden-value learned K-Mean Clustering algorithm (EMK)
Input: K Color quantized image
Output: EMK Color quantized image

1. Begin
2. Let the size of the image be (I x J)
3. Let Q(i,j) be the pixel at the (i,j) position where i represents the x-coordinate of the image and j represents the y-coordinate of the image
4. Let the RGB pixel value at the (i,j) position can be represented as Q[i,j,L], where L = 0 for R pixel, L = 1 for G pixel and L = 2 for B pixel.
5. Let C_K be the 16 different color value [RGB] as the cluster mean value determined from Algorithm 1
6. Let the hidden value Z be C_K and the observed value be the original Image
7. In Expectation step, determine the maximum likelihood probability distribution of the each cluster mean value with the original pixel value.
8. Repeat until identified each cluster got convergence
9. for k : = 0 to 15
10. for i : = 0 to I
11. for j : = 0 to J
12. for L: = 0 to 2
13. Probability Model $P(i,j,k)=P\left(\frac{Q(i,j,L)}{C(i,j,L)}\right)$
14. end
15. end
16. end
17. end
18. The Probability P(I,J,K) = 1 if the clusters centroid will not be changed, otherwise new centroid pixel(NC) is calculated in maximization step
19. for k : = 0 to 15
20. for i : = 0 to I
21. for j : = 0 to J
22. for L: = 0 to 2
23. If (P(I,j,K) != 0)
24. $NC(k)= P\left(\frac{P(i,j,k)*Q(I,j,k)}{\text{Number of pixel in that k Cluster}}\right)$
end
25. end
26. end
27. end
28. end
29. Until the point between cluster is minimum the cluster will not be redefine otherwise a adjacent value is updated as the Mean value of the cluster
30. End

III. COLOR EXTRACTION

The most dominant color on an image can be extracted using the RGB pixel values. The HSV kind of model provides better Hue extraction because the values are depend on the Saturation and Brightness of the image. So, here the color quantized image is converted to HSV model. The conversion is mainly depending upon the saturation value S. Algorithm for this conversion [14] is:

Algorithm 3: Prevalent Color Indexing
Input: EMK Color quantized image
Output: A Set of Image indexing in relevant color class {Red, Yellow, Green, Cyan, Blue, Magenta }

1. Begin
2. Normalize the EMK quantized(I x J) image, RGB values between [0 – 1] by dividing the values with 255
3. Let P_x be the maximum of normalized RGB value and P_n be the minimum value of normalized RGB
4. The Brightness value V be P_x ; Define $D = M_x - M_n$
5. If $P_x = 0$ The Saturation value $S = 0$; At this condition Hue is undefined and the pixel is Grey in color it won't reveal the true color of the pixel
6. Else $S = D/ P_x$
7. H is also undefined if $D = 0$
8. H is computed in degree by
    a. If $P_x = R$ and $G > B$ then $H = (60*(G-B))/D$
    b. If $P_x = R$ and $G < B$ then $H = (360 +60*(G-B))/D$ (adding the 360 degree to the value)
    c. If $P_x = G$ then $H = (60*(2+(B-r))/D$
    d. Else $H = (60*(4+(R-G))/D$
9. H can also be computed in decimal value by
    a. $(\sqrt{3}\ (G - B))\ /\ ((2 * R) - G - B)$
10. Compute the total number of color band pixel for 6 color class {Red, Yellow, Green, Cyan, Blue, Magenta } using the H matrix value as shown:
    a. Let P = Total number of pixels in the image
    b. for i : = 0 to I
    c. for j : = 0 to J
    d. if (H(i,j) = = (- 1 to .083) || (.9167 to -1) R_x++
    e. else if (H(i,j) = = (.083 to 0.25) Y_x++
    f. else if (H(i,j) = = (.25 to 0.416) G_x++
    g. else if (H(i,j) = = (.416 to 0.5833) C_x++
    h. else if (H(i,j) = = (.5833 to 0.75) B_x++
    i. else if (H(i,j) = = (.75 to 0.91673) Y_x++
    j. end
    k. end
    l. end
    m. Then compute the total number of each color pixel as
    n. Red color pixel = R_x/P
    o. Yellow color pixel = Y_x/P
    p. Green color pixel = G_x/P
    q. Cyan color pixel = C_x/P
    r. Blue color pixel = B_x/P
    s. Magenta color pixel = M_x/P
11. Prevalent color =Maximum {Red, Yellow, Green, Cyan, Blue, Magenta } of the HSV converted image
12. Save the images as per there maximum color band value in their respective color classes
13. End

Using this algorithm the RGB value of an image is converted into an HSV whose values are normalized between [0 –1] . From these values the total number of pixels in any color band can be computed if the HSV of that particular band is known. In this paper, the prevalent color extraction concept was tested on a Asteroideae flower family datasets. So, the color band choices are depends upon the Asteroideae, a daisy flower family

images. Thus colors such as yellow, red, white, green, cyan, blue and magenta are identified form the given set of images. The Hue value range for the particular colors are calculated as per the given tabulated in Table 1

TABLE I
Color based Hue range

SN	Color	Hue value range
1	Red	H = (-1 to .083) or (.9167 to -1)
2	Yellow	H = .083 to .25
3	Green	H = .25 to.416
4	Cyan	H = .416 to .5833
5	Blue	H = .5833 to .75
6	Magenta	H = .75 to .9167

From the Table I [12], a circular kind of range can be identifies, the reason behind this is that the Hue is an 360^{o} representation of color. The White pixels in the images can be identified by S and V values. If the (S < .05) and (V > .85) those pixel can be calculated as white color band pixel. Thus from the Histogram the color band with maximum value will be labeled as the Prevalent Color of that image.

IV. IMPLEMENTATION ANALYZES

For Implementation purpose Asteroideae flower family image dataset is created with the help of World-wide Biological Universities' lab images. This particular family has 1130 genera [17] and totally more than 16,000 species. For the experimental purpose we have taken nearly 100 species.

Fig.1: Color Quantization

Fig.1 shows the quantization result obtained from each algorithm. To analyze the efficiency of Hidden-value learned K-mean clustering quantization algorithm; the Mean square error, Peak signal to noise, MNormalized cross correlation, Structure content and Normalized absolute error is determined for all the shown algorithms. Fig. 2 shows a flower family Xerunthemum in quantized state.

(a)

(b)

(c)

Fig. 2 (a) Original Image (b) K-Mean quantized image (c) EMK Quantized image

This image quality measure is taken between the original input images to its quantized image. The Mean square error will provide the average square error between the images which has to be low. The Peak signal to noise calculates the signal noise ration which has to be high. The MNormalized cross correlation provide how the images are been align after quantization which should be less than 1. The Structure content check the changes in the lines and edges of the original to quantized image which should also be less than 1 so the originality will not be ruined. The Normalized absolute error checks the scale error between the images. The Table2 provide the average of all this metrics for all the 100 images.

TABLE II
Quality analysis of quantization algorithm

Quantization Methods	**Average**				
	Mean Square Error	**Peak Signal to Noise Ratio**	**MNormalized Cross-Correlation**	**Structural Content**	**Normalized Absolute Error**
Scalar Quantization	218.29	24.986	1.0002	0.981658042	0.136014153
OcTree Quantization	42.620	32.0511	0.9966476	1.003184716	0.052653542
K-Mean Quantization	33.201	33.2560	0.99671	1.003961257	0.042612075
EMK learned Quantization	36.111	32.9487	0.99486	1.007457257	0.035812388

From the Table II the marginal difference between K-Mean and K-Mean with EM can be notified. For certain kind of image the Hidden-value learned K-mean clustering quantization algorithm provides better color extraction metrics as shown in Table III. The chart representation of these values is shown in Fig. 3.

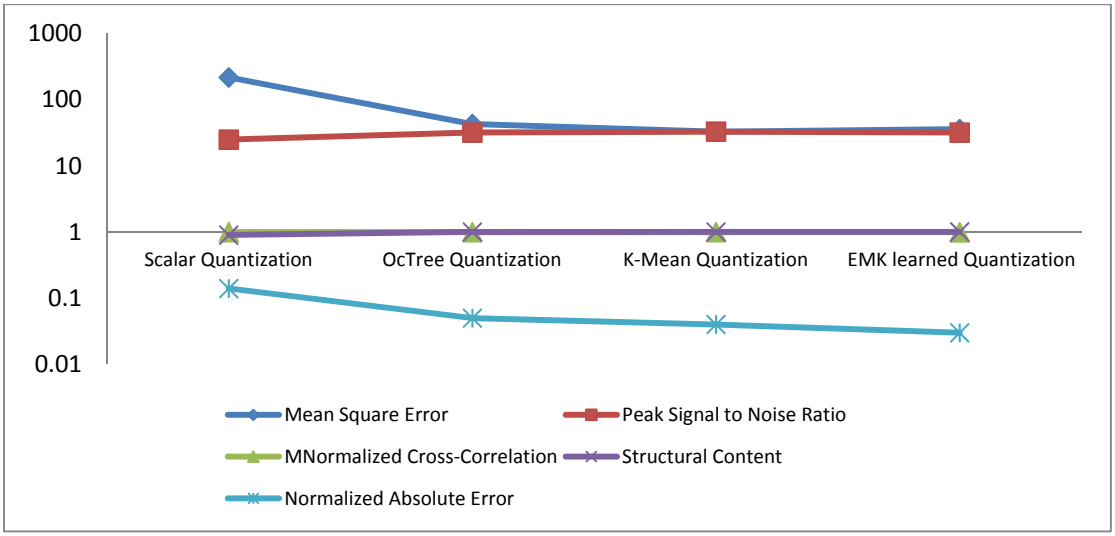

Fig. 3: Quality analysis of quantization algorithm

Hypothetical flower image sets are used to determine the marginal efficiency of K –Mean and EMK color quantization techniques. The flower sets used are {Xerunthemum, Allardia, Oncosiphon and Pentzia}. The color histogram generated from both the quantized method are more or less the same as shown in Fig. 4 but for the flower Allardia K Mean quantized image shows both the red and magenta as the dominant color , which cannot be. As shown in Table I both have different set of H bands.

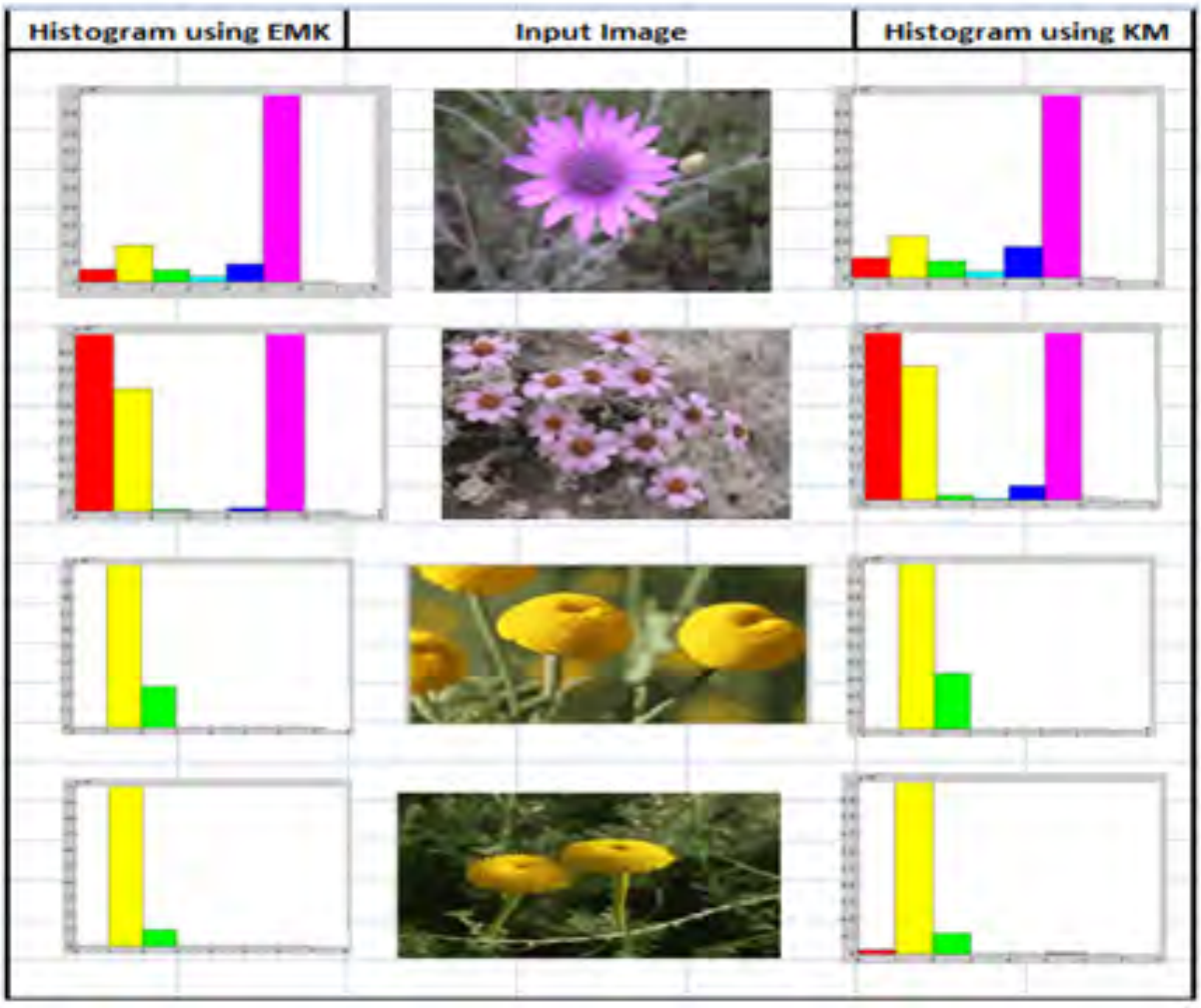

Fig. 4: Histogram analysis of Xerunthemum, Allardia, Oncosiphon and Pentzia

Table III shows the Histogram values of corresponding colors for Allardia flower image, from the variation of Red and magenta histogram values we can justify that by quantizing the image using Hidden-value learned K-mean clustering quantization algorithm the prevalent color is extracted effectively for any user defined images.

TABLE III
Histogram value of Allardia

	Red	Yellow	Green	Cyan	Blue	Magenta	White
EMK	0.00375	0.000692	8.33E-06	0	1.67E-05	0.003742	0
K-Mean	0.006475	0.0008	0.000025	8.33E-06	8.33E-05	0.003517	1.67E-05

So, to index an image as per its prevalent color, classes of 7 different colors are to be created. If an image has yellow as its prevalent color then the image will be saved under the yellow class. Such a way the Asteroideae flower image dataset is indexed as per there dominating colors. Fig.5 shows the indexed images as per there color domain.

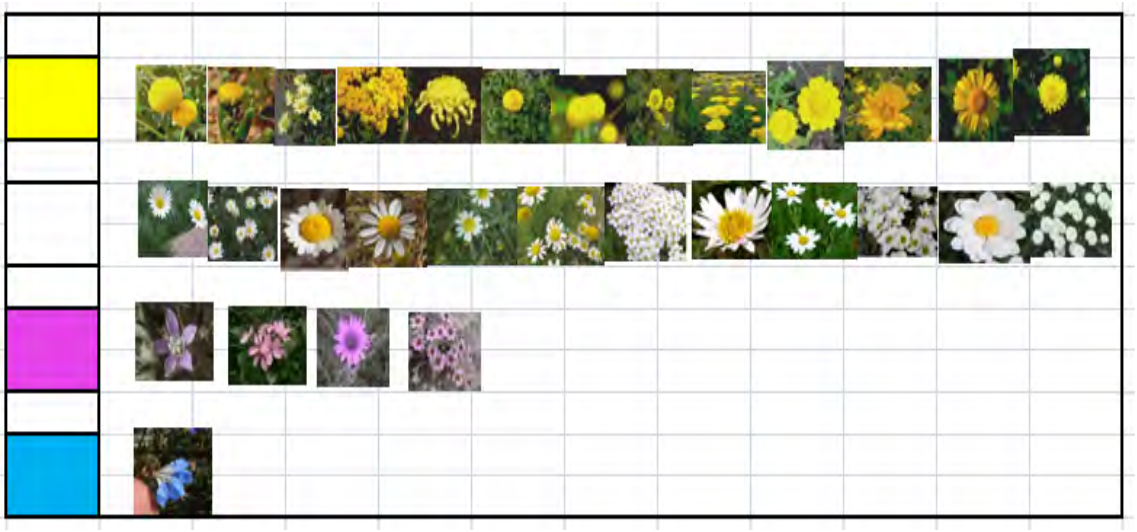

Fig. 5: Indexed images

V. CONCLUSION

In this paper of prevalent color extraction and indexing, Asteroideae flower images are quantized using Hidden-value learned K-mean clustering quantization algorithm which is adaptable for any kind and nature of images. From the quantized image the seven major prevalent colors are identified from the HSV space mode images. Images are then indexed as per there highest color histogram values. This indexed class images can be further used in an ontological structure for the Asteroideae flower domain feature ontology in our forthcoming work. Once the dominant color of an image is identified the images can be segmented and analysis for image retrieval purpose.

REFERENCES

[1] Guang Hai Lie and Jing Yu Yang "Content based image retrieval using color difference histogram" *Pattern Recognition* Vol:46 issue 1,Jan 2013 Pg 188 – 198,2013.

[2] Yong Rui and Thomas S.Huang "Image Retrieval:Current Techniques, Promising Directions and Open Issues" *Journal of Visual communication and Image Representtion*,No:10,Pg:39-62,1999.

[3] R. Brunelli and O. Mich, "Histograms Analysis for Image Retrieval", *Pattern Recognition* 34, pp., 1625-1637,2001.

[4] Y. Deng, B. S. Manjunath, C. Kenney, M. S. Moore and H. Shin "An Efficient Color Representation for Image Retrieval", *IEEE Transactions on Image Processing*, 10, pp., 140-147,2001.

[5] B.S.Manjunath,Jens-Rainer Ohm,Vinod V.Vasudevan and Akio Yamada "Color and Texture Descriptor" *IEEE Transaction on Circuits and Systems for video Technology*,Vol.11,No.6 Pg:703-715,2001.

[6] John P.Kakins "Towards intelligent image retrieval" *Pattern Recognition* , No:35 Pg:3-14,2002

[7] H.Yu, M. Li, H J. Zhang and J. Feng,"Color Texture Moments for Content-Based Image Retrieval", *Proc. Int. Conference on Image Processing*, Volume III, pp., 929-931, 2002.

[8] Mathias Lux"Revisting the Vector Retrieval Model in Context of the MPEG 7 Semantic Description Scheme" *Ninth International Workshop on Image Analysis for Multimedia Interactive Service*,2008.

[9] Mathias Lux, Michael Granitzer and Werner Klieber "Caliph & Emir: Semantics in Multimedia Retrieval and Annotation" *Proceedings for the 19th International CODATA Conference The Information Society: New Horizons For Science*,2004.

[10] Yuchou Chang, Dah-Jye Lee, Yi Hong, James Archibald and Dong Liang " A Robust color Quantization Algorithm based on Knowledge reuse of K-Means clustering Ensemble" *Journal of Multimedia*,Vol:3, No2, Pg:20 – 27,2008.

[11] Stuart Russell and Peter Norvig "*Artificial Intelligence A Modern Approach*" II Ed. Pearson Education Asia,2011.

[12] S.Jayaraman,S.Esakkirajan and T.Veerakumar "*Digital Image Processing*" Tata McGraw Hill Education Private Limited,2009.

[13] S.Sridhar "*Digital Image processing*" Oxford University Press,2011

[14] R. C. Gonzalez and R. E. Woods, "*Digital Image Processing* ", II Ed. Pearson Education Asia, First Indian Reprint,2002
[15] Heckbert P "*Color image quantization for frame buffer display*" Proceedings of the 9th annual conference on Computer graphics and interactive techniques Pages 297-307,1982
[16] Henning Eberhardt, Vesa Klumpp, Uwe D. Hanebeck,"Density Trees for Efficient Nonlinear State Estimation", *Proceedings of the 13th International Conference on Information Fusion, Edinburgh, United Kingdom*, July, 2010
[17] Angiosperm Phylogeny Group . "An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II". *Botanical Journal of the Linnean Society* 141 (4): 399–436,2003.
[18] Esteban J.Palomo and Enrique Dominguez"Hierarchical color quantization based on Self organization" *Journal of Mathematical Imaging and vision*, March 2013
[19] Brijnesh J.Jain "Mixtures of Radial Densities for clustering Graphs" *Computer Analysis of Images and Patterns* LNCS Vol 8047, 2013, PP110-119,2013
[20] K.K.Thyagharajan and R.Harikrishnan "Contentbased Bandwidth Aware Hierarchical video summarization" *International Journal of computer science,Systems engineering and Information Technology*" Vol:2 Number:2 PP:197-203,2009.
[21] K.K.Thyagaharajan , G.Nagarajan "Semantically Effective Visual Concept Illustration for Images" *5th International Conference on Information and Multimedia Technology* (ICIMT 2013), December 6-7, 2013, Sydney, Australia .
[22] G.Nagarajan, Dr.K.K.Thyagaharajan(2013) "Conceptual Visual Image Classification using Statistical Learning Model", National Conference on Recent Trends in Mathematical Computing(NCRTMC'13),VIT University.ISBN:97893-82338-680.August 2013pp 410-416.
[23] R.I.Minu, Dr.K.K.Thyagharajan(2012) " A Novel Approach to Build Image Ontology using MPEG 7" "INTERNATIONAL SYMPOSIUM ON INTELLIGENT INFORMATICS" published by Springer's INTELLIGENT INFORMATICS, Advances in Intelligent Systems and Computing,2013,Volume 182 Pg:333-339
[24] http://en.hortipedia.com/wiki/Main_Page

Dr. K.K. Thyagharajan obtained his B.E., degree in Electrical and Electronics Engineering from PSG College of Technology (Madras University) and received his M.E., degree in Applied Electronics from Coimbatore Institute of Technology. He also possesses Post Graduate Diploma in Computer Applications from Bharathiar University. He obtained his Ph.D. degree in Information and Communication Engineering (Computer Science) from College of Engineering Guindy, Anna University. He has twenty five years of teaching experience. Now he is the Dean (Academic) of R.M.D. Engineering College. He has written 5 books in Computing including "Flash MX 2004" published by McGraw Hill (INDIA) and it has been recommended as text and reference book by universities and Polytechnics. He has published more than 70 papers in National and International Journals and Conferences. He is a grant recipient of DST & Tamil Nadu State Council for Science and Technology. His biography has been published in the 25th Anniversary Edition of Marquis Who's Who in the World. He has been invited as chairperson and delivered special lectures in many National and International conferences and workshops. His current interests are Multimedia Networks, Content Based Information Retrieval, Web services, Data Mining, e-learning, Image Processing,. He is reviewer for many International Journals and Conferences. He is a recognized supervisor for Ph.D by Anna University Chennai,MS University, JNTU and Sathyabama University, and now 11 students are doing Ph.D. under his supervision.

Mrs.R.I.Minu obtained her B.E., degree in Electronics and Communication Engineering from Mookambigai College of Engineering (Bharathidasan University) and received her M.E., degree in Computer Science and Engineering from MNM Jain Engineering College (Anna University) .She is currently pursuing Ph.D(Part Time) at Anna university. She has 6 year of teaching experience. Now she is working as Assistant Professor in the Department of Computer Science and Engineering in Jerusalem College of Engineering. She has published more than 15 papers in National/International Journals and Conferences. She has involved in many UG and PG projects in the area of Image Processing, E-Learning, Web Services and Artificial Intelligent.