Color Feature Extraction in Content based Image Retrieval based on Quaternion Space

Vinayak Gajanan Kottawar
Assistant Professor, Department of Computer Science & Engineering
M.G.M's College of Engineering, Nanded
Maharashtra, India

A.M.Rajurkar, Ph.D
Professor & Head, Department of Computer Science & Engineering
M.G.M's College of Engineering, Nanded
Maharashtra, India

ABSTRACT

Interest in the digital images has increased a lot over the last few years, but the process of locating a desired image in such a large and diverse image collection becomes very difficult. Traditionally text in different languages is used for efficient retrieval of images; it has several drawbacks such as language constraint and subjectivity of human perception.

Content-based image retrieval is a technique which uses visual contents such as color texture and shape to search images from large image databases according to user's desire. Color is the most commonly used feature for content based image retrieval. In many applications color histogram is used to represent extracted color features. The important drawback of usual color histogram based method is that, it does not take image color distribution into consideration and inflexibly partition the color spaces into a fixed number of bins.

In this paper we propose a moment-preserving technique based on binary quaternion space for feature extraction. It aims to extract color features according to the image color distribution that effectively reduces the distortion incurred in the feature extraction process. We also propose an efficient clustering based algorithm to compare similarity between two histograms.

It is observed that minimizing the distortion incurred in the extraction process can improve the accuracy of retrieval. Our experimental results show that the proposed extraction methods can improve the average retrieval precision rate by a factor of 25% over that of a color histogram based feature extraction method (binning method). It is also observed that, this technique effectively reduces the average retrieval time.

Keywords

Content based image retrieval, Quaternion moment, Histogram, Clustering based histogram comparison

1. INTRODUCTION

Many times we imagine of an image which we desire, but sometime we can't express this desire in precise words. It is easier to find such an images by looking through the collection of images and selecting one which matches with the image drawn by imagination. The textual descriptions fail to confine the perfection in such a situation.

Content-based image retrieval is a technique which uses visual contents of an image to search the preferred images from large scale image databases according to user's desire. The visual content of an image is extracted and described by multidimensional feature vectors. Users provide example image to the retrieval system, the system then converts these examples into representation of feature vectors. The similarity distances between the feature vectors of the query example and those of the images in the database are then calculated and retrieval is performed.

The years 1994-2000 is considered as the early phase of research and development of image retrieval by content. Many techniques and approaches were proposed for fast and efficient retrieval of images. Color is the common most features for content based image retrieval. QBIC [12], Pictoseek [3] are some of the popular systems which uses color histograms as one of the feature for feature extraction and image retrieval.

In color histogram based method, color features are extracted to represent image color content and represented as color histograms. The histogram based methods inflexibly partition the color spaces into a fixed number of bins, each of which corresponds to a bin in the histogram. The extraction process maps pixels into their closest color bin. The weight of a bin in turn denotes the percentage of pixels in an image belonging to that bin. Therefore, a color histogram can be considered as a quantized color distribution of an image.

The important drawback of these methods is that it do not take the image color distribution into consideration while deciding the number of bins in a color histogram, it uses the same set of representative colors for every image. Therefore, color histogram based method provides little compliance to the color content of an image, and color features may get heavily distorted if a less number of bins are used. To overcome this problem, we proposed a new feature extraction technique, Quaternion moment preserving [QMP], which is based on quaternion space. It takes image color distribution into consideration during feature extraction process and reduces the distortion which may occur during this process, by preserving a moments up to third moment.

If we consider the value of a pixel in an image as a random vector, the color distribution of this image will be the probability distribution of this random vector [5].With this view, we applied the quaternion-moment-preserving technique [QMP] to the problem of feature extraction.

After extraction of features from an image, feature vector are compared to find the distance between two feature vectors. Usually earth mover's distance (EMD)[19] is used to calculate the distance between two distributions. The EMD method models the problem of finding distances as the transportation problem in linear programming [9] and the distance is then calculated by finding the best flows which transform one distribution to the other. However, the computation of the EMD is costly It has an exponential time complexity [12]. To lessen this problem, we devised a new distance measure based on clustering, to improve the efficiency of measuring distances between histograms. The proposed distance measure, Clustering Based Histogram Comparison (CBHC) is an effective greedy approach, utilizes the complete link method [11] to group similar colors in histograms into a cluster and defines the inconsistency in a cluster. The total distance between two histograms is defined as the normalized sum of the inconsistency over all clusters.

The main contribution of this paper is that we propose an adaptive color feature extraction scheme by preserving color distributions up to the third moment and we devise an efficient and effective distance measure to find the similarity between the resulting color histograms.

The rest of this paper is organized as follows. The new color extraction methods are proposed in Section II, the new distance measure is introduced in Section III. Section IV contains the experimental results. We conclude this paper in Section V.

2. COLOR FEATURE EXTRATION USING QMP

In this section, we will define the problem of color feature extraction and describes the QMP thresholding technique. Let *I* is an image and *a* be a pixel. The color feature extraction can be defined as a function *F: I→Q* where *Q* is a set of representative colors. *F* maps a pixel *a* to representative color. Color histogram represents the percentage of pixels in *I* which are mapped into *Q*. If we consider the value of pixel in *I* as a random vector, *H* represents the quantized probability distribution of *I*.

Conventional extraction methods fix *Q* for all images without considering their color distribution. The main idea of this paper is to find a proper function *F* for each image according to their color distribution which is derived with a moment preserving scheme in quaternion space.

2.1 Quaternion Space

Each quaternion number can be denoted as

$$q=q_0+q_1.i+q_2.j+q_3.k \quad (1)$$

Where *I, j* and *k* are the operation units of quaternion number.

In our problem, color values R, G, B can be treated as a quaternion with q_1 = R, q_2=G, q_3=B and q_0=0. Based on the definition of the quaternion, the first three orders of quaternion moments are defined as follows:

$$m1=E\ (q_0)\ +E\ (q_1).i\ +E\ (q_2).j\ +\ E\ (q_3).k \quad (2)$$

$$m2=E\ (q_0^2+\ q_1^{2+}\ q_2^2+\ q_3^2\ ) \quad (3)$$

$$\begin{aligned} m3=E\ &(q_0^3+q_0q_1^2+q_0q_2^2+q_0q_3^2) \\ &+ E\ (q_1q_0^2+q_1^3+q_1q_2^2+q_1q_3^2)*i \\ &+ E\ (q_2q_0^2+q_2q_1^2+q_2^3+q_2q_3^2)*j \\ &+ E\ (q_3q_0^2\ +q_3q_1^2+q_3q_2^2+q_3^3)*k \end{aligned} \quad (4)$$

Where E represents the sample mean

The problem of QMP thresholding in a quaternion valued data set is to select a hyperplane A as a threshold, such that if those below-threshold data points and those above threshold data points are replaced by the representative Z_0 and Z_1 respectively, and the first three quaternion moments are preserved in the resultant two-level data set.

$$Z_o=Z_{00}+Z_{01}*\ i+Z_{02}*j+Z_{03}*k$$

$$Z_1=Z_{10}+Z_{11}*i+Z_{12}*j+Z_{13}*k \quad (5)$$

The authors in [18] derived the closed form solution of z_0 and z_1 and assumed that the first moment is 0.

2.2 QMP Feature Extraction Technique

The QMP thresholding technique extracts color features in two ways. In fixed cluster technique (FC), we will extract fixed number of pixel clusters from an image. In Variable cluster technique (VC), we will extract a variable number of pixel clusters from an image, and the number of pixel clusters extracted depends on the intra-cluster variances.

The procedure of fixed cluster (FC) and variable cluster

(VC) can be summarized in the following three steps:

1) Take data set as a input

2) Find a sub cluster which can be divided and whose variance is maximum and then use the QMP thresholding technique to divide the sub cluster into two new sub clusters. If further splitting is not possible, mark it.

3) If there exist a cluster, which can be further divided:

(a) In the case of FC, repeat Step 2 until exactly N clusters are formed;

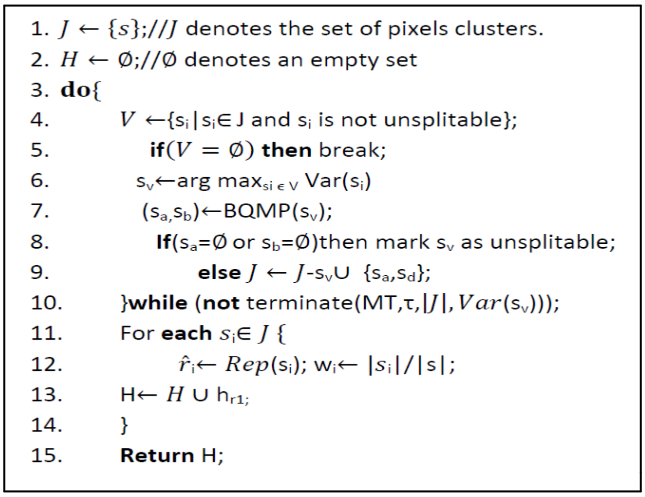
(b) In the case of VC, repeat Step 2 until the variance in each sub cluster is below a variance threshold.

Splitting process of FC and VC resembles to the splitting process of a binary tree. Splitting process of clusters resembles with a splitting process of a leaf node in a binary tree. The root corresponds to the initial multiset of all pixels. For each non terminal node, its left child represents split sub cluster whose representative is Z_0 and its right child represents for the other sub cluster whose representative Z_1 .FC and VC differ from each other in their conditions of termination of extraction process. FC completes the process when a fixed number of pixel clusters have been extracted. VC completes this process when the numbers of pixel clusters extracted so far are sufficient to represent the image. We define such sufficiency by the condition that the variance in each pixel cluster is below a variance threshold T_V, where the value of T_v is predefined and empirically calculated.

Algorithm for QMP based feature extraction:

Input: type of extraction method, multiset s and termination parameter

Output: A color histogram.



```
$J \leftarrow \{s\}$;//$J$ denotes the set of pixels clusters.
$H \leftarrow \emptyset$;//$\emptyset$ denotes an empty set
do{
        $V \leftarrow \{s_i | s_i \in J$ and $s_i$ is not unsplitable};
            if($V = \emptyset$) then break;
          $s_v \leftarrow \arg\max_{si \in V} Var(s_i)$
           $(s_{a,}s_b) \leftarrow BQMP(s_v)$;
             If($s_a=\emptyset$ or $s_b=\emptyset$)then mark $s_v$ as unsplitable;
                 else $J \leftarrow J$-$s_v \cup \{s_a,s_d\}$;
        }while (not terminate(MT,τ,$|J|$,$Var(s_v)$));
        For each $s_i \in J$ {
                $\hat{r}_i \leftarrow Rep(s_i)$; $w_i \leftarrow |s_i|/|s|$;
        H← $H \cup h_{r1;}$
        }
        Return H;
```

2.3 Comparison with the Binning Methods

FC and VC preserve the color distribution of images, and therefore, color features extracted by them are less distorted than those extracted by the binning method.

FC and VC finds the representative colors based on the color distribution of an image while the binning methods do not provide such litheness. An image is usually divided into several blocks and regional color features are extracted. In this case, our extraction methods will be more suitable than binning methods because the color content in a sub block tends to be conquered by only a few colors. Our extraction methods are able to flexibly extract those colors while there is no clear extension that will make possible the binning methods to do that.

For a binning method, the extraction completes in one iteration and has a running time of O(S). In contrast, FC extracts color features in time O (NS). However, as discussed earlier, the average computing time of FC is in fact much smaller than that in the worst case. Moreover, when the regional color features, whose color contents are highly homogeneous, are extracted, the computing time of FC and VC will be further closer to that of the binning method.

3. DISTANCE MESURE : CLUSTRING BASED HISTOGRAM COMPARISON

The resulting histograms from FC or VC consist of different sets of extracted colors derived from the extraction process. The EMD measure is applicable to these color features. However, the computation of EMD is costly because It has an exponential time complexity [12]. To ease this problem, we devise a new distance measure, *Clustering based histogram comparison (CBHC)*, abbreviated as *CBHC*. The time complexity of *CBHC* is $O\ (n^2\ log\ n)$ where n is the sum of cardinalities of two histograms to be compared.

CBHC is a clustering based algorithm. The distance between two points P_u and P_v, is defined as the Euclidean distance between their representative colors r_u and r_v

$$dist\ (p_u,p_v)=||r_u-r_v|| \qquad (6)$$

CBHC utilizes the complete link method [11] to group similar clusters. The inter-cluster distance in a complete link method is defined as the maximum distance between one member in one cluster and another in the other cluster

$$dist\ (c_u, c_v)=max(dist(p_u, p_v)|\ p_u € c_u, p_v € c_v) \qquad (7)$$

Where c_u and c_v are two clusters and p_u and p_v are points in them. With each iteration, two clusters with the minimum distance among all pairs are merged. As a result, the complete link method manages to minimize the diameter of a cluster during the clustering process and the diameter of a cluster represents the closeness of points in that cluster. In view of this, CBHC limits the diameter of a cluster within a predefined threshold T_d by terminating the clustering process when the minimum distance between clusters is about to exceed T_d. Hence, T_d is so defined to be the distinguishing power between similar colors.

The procedure of CHIC can be summarized as follows:

1. Map each H_r in H_q and H_t to a point $P = (r,w,img)$. In the clustering problem , where $w=hr$ and img =
   a. $‘q’ , h_r € H_q$
   b. $‘t’ , hr € H_t$
2. Cluster $\{p1,p2,p3….p|H_q|+|H_t|\}$, by complete link method under the constraint that the diameter of each cluster is not larger than T_d
3. Return the distance between H_q and H_t , denoted by $D(H_q, H_t)$, by the normalized sum of discrepancies in all cluster

3.1 Comparison between CBHC and EMD

CBHC is a more efficient distance measure than the EMD because CBHC is greedy in nature. The EMD finds the optimal flow by iteratively testing a feasible set of flows and the total number of iterations may grow exponentially with the input size. In contrast, the running time of CBHC is bounded by O(n^2logn) .For the physical meaning of these two distance measures, the EMD is defined as the amount of work to transform one distribution to the other, and the transformation can be made by fractionally moving earth. In contrast, CBHC in a sense only moves a lump of earth because each point p_i in the distributions is an entity to be clustered. Therefore, in the case of histograms with a small cardinality, each p_i tends to have a large weight w and fractionally moving makes the EMD more accurate. On the contrary, in the case of histograms with a large cardinality, each p_i tends to have a small weight, and the advantage of fractionally distributing weights diminishes. As observed in our experiments, CBHC, which defines a similarity threshold on colors, emerges as a more effective distance measure.

4. EXPERIMENTAL RESULTS

To evaluate the performance of a CBIR system, we have used a database of 1000 images collected from image albums published by Corel Corp. These images consist of six image classes, each of which contains around 200 images. In other words, each image has a known class identity. Such as (a) red roses (b) seas, (c) farms, (d) mountains (e) lands & skies. The experiments are conducted on a PC with a Pentium-I5 with 2.8 GHz CPU and 4 GBytes RAM running the Windows-XP OS.

The performance of a color extraction method is measured in terms of evaluation measures such as F-measure, NDCG and average retrieval precision [26]. Each time a query image is selected from the database to retrieve 25 best matched images, excluding the query image itself, from the database. Table 1 summarized the performance of fixed clustering method, where N is the fixed number of clusters formed during feature extraction process.

Table 1.Retrieval performance of fixed cardinality approach

Parameters/N	**16**	**32**	**64**	**128**	**256**
Precision	0.52	0.56	0.56	0.56	0.60
Recall	0.81	0.88	0.88	0.88	0.90
F measure	0.63	0.68	0.68	0.68	0.73
$f_β$	0.56	0.60	0.60	0.60	0.65
Average Precision	0.58	0.65	0.71	0.74	0.79
Execution time(ms)	6	7.2	9.45	12.32	14.20
Ndcg	0.84	0.85	0.88	0.88	0.89

Table 2 summarized the performance of variable clustering method, which is calculated on different concepts such as red rose, sea, farms, mountains & rivers and lands and skies.

We first compared the average retrieval precision rates of fixed clustering and variable clustering methods. The quadratic-form distance [5] is used as a distance measure for color histograms obtained by the binning method. As suggested in [24], the L1 metric is used as a distance measure for color channel moments. We set the diameter threshold T_d of CBHC as 30 when the cardinality $N=16$ and 32, and T_d as 20 when $N=64$, 128 and 256. Note that for the technique of color channel moments, only the moment values are used to represent color features, and the cardinality value shown in Fig. 1 does not apply to it. We summarize the results obtained in Fig. 1 as follows:

Table 2 Retrieval performance of Variable cardinality approach

Parameter/ Concept	Red Rose	Sea	farms	Mountain& Rivers	Land & Skies
Precision	0.56	0.60	0.68	0.40	0.44
Recall	0.88	0.83	0.85	0.83	0.85
F measure	0.68	0.7	0.76	0.54	0.58
fB	0.60	0.64	0.71	0.45	0.49
Average Precision	0.73	0.67	0.79	0.69	0.74
NDCG	0.88	0.90	0.89	0.92	0.88
Average Threshold	110	102	154	132	122
Execution Time	10.2	9.88	11	10.8	10.3

- It is observed that FC increases the precision rate by the factor of 25 % as compare to the binning method when 16 bins are extracted. Even when 256 colors are extracted, FC still achieves a higher precision rate than the binning method by 13%. Therefore, the precision rate is mainly conquered by the extraction methods instead of the distance measures. This result shows that representing color features in a precise manner is very helpful for the image retrieval applications.
- When cardinality (N) is below 64, the method of channel moments perform better than the binning method. These shows that the moments are a compact representation for color features

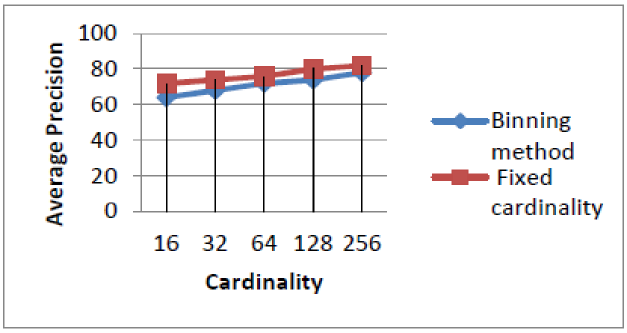


Fig. 1 Comparison of average precision of Binning method and fixed cardinality approach

Fig 2 shows the query image and the resultant landscape images retrieved by the system.

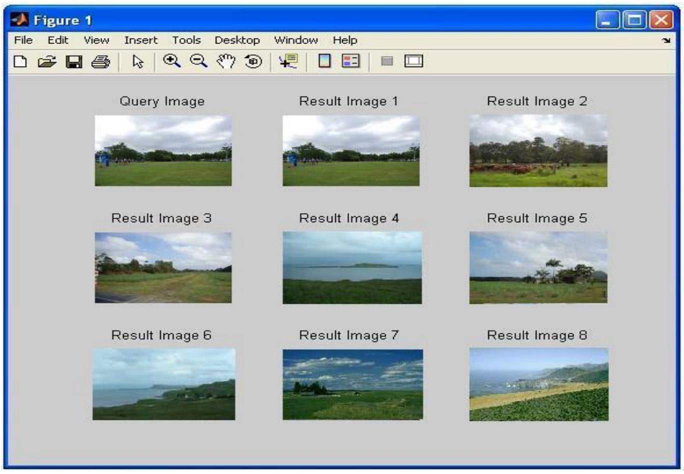


Fig. 2 Query image and retrieved landscape images

The execution time of CBHC and that of the EMD are compared in Fig. 2. The execution time is in a logarithmic scale. We set the two histograms which we compared to have same cardinality. The execution time of the EMD grows extremely fast as the cardinality grows while CBHC significantly reduces the execution time required. As expected, CBHC is a more efficient distance measure than the EMD.

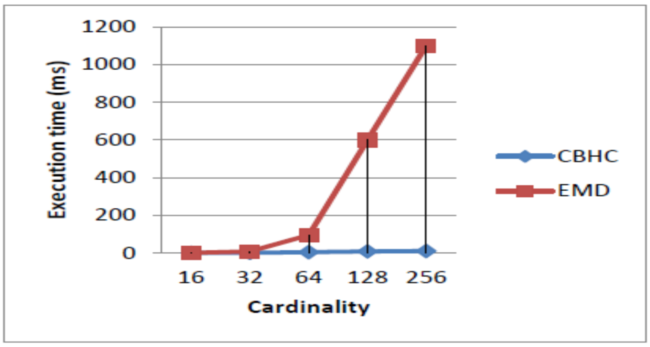


Fig.3 Comparison of execution time of CBHC and E

4. CONCLUSION

We proposed a color feature extraction scheme (QMP), which preserves the color distributions up to the third moment. It is observed that reducing the distortion incurred in the feature extraction process can improve the accuracy of image retrieval system. The experimental results show that the new extraction methods can achieve a significant improvement over the conventional color feature extraction methods.

We also proposed an effective distance measure (CBHC), which is greedy in nature and utilizes complete link method to group similar clusters to find the similarity between the two feature vectors.

An image is generally divided into several sub blocks and the regional color features are extracted. In this case, our extraction method and distance measure is observed to be more effective than the conventional binning methods because the color content in a sub block tends to be conquered by only few colors. Further, the variable clustering extraction method, which terminates the extraction process based on the heterogeneity of the pixel values, is achieving a balance between articulacy and compactness.

5. REFERENCES

[1] S. Antani, and R. Jain, “A survey on the use of pattern recognition methods for abstraction, indexing, and retrieval of images and video,” Pattern Recognit., vol. 35, no. 4, pp. 945–965, 2002.

[2] L. Brown, “Tree-based indexes for image data,” Vis. Commun. Image Represent., vol. 9, no. 4, pp. 300–313, 1998.

[3] W. H. Day, “Efficient algorithms for agglomerative hierarchical clustering methods,” J. Classificat., vol. 1, pp. 1–24, 1984.

[4] D. Defays, “An effecient alogrithm for a complete link method,” Comput. J., vol. 20, no. 4, pp. 364–366, 1977.

[5] M. Flickner, H. Sawhney, “Query by image and video content: The QBIC system,” IEEE Computer, vol. 28, no. 9, pp. 23–32, Sep. 1995.

[6] J. B. Fraleigh, “A First Course in Abstract Algebra” Reading, MA: Addison-Wesley, 1982.

[7] H. Frigui, “Visualizing and browsing large image databases,” in Proc. Int. Conf. Information and Knowledge Engineering, 2004, pp. 68–74.

[8] R. M., “Quantization,” IEEE Trans. Inf. Theory, vol. 44, no. 6, pp. 2325–2383, Nov. 1998.

[9] F. S. Hiller, “Introduction to Mathematical Programming”. New York: McGraw-Hill, 1990.

[10] J. Huang, “An automatic hierarchical image classification scheme,” in Proc. ACM Int. Conf. Multimedia, 1998, pp. 219 228.

[11] B. King, “Step-wise clustering procedures,” J. Amer. Statist. Assoc., vol. 69, pp. 86–101, 1967.

[12] V. Klee, “How good is the simplex algorithm,” in Inequalities, 1972, vol. 3, pp. 159–175.

[13] A. Kushki, “Query feedback for interactive image retrieval,” IEEE Trans. Circuits Syst. Video Technol., vol. 14, no. 5, pp. 644–655, May 2004.

[14] F. Long, H. Zhang, “Fundamentals of content-based image retrieval,” in Multimedia Information Retrieval and Management Technological Fundamentals and Applications. New York: Springer-Verlag, 2003.

[15] M. Oge and F. Borko, “Muse: A content-based image search and retrieval system using relevance feedback,” Multimedia Tools Appl., vol. 17, pp. 21–50, 2002.

[16] A. Papoulis, “Probability, Random Variables, and Stochastic Processes”. New York: McGraw-Hill, 2002.

[17] S.-C. Pei and C.-M. Cheng, “Color image processing by using binary quaternion moment-preserving thresholding technique”, IEEE Trans. Image Process., vol. 8, no. 5, pp. 614–628, May 1999.

[18] Y. Rubner, C. Tomasi, and L. J. Guibas, “A metric for distributions with applications to image databases,” in Proc. IEEE Int. Conf. Computer Vision, 1998, p. 59.

[19] Y. Rui, T. S. Huang, and S.-F. Chang, “Image retrieval: Current techniques, promising directions, and open issues,” J. Vis. Commun. Image Represent., vol. 10, no. 1, Mar. 1999.

[20] A. Smeulders, M.Worring, S. Santini, A. Gupta, and R. Jain, “Contentbased image retrieval at the end of the early years,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 12, pp. 1349–1380, Dec. 2000.

[21] J. R. Smith and S.-F. Chang, “VisualSeek: A fully automated contentbased image query system,” in Proc. ACM Int. Conf. Multimedia, Nov. 1996, pp. 87–98.

[22] J. R. Smith and C. Li, “Image classification and querying using composite region templates,” Comput. Vis. Image Understand., vol. 75, pp. 165–174, 1999.

[23] M. Stricker and M. Orengo, “Similarity of color images,” in Proc. Storage and Retrieval for Image and Video Databases (SPIE), 1995, pp. 381–392.

[24] M. J. Swain, “Color indexing,” Int. J. Comput. Vis., vol. 7, pp. 11–32, 1991.

[25] E. M. Voorhees, "The Philosophy of Information Retrieval Evaluation," in: Evaluation of Cross-Language Information Retrieval Systems, Lecture Notes in Computer Science 2001, pp. 143–170.