A SIMILARITY MEASURE FOR COLOR IMAGE RETRIEVAL AND INDEXING BASED ON THE MULTIVARIATE TWO SAMPLE PROBLEM

Christos Theoharatos, Nikolaos Laskaris, George Economou and Spiros Fotopoulos

Electronics Laboratory, Department of Physics, University of Patras
Patras, Rio 26500, Greece (Europe)
phone: +30 2610 997287, fax: +30 2610 997456, email: {htheohar, nlaskar, economou, spiros}@physics.upatras.gr
web: http://www.ellab.physics.upatras.gr/Personnel/Faculty/Homepages/SFotopoulos/spiros.html

ABSTRACT

In this work, a similarity measure in the feature space is proposed for color retrieval and indexing based on the *"Multivariate Two-Sample Problem"*. Color information is extracted via random selection of image pixels from high-density regions. The proposed scheme has a global nature due to its randomness and is easy to implement. It makes uses of the minimal spanning tree (MST) structure and properties, providing the retrieval results with a statistical measure of their significance level. The main advantages of our proposal are its computational efficiency and the fact that it is generally applicable to natural image collections.

1. INTRODUCTION

During the past decade large collections of digital libraries are being created due to the low cost of digital storage and the rapid growth of computational power. The increasing number of images has led to the need for developing powerful tools for searching through such image databases. Automatic retrieval of an image from a whole data set using representative text-annotation comes from the early 1970's. However, the limitations of this conventional method forced the development of content-based retrieval systems [1]. These systems make use of low-level features such as color, shape and texture to represent the image content. Among others, MPEG-7 standard [2] defines a set of visual descriptors for image content representation designed to meet the requirements of different application domains.

Considerable research has been carried out on the basis of color features [1,3], since they are scale and rotation invariant and due to their robustness to background complications. The most popular technique for representing color information is the global histogram. A lot of similarity measures have been proposed in the literature for comparing color distributions, such as histogram intersection [4], L_2-norm and cumulated color histogram [5]. Furthermore, several other color features have been used as color descriptors, including color moments [5] and color sets [1]. These last approaches, although providing the user with better performance, are time consuming regarding the preprocessing stage.

In this work, the use of a nonparametric test dealing with the *"Multivariate Two-Sample Problem"* is being adopted and used for expressing color image similarity. The specific test is a multivariate extension of the *Wald-Wolfowitz test* and compares two different samples of vectorial observations (i.e. two sets of points in $\mathbf{R^P}$) by checking whether they form different branches in the overall MST [6]. The output of this test can be expressed as the probability that the two point-samples are coming from the same distribution. Its great advantage is that no a-priori assumption about the distribution of points in the two samples is a prerequisite.

The remainder of this paper is organized as follows. In Section 2, the theoretical framework of MST and the multivariate Wald-Wolfowitz test for the two-sample problem are described. The feature extraction process is explained in Section 3. Experimental results including a short discussion are presented in Section 4. Finally, conclusions are drawn in Section 5, along with an outline of our future objectives.

2. THEORETICAL FRAMEWORK

2.1 Minimal Spanning Tree (MST)

Graph theory sketches the MST structure with the following definitions. A *graph* is a structure for representing pairwise relationships among data. It consists of a set of *nodes* $V=\{V_i\}_{i=1:N}$ and a set of links $E=\{E_{ij}\}_{i\neq j}$ between nodes called *edges*. The *degree* d_i of a node is the number of edges incident to it. When a weight e_{ij} is assigned to each link, a weighted-graph is formed and in the particular case that $e_{ij}=e_{ji}$ this graph is called *undirected weighted graph*. A *tree* is a connected graph with no cycles. A *spanning tree T* of a (connected) weighted graph G(V,E) is a connected subgraph of G(V,E) such that: (i) it contains every node of G(V,E), and (ii) it does not contain any cycle. The *MST* is a spanning tree containing exactly (N-1) edges, for which the sum of edge weights is minimum.

When the previous notions are employed for the color description of an image, N pixels are selected and based on the corresponding RGB-vectors these pixels are represented as points in $\mathbf{R^3}$. The specific points are then used as the nodes of the original (fully-connected) graph, while the interpoint Euclidean distances as the weights of the corresponding edges. Finally, using a standard algorithm, the MST is delineated from the original graph and used as a parsimonious description of the color variation in the image. Given two images, the contrast of their color content is then transformed to a comparison of the corresponding MST-graphs. To perform such a comparison, a well-defined statis-

tical test is available in the literature of multivariate statistics. A short description of this test is provided in the sequel.

2.2 The multivariate Wald-Wolfowitz test (WW-test)

Given two multidimensional point samples $\{X_i\}_{i=1:m}$ and $\{Y_i\}_{i=1:n}$, the hypothesis H_o to be tested is whether they are coming from the same multivariate distribution. At first, the sample identity of each point is not encountered and the MST of the overall sample is constructed. Then, based on the sample identities of the points, a test statistic R is computed. R is the total number of *runs*, while a *run* is defined as a consecutive sequence of identical sample identities. Rejection of H_o is for small values of R. The null distribution of the test statistic has been derived, based on combinatorial analysis [6].

Consider samples of size *m* and *n* respectively from distributions F_x and F_y, both defined in $\mathbf{R}^p$. Let N=m+n, C be the number of edge pairs of MST sharing a common node and d_i be the degree of the i^{th} node. Under H_o, the mean and variance of R can be calculated as follows:

$$E[R]=\frac{2mn}{N}+1 \text{ and}$$

$$Var[R\,|\,C]=\frac{2mn}{N(N-1)}\times$$

$$\times\left\{\frac{2mn-N}{N}+\frac{C-N+2}{(N-2)(N-3)}\left[N(N-1)-4mn+2\right]\right\}.$$

It has been shown that the quantity:

$$W=\frac{R-E[R]}{\sqrt{Var[R]}},$$

approaches (asymptotically) the standard normal distribution while E[R] and Var[R] are given in closed form based on the size of the two samples [6]. This enables the computation of the *significance level* (and *p-value)* for the acceptance of the hypothesis H_o.

In our case, the above test is utilized as follows. Point-samples are drawn from the two images (the color content of which is to be compared) by selecting some pixels from them. W is then computed and used as a similarity measure in a way that the more positive its value is, the more similar the two images are. Under this view, the procedure is directly incorporated in retrieval processes from a large image library and/or indexing. The only point that needs further consideration is the selection of pixels participating in the MST-graph representation of each image. Following the classical pattern-analytic convention, the pixel-selection can be thought as the feature extraction step.

3. FEATURE EXTRACTION

In order to perform the multivariate WW-test for image similarity, a set of m=n=N/2 points must be extracted from each database image as representative features. Many techniques can be adopted in this step, including local-based approaches such as salient-point based methods [7] and region-point based ones [8]. Since clustering and segmentation techniques are time consuming, difficult to implement and still an open issue in the field of image processing (especially when natural images are encountered), random sampling is used for mining color vectors. In this way, a simple and generic framework emerges.

3.1 Random sampling

With random sampling, the goal is to choose a representative set of cases from the full population under consideration. In statistical terms, *simple random sampling* is the basic sampling technique where we select a group of items (a sample) from a larger group (a population). Each individual is chosen entirely by chance and each member of the population has an equal chance of being included in the sample. Every possible sample of a given size has the same chance of selection; that is, each member of the population is equally likely to be chosen at any stage in the sampling process.

Fig. 1 illustrates the performance of WW-test for a pair of dissimilar and a pair of similar RGB-images (shown with the associated label attached). For visualization purposes, only the red and green components are shown. In both panels, 15 vectors were randomly selected from each image and labelled accordingly from {1-15} and {16-30}. In Fig. 1(a), 3 edges having differently labelled nodes as endpoints are found, splitting the overall MST into 4 subgraphs, thus *R=4*.

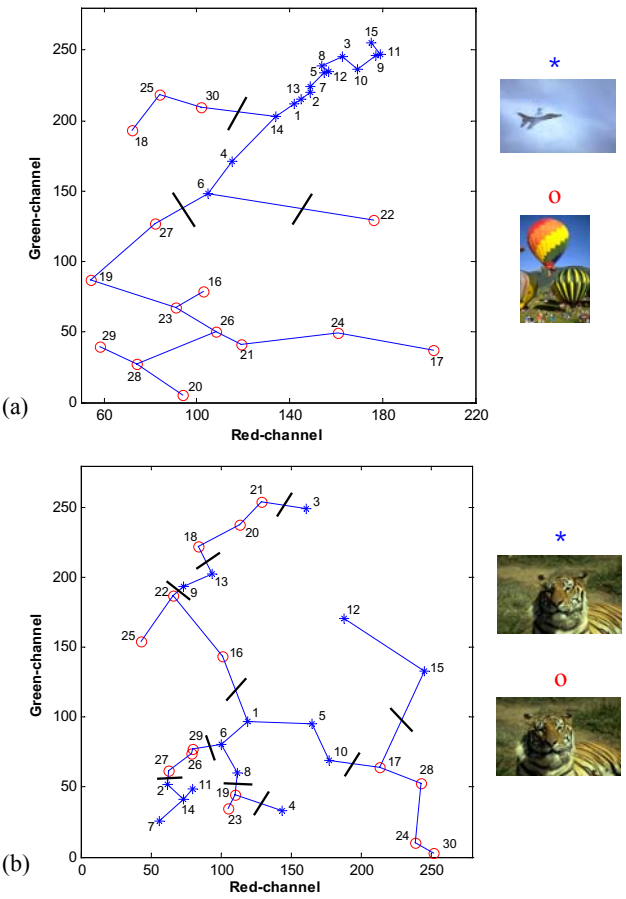

Figure 1: WW-test using 15+15 selected pixels ('o' and '∗' labels are indicating pixels from different images) from a pair of dissimilar images (a) and similar images (b).

In Fig. 1(b), 10 edges having differently labelled nodes as endpoints are apparent, i.e. *R=11*. In section 2, it was mentioned that the bigger the R the more similar two distributions (and thus images) are. It is therefore obvious that the samples selected from the two images in Fig. 1(a) are coming from two totally different color images, while the ones in Fig. 1(b) are coming from two similarly colored images. Finally, the similarity measure in the presented example is *W=-5.25* in the former case (Fig. 1(a)) and *W=-1.08* in the latter case (Fig. 1(b)).

3.2 Using local density to guide point selection

The main advantage of the proposed similarity measure is that it can be robust even for random samples of small size, as it will be indicated in Fig. 2. However, a major restriction might be the fact that in the process of random selection of pixels one cannot avoid those coming from the transition regions between two distinct color areas in the image plane. Since the number of randomly selected samples is sufficiently small and the pixels belonging to the edges or contour are not representative for the image color information, they must be excluded from the selection process.

To achieve this, density was estimated locally in the image, using a 3x3-sliding window, based on the technique of potential functions [9,10]. In this way the most homogeneous regions in the image can be easily identified [10]. The density-value or potential $p_L(\mathbf{X})$ produced by the L sample vectors X_i (included in the window L) at the position X (corresponding to the central pixel), was computed as follows:

$$p_L(X) = \frac{1}{Lh^3}\sum_{i=1}^{L} K\left(\frac{X - X_i}{h}\right),$$

where K is the multivariate Gaussian kernel and h is its bandwidth (set to a global value). Using the computed values of density (estimated for each pixel locally from its neighbourhood), the pixels were ordered globally based on the attached potentials and a proportion of them having a low-density value were excluded from the subsequent random sampling.

4. EXPERIMENTS

For all the experiments reported in this section, 200 typical images from the Macmillan Image Collection were used as the test data set. The images were selected so as to form 10 categories of 20 images each.

In order to estimate the number of random pixel that should be selected each time for our experiments, 20 extra images were used as queries (2 from each category). Using *W* as the similarity measure, it can be seen from Fig.2 that the system performs well enough even for a small number of random pixels. Moreover, there is a plateau in the included curve after the sample-size of 30 pixels, which justifies the computational efficiency of the proposed technique. Taking benefit from this experimental fact, in the rest of the experiments 30 pixels were used from each image.

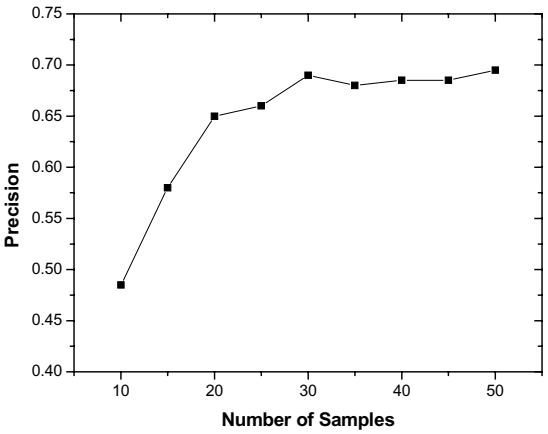

Figure 2: Retrieval precision vs. # of pixels from each image

The performances of WW-test and the global histogram intersection (HI) metric proposed by Swain and Ballard [4] for retrieval are compared. The estimated precision of retrieval for both measures is shown in Table 1. For better comparison of these two approaches, a plot of precision vs. recall Pr(Re) has also been included in Fig. 3.

Table 1: Comparison between Histogram Intersection (HI) and WW-test

Precision	Return Top 5	Return Top 8	Return Top 10	Return Top 15
WW-test	0.71	0.64	0.62	0.56
HI	0.62	0.57	0.49	0.46

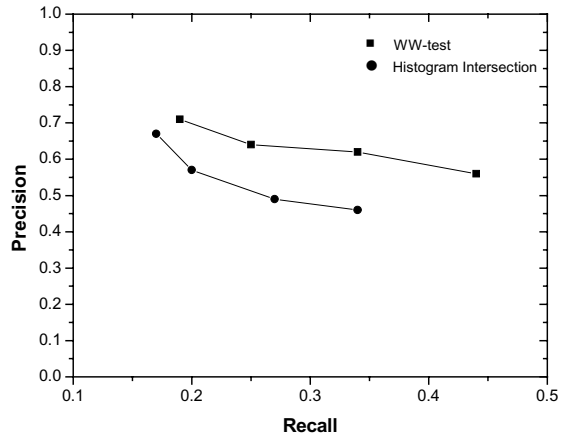

Figure 3: Retrieval performance Pr(Re) comparison

Next, the performance of the two different similarity measures when used for indexing in databases was compared. For visualization purposes, traditional systems display the retrieved images in a simple 1-D scheme sorted by decreasing similarity to the query. In this way, even though images are ranked proportional to the query, some relevant ones may appear at totally different places in the list. In order to get a global view reflecting the relations among the database images based on their mutual information similarities, Rubner [11] proposed a 2-D visualization scheme based on Multi-Dimensional Scaling (MDS) problem, discussed by

Kruskal [12]. In Fig. 4, images from 4 different categories are displayed, comparing histogram intersection to WW-test.

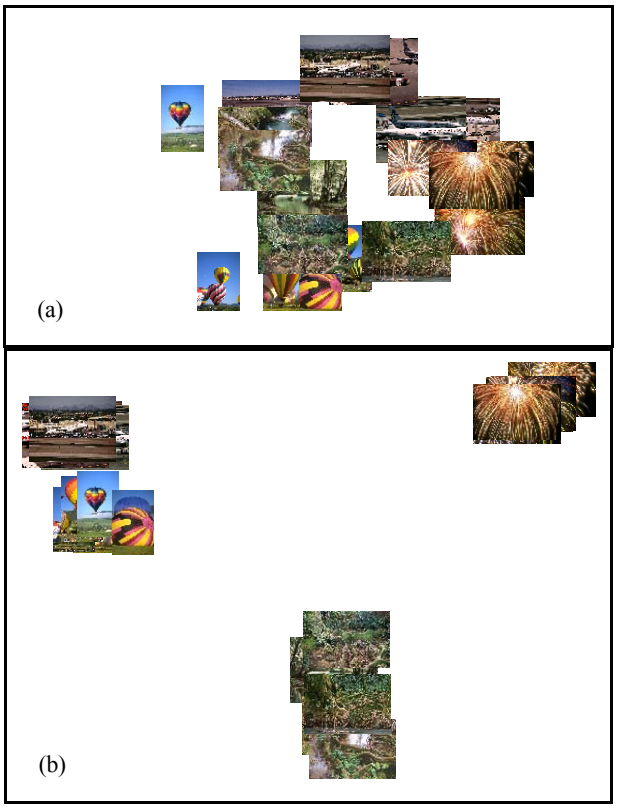

Figure 4: Multidimensional scaling results for 20 images using (a) Histogram Intersection and (b) WW-test

From the above figure it is obvious that, similarity based on WW-test forms separate clusters in the multidimensional scaling display characterized by high concentration of inter-cluster distance. On the other hand, color histogram is inappropriate for separating images coming from different categories.

5. CONCLUSIONS AND DISCUSSION

In this paper a similarity measure is presented for color image retrieval and indexing based on the ‘‘multivariate two-sample problem’’-related test. In order to estimate its statistical metrics, a small number of random samples is required from the two images to form two sets of points in the RGB color space, justifying the computational efficiency of the proposed technique. To avoid pixels from the transition region between two areas in the image plane of different color, we use the computation of its pixel density from its neighbourhood and exclude the low potential ones from the random sampling. Afterwards, statistical measures W and R are calculated as described in Section 2.

In this approach, W was used as the similarity measure for retrieval and visualization, given that the greater the W the more similar the two images we have sampled from. Since W is indeed a statistical metric, we might want to benefit from the fact that W can be transformed to a significance level, so as to filter out recalls corresponding to images that although are ranked (relatively) among the most similar, they are indeed different from the query image in term of color distribution.

It is among our future objectives to form sets of vectorial observations in higher space by incorporating the local context of a randomly sampled pixel, embedding it along with its 8-connected neighbours in a corresponding 3x9 RGB color-space. Moreover, pixels coming from points with greatest color interest might be inserted in the proposed test in order to improve the retrieval accuracy.

6. ACKNOWLEDGEMENTS

This work was supported by the European Social Fund (ESF), Operational Program for Educational and Vocational Training II (EPEAEK II), under the Program HERAKLEITOS.

REFERENCES

[1] Y. Rui, T. S. Huang and S. F. Chang, “Image retrieval: Current techniques, promising directions and open issues”, *Journal of Visual Communication and Image Representation*, vol. 10, pp. 39-62, Jan. 1999.

[2] B. S. Manjunath, P. Salembier and T. Sikora, *Introduction to MPEG-7: Multimedia Content Description Interface*, Wiley, 2002.

[3] V. Castelli and L. D. Bergman, *Image Databases: Search and Retrieval of Digital Imagery*, New York, Wiley, 2002.

[4] M. J. Swain and D. H. Ballard, “Color indexing”, *International Journal of Computer Vision*, vol. 7 (1), pp. 11-32, 1991.

[5] M. Stricker and M. Orengo, “Similarity of color images”, *Proc. SPIE Storage and Retrieval for Image and Video Databases*, San Jose, pp. 381-392, 1995.

[6] J. H. Friedman and L. C. Rafsky, “Multivariate Generalizations of the Wald-Wolfowitz and Smirnov Two-Sample Tests”, *Annals of Statistics*, vol. 7, pp. 697-717, Jul. 1979.

[7] V. Gouet and N. Boujemaa, “Object-based queries using color points of interest”, *Proc. IEEE Workshop on Content-Based Access of Images and Videos (CBAIVL)*, Hawaii, USA, Dec 2001.

[8] B. Moghaddam, H. Biermann and D. Margaritis, “Image retrieval with local and special queries”, *Proc. IEEE International Conference on Image Processing (ICIP)*, Vancouver, Canada, pp. 542-545, Sept. 2000.

[9] B. Smolka, K. N. Plataniotis, R. Lukac and A. V. Venetsanopoulos, “Kernel density estimation based multichannel impulsive noise reduction filter”, *Proc. IEEE International Conference on Image Processing (ICIP)*, Barcelona, Spain, 7-11 September 2003.

[10] G. Economou, A. Fotinos and S. Fotopoulos, “Color image edge detection based on nonparametric density estimation”, *Proc. IEEE International Conference on Image Processing (ICIP)*, Thessaloniki, Greece, vol. 1, pp. 922-925, 7-10 October 2001.

[11] Y. Rubner, *Perceptual metrics for image database navigation*, PhD dissertation, Stanford University, 1999.

[12] J. B. Kruskal, “Multi-dimensional scaling by optimizing goodness-of-fit to a nonmetric hypothesis”, *Psychometrica*, vol. 29, pp. 1-27, 1964.