AUTOMATIC IMAGE ANNOTATION AND RETRIEVAL USING THE JOINT COMPOSITE DESCRIPTOR.

Konstantinos Zagoris, Savvas A. Chatzichristofis, Nikos Papamarkos and Yiannis S. Boutalis
Department of Electrical & Computer Engineering
Democritus University of Thrace, Xanthi, Greece
{*kzagoris,schatzic,papamark,ybout*}*@ee.duth.gr*

***Abstract*—Capable tools are needed in order to successfully search and retrieve a suitable image from large image collections. Many content-based image retrieval systems employ low-level image features such as color, texture and shape in order to locate the image. Although the above approaches are successful, they lack the ability to include human perception in the query for retrieval because the query must be an image. In this paper a new image annotation technique and a keyword-based image retrieval system are presented, which map the low-level features of the Joint Composite Descriptor to the high-level features constituted by a set of keywords. One set consists of colors-keywords and the other set consists of words. Experiments were performed to demonstrate the effectiveness of the proposed technique.**

***Keywords*-Image Retrieval; Image Annotation; Joint Composite Descriptor; CCD; JCD**

I. INTRODUCTION

One major source of rapidly increasing user-created media is the image. This increase is sparked by the easiness of creating such images through the use of mobile phones, digital cameras and scanners. Thus, capable tools are needed in order to successfully search and retrieve the images. Content-based image retrieval techniques have been used such as img(Anaktisi) [1] and img(Rummager) [2] which employ low-level image features such as color, texture and shape in order to locate similar images. Although the above approaches are successful, they lack the ability to include human perception in the query of the retrieval because the query must be an image. In this paper we present a new image annotation technique and a keyword-based image retrieval system that attempts to solve this problem. The advantages of automatic image annotation versus content-based image retrieval are that queries can be more naturally specified by the user [3]. Many automatic image annotation systems have been developed since the early 1990s. In these systems, the images are represented by either global features, or block-based local and spatial properties, or region-based local features [4][5]. Global image low level features have been widely used in the literature [4].

The algorithm proposed in [6] classifies images by using the spatial correlation of colors. In [7] the color histograms are used in order to discriminate between indoor and outdoor images. In [8], support vector machines (SVMs) are employed to categorize images using the HSV color histograms. Bayesian classifiers on the color and edge direction histograms are used in order to classify sunset/forest/mountain images and city/landscape images [9]. Color features, shape features and wavelet-based texture features are used for automatic image annotation in [10] using SVMs and Bayes point machines (BPS). Comprehensive surveys of the automatic image annotation methods are available at [11] and [12]. On-line image annotation retrieval systems are available at [5] and [13].

The proposed method employs the Joint Composite Descriptor (JCD) [14] and utilizes two sets of keywords in order to map the low-level features of the descriptor to the high-level features constituted by these keywords. One set consists of colors-keywords and the other set consists of words. Experiments performed on the WANG [15] and the NISTER [16] databases demonstrate the effectiveness of the proposed technique.

The rest of the paper is organized as follows: Section 2 describes in details how the JCD is extracted. The proposed keyword based image retrieval system is presented in section 3. Section 4 illustrated the implementation method and the experimental results of the proposed technique. Finally, the conclusions are given in Section 5.

II. JOINT COMPOSITE DESCRIPTOR

The schemes which include more than one features in a compact vector can be regarded that they belong to the family of Compact Composite Descriptors (CCD) [17]. In [18] and [19] two descriptors are presented, those contain color and texture information at the same time in a very compact representation: the Color and Edge Directivity Descriptor (CEDD) [3] and the Fuzzy Color and Texture Histogram (FCTH) [19]. The structure of these descriptors consists of n texture areas. In particular, each texture area is separated into 24 sub regions, with each sub region describing a color. CEDD and FCTH use the same color information, as it results from 2 fuzzy systems that map the colors of the image in a 24-color custom palette. To extract texture information, CEDD uses a fuzzy version of the five digital filters proposed by the MPEG-7 EHD [20]

[21], forming 6 texture areas. In contrast, FCTH uses the high frequency bands of the Haar wavelet Transform in a fuzzy system, to form 8 texture areas. The types of texture areas adopted by each descriptor are illustrated in Figure 1. Observing the CCD results in various queries, it is easy to ascertain that in some of the queries, better retrieval results are achieved by using CEDD, while in others by using FCTH [14].

	0	1	2	3	4	5	6	7
CEDD	**Linear**	**Non Directional**	**Horizontal Activation**	**Vertical Activation**	**45 Degree Diagonal**	**135 Degree Diagonal**	**-**	**-**
FCTH	**Linear Low Energy**	**Horizontal Low Energy**	**Vertical Low Energy**	**Both Directions Low Energy**	**Linear High Energy**	**Horizontal High Energy**	**Vertical High Energy**	**Both Directions High Energy**

Figure 1. Compact Composite Descriptors Texture Areas.

Joint Composite Descriptor (JCD) combines CEDD and FCTH. This new descriptor is made up of 7 texture areas, with each area made up of 24 sub regions that correspond to color areas. The colors that represent these 24 sub regions are: (0) White, (1) Grey, (2) Black, (3) Light Red, (4) Red, (5) Dark Red, (6) Light Orange, (7) Orange, (8) Dark Orange, (9) Light Yellow, (10) Yellow, (11) Dark Yellow, (12) Light Green, (13) Green, (14) Dark Green, (15) Light Cyan, (16) Cyan, (17) Dark Cyan, (18) Light Blue, (19) Blue, (20) Dark Blue, (21) Light Magenta, (22) Magenta, (23) Dark Magenta. The texture areas are as follows: JCD(0) Linear Area, JCD(1) Horizontal Activation, JCD(2) 45 Degrees Activation, JCD(3) Vertical Activation, JCD(4) 135 Degrees Activation, JCD(5) Horizontal and Vertical Activation and JCD(6) Non directional Activation

In order to make the combination process of CEDD and FCTH clear, we model the problem as follows: Let CEDD and FCTH be available for one image (j). The indicator $m \in [0, 23]$ symbolises the bin of the color of each descriptor while $n \in [0, 5]$ and $n' \in [0, 7]$ determine the texture area for the CEDD and FCTH respectively. Each descriptor can be described in the following way: $\mathrm{CEDD}\,(\mathrm{j})_{\mathrm{n}}^{\mathrm{m}}, \mathrm{FCTH}\,(\mathrm{j})_{\mathrm{n}'}^{\mathrm{m}}$.

For example, the symbol $\mathrm{CEDD}(\mathrm{j})_2^5$ corresponds to the $\mathrm{bin}(2 \times 24 + 5 = 53)$ of the CEDD descriptor of image (j). The algorithm for the Joint Composite Descriptor can be analyzed as follows:

$$\mathrm{JCD}\,(\mathrm{j})_0^{\mathrm{i}} = \frac{\mathrm{FCTH}\,(\mathrm{j})_{\mathrm{o}}^{\mathrm{i}} + \mathrm{FCTH}\,(\mathrm{j})_4^{\mathrm{i}} + \mathrm{CEDD}\,(\mathrm{j})_{\mathrm{o}}^{\mathrm{i}}}{2} \quad (1)$$

$$\mathrm{JCD}\,(\mathrm{j})_1^{\mathrm{i}} = \frac{\mathrm{FCTH}\,(\mathrm{j})_1^{\mathrm{i}} + \mathrm{FCTH}\,(\mathrm{j})_5^{\mathrm{i}} + \mathrm{CEDD}\,(\mathrm{j})_2^{\mathrm{i}}}{2} \quad (2)$$

$$\mathrm{JCD}\,(\mathrm{j})_2^{\mathrm{i}} = \mathrm{CEDD}\,(\mathrm{j})_4^{\mathrm{i}} \quad (3)$$

$$\mathrm{JCD}\,(\mathrm{j})_3^{\mathrm{i}} = \frac{\mathrm{FCTH}\,(\mathrm{j})_2^{\mathrm{i}} + \mathrm{FCTH}\,(\mathrm{j})_6^{\mathrm{i}} + \mathrm{CEDD}\,(\mathrm{j})_3^{\mathrm{i}}}{2} \quad (4)$$

$$\mathrm{JCD}\,(\mathrm{j})_4^{\mathrm{i}} = \mathrm{CEDD}\,(\mathrm{j})_5^{\mathrm{i}} \quad (5)$$

$$\mathrm{JCD}\,(\mathrm{j})_5^{\mathrm{i}} = \mathrm{FCTH}\,(\mathrm{j})_3^{\mathrm{i}} + \mathrm{FCTH}\,(\mathrm{j})_7^{\mathrm{i}} \quad (6)$$

$$\mathrm{JCD}\,(\mathrm{j})_6^{\mathrm{i}} = \mathrm{CEDD}\,(\mathrm{j})_1^{\mathrm{i}} \quad (7)$$

with $\mathrm{i} \in [0, 23]$.

Table 1 presents the efficiency of the JCD in relation to that of the CEDD and FCTH is several known image databases. ANMRR[21] is employed to evaluate the performance of the descriptors.

	WANG [15]	UCID[22]	NISTER[16]
CEDD	0.25283	0.28234	0.11297
FCTH	0.27369	0.28737	0.09463
JCD	0.25606	0.26832	0.08548

Table I
ANMRR RESULTS IN THREE BENCHMARKING IMAGE DATABASES

The ANMRR is always in range of 0 to 1, and the smaller the value of this measure is, the better the matching quality of the query. ANMRR is the evaluation criterion used in all of the MPEG-7 color core experiments.

III. SYSTEM OVERVIEW

Figure 2 depicts the overall structure of the proposed system, which consists of two parts: the Offline and Online operations.

In the Offline operation the images are examined and the JCD is calculated.

The proposed image annotation technique and a keyword-based image retrieval system utilizes a two set of keywords in order to map the low features of the descriptor to the high level features that constitute by these keywords. Each set employs a different use of the JCD in order to annotate the image. The first set uses the inner workings of the descriptor in order to annotate images based on the quantity of each corresponding color in them. The second set uses SVMs to map the JCD image depiction with an equivalent word. The first set (Set A) is comprised of keywords that represent colors (Black, White, Yellow, Orange, Green, etc) while the second set (Set B) consists of simple words (Animal, Bird, Boat, Buildings, Car, etc).

The Color Similarity Grade (CSG) defines the amount of corresponding color in the image. It is calculated from the JCD for each color keyword based on the following equations:

$$\{black\} = \sum_{k=0}^{6} JCD\,(j)_k^2 \quad (8)$$

$$\{white\} = \sum_{k=0}^{6} JCD\,(j)_k^0 \quad (9)$$

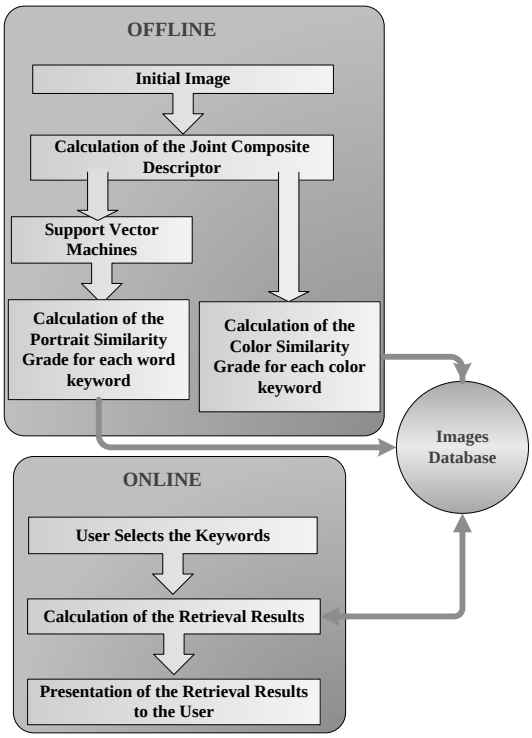

Figure 2. The overall structure of the proposed Keyword Annotation Image Retrieval System.

$$\{gray\} = \sum_{k=0}^{6} JCD\,(j)_k^1 \tag{10}$$

$$\{yellow\} = \sum_{k=0}^{6} JCD\,(j)_k^{10} \tag{11}$$

$$\{orange\} = \sum_{k=0}^{6} \left(\begin{array}{c} JCD\,(j)_k^6 + JCD\,(j)_k^7 + \\ +JCD\,(j)_k^8 + JCD\,(j)_k^{11} \end{array} \right) \tag{12}$$

$$\{green\} = \sum_{k=0}^{6} \left(\begin{array}{c} JCD\,(j)_k^9 + JCD\,(j)_k^{12} + \\ JCD\,(j)_k^{13} + JCD\,(j)_k^{14} + \\ +JCD\,(j)_k^{15} \end{array} \right) \tag{13}$$

$$\{cyan\} = \sum_{k=0}^{6} JCD\,(j)_k^{16} \tag{14}$$

$$\{blue\} = \sum_{k=0}^{6} \left(\begin{array}{c} JCD\,(j)_k^{17} + JCD\,(j)_k^{18} + \\ +JCD\,(j)_k^{19} + JCD\,(j)_k^{20} \end{array} \right) \tag{15}$$

$$\{magenta\} = \sum_{k=0}^{6} \left(\begin{array}{c} JCD\,(j)_k^{21} + JCD\,(j)_k^{22} + \\ +JCD\,(j)_k^{23} \end{array} \right) \tag{16}$$

$$\{red\} = \sum_{k=0}^{6} \left(JCD\,(j)_k^3 + JCD\,(j)_k^4 + JCD\,(j)_k^5 \right) \tag{17}$$

The Portrait Similarity Grade (PSG) defines the connection of the image depiction with the corresponding word. It is calculated from the normalization of a trained Support Vector Machines SVM Decision Function using as training samples the JCD values from a small subset of the available image database.

The Support Vector Machines (SVMs) are based on statistical learning theory and recently have been applied to many and various classification problems. The SVMs separate the space that the training samples are resided in two classes. The new sample is classified depending where in the space residues.

Let D is a given training dataset $\{(x_i, y_i)\}_{i=1}^{n}$, where x_i is the i input vector and y is the label correspond to the x_i.

If the training data are not linear separable (as in our case) then they mapped from the input space X to a feature space F using the kernel method where the training data become linearly separable. Our experiments showed the Radial Basis Function ($\exp\{-\gamma \|x - x'\|\}$) as the most robust kernel.

In practice sometimes the classifier must misclassify some data points (for instance to overcome the over fitting problem) using slack variables. The maximum margin classifier is calculated by solving the following constrained optimization problem which is expressed in terms of variables α_i:

$$\begin{aligned} \underset{a}{\text{maximize}} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j x_i^T x_j \\ \text{subject to :} \quad & \sum_{i=1}^{n} y_j \alpha_i = 0, \;\; 0 \leq \alpha_i \leq C \end{aligned}$$

The constant $C > 0$ defines the tradeoff between the training error and the margin. The training data x_i for which $a_i > 0$, are called support vectors.

One of the difficulties of the SVM consists of finding the correct parameters to train them. In our case, we have two parameters: the C from the maximum margin classifier and the γ from the Radial Basis Function kernel. The goal is to find the values of the two parameters C and γ so that the classifier can accurately predict the unknown data. This is achieved through a cross-validation procedure by using a grid search for the two parameters. The values of the above parameters for our keyword based image retrieval are calculated as: $C = 3, \gamma = 0.006$.

Finally the decision function which defines the classification of a new data sample x to one of the two classes are:

$$f\,(x) = sign\left(\sum_{i=1}^{n} a_i y_i \left(\varphi\,(x_i)\,, \varphi\,(x)\right) - b \right) \tag{18}$$

If $f\,(x) > 0$ then the sample x is classified to the class 1 otherwise to the class 0. In this work, we used the following equation to determine the membership value $R(x)$ of the sample x to the class 1:

$$R\,(x) = \left\{ \begin{array}{ll} 100 \times \max\left\{ \frac{1}{1+\frac{1}{3}e^{f(x)}}, \frac{1}{1+\frac{1}{3}e^{-f(x)}} \right\} & f\,(x) > 0 \\ 100 \times \left(1 - \max\left\{ \frac{1}{1+\frac{1}{3}e^{f(x)}}, \frac{1}{1+\frac{1}{3}e^{-f(x)}} \right\} \right) & f\,(x) < 0 \end{array} \right. \tag{19}$$

For each word a SVM is trained using as training samples a small sub set from the available image database. The input of a trained SVM for the word "Animal" is the image JCD and its output is the membership value which it is the PSG of the corresponding word for this image.

In the Online Operation the user can employ the two distinct keyword sets in order to retrieve an image. When the user selects multi keywords as a query, the images are displayed according to the Borda count rank. Initially, the Borda count was proposed as a single-winner election method whereby voters rank candidates in order of preference. The way in which the Borda count was applied in order to combine the results of the Portrait Similarity Grade and Color Similarity Grade is described as follows:

Let the keyword be A. The search is performed on a database according to the Color Similarity Grade. The results are sorted according to the distance D, which each image presents from the keyword A. Each image l, depending on the position shown in the results, is scored as follows:

$$Rank'(l) = \frac{N - RA}{N} \tag{20}$$

where N is the total number of the images in the database and RA is the Rank of the image l after the classification.

The same procedure is followed for a keyword B with the results that come from the Portrait Similarity Grade. The results are classified and each image is scored with $Rank(l)''$.

Finally, for each l image the $Rank(l) = Rank(l)' + Rank(l)''$ are calculated and a final classification of the results according to the $Rank$ of each image is done.

On the completion of the process, user can select one (or more) of the resultant images and the systems returns visually similar images from the database using once more the Joint Composite Descriptor.

IV. Implementation And Evaluation

The proposed work is implemented as part of the img(Anaktisi) web application at http://www.anaktisi.net and it is developed in C# .NET Framework 3.5 and it requires a fairly modern browser to use it. Figure 3 depicts the implemented interface of the proposed keyword based image retrieval technique.

The keyword based image retrieval is tested in the Wang database and the NISTER database. Figure 4 depicts the first 9 images from the retrieval results using the keyword "orange" (Fig. 4(a)) and the keywords "orange" and "black" (Fig 4(b)). It shows the impact of the keyword "black" to the initial retrieval results.

For the B' Set, we trained 13 SVMs (for each 13 word of B' Set) with 10 sample images for each one. Figure 5(a) depicts the retrieval results for the keyword "mountain" and Figure 5(b) shows the retrieval results using as query the JCD from the first image. This shows that the keyword

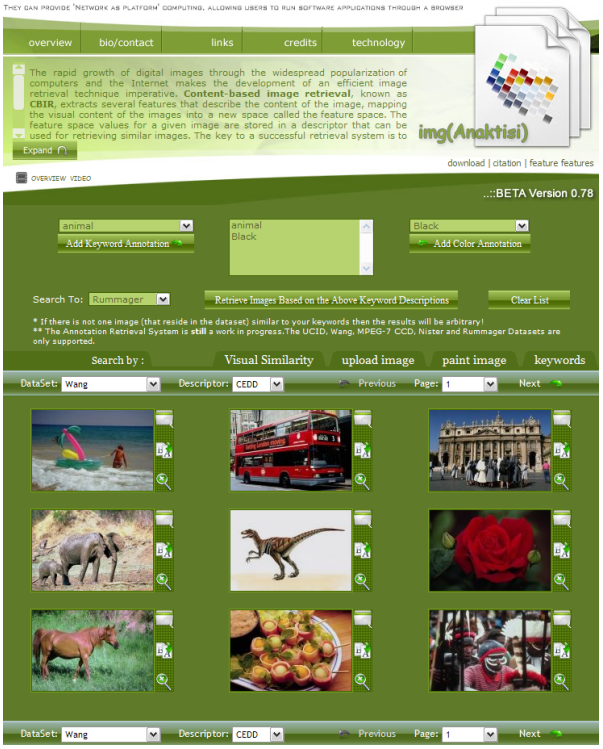

Figure 3. The implemented interface of the proposed keyword based image retrieval technique.

technique manage to retrieve more coherent results than using only the JCD feature.

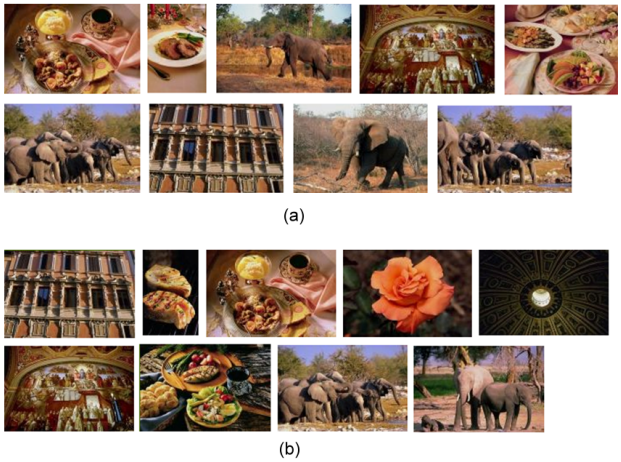

Figure 4. Image retrieval results using the keywords from the A' Set. (a) Image Retrieval Results using the keyword "orange", (b) Image retrieval results using the keywords "orange" and black.

V. Conclusions

An automatic image annotation method which uses the Joint Composite Descriptor is presented. This method provides two distinct sets of keywords: colors-keywords and

Figure 5. Image retrieval results using keyword from the B' Set and a query using only the JCD. (a) Image Retrieval Results using the keyword "mountain", (b) Image retrieval results using the JCD of the first image as query.

common words such as animal, bird, etc. The proposed project has been implemented and evaluated on the Wang database and the NISTER database. The results demonstrate the method's effectiveness as it retrieved results more consistent with human perception.

REFERENCES

[1] K. Zagoris, S. Chatzichristofis, N. Papamarkos, and Y. Boutalis, "Img (anaktisi): a web content based image retrieval system," in *SISAP*, 2009, pp. 154–155.

[2] S. A. Chatzichristofis, Y. S. Boutalis and Mathias Lux, "Img(rummager): An interactive content based image retrieval system," in *SISAP*, 2009, pp. 151–153.

[3] R. Datta, D. Joshi, J. Li, and J. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Computing Surveys*, vol. 40(2), pp. 1–60, 2008.

[4] X. Qi and Y. Han, "Incorporating multiple svms for automatic image annotation," *Pattern Recognition*, vol. 40, no. 2, pp. 728–741, 2007.

[5] J. Li and J. Wang, "Real-time computerized annotation of pictures," in *Proceedings of the 14th annual ACM international conference on Multimedia*. ACM, 2006, p. 920.

[6] J. Huang, S. Kumar, and R. Zabih, "An automatic hierarchical image classification scheme," in *Proceedings of the sixth ACM international conference on Multimedia*. ACM New York, NY, USA, 1998, pp. 219–228.

[7] M. Szummer and R. Picard, "Indoor-outdoor image classification," in *Proceedings of the 1998 International Workshop on Content-Based Access of Image and Video Databases (CAIVD'98)*, 1998, p. 42.

[8] O. Chapelle, P. Haffner, and V. Vapnik, "Svms for histogram-based image classification," *IEEE transactions on Neural Networks*, vol. 10, no. 5, p. 1055, 1999.

[9] A. Vailaya, M. Figueiredo, A. Jain, and H. Zhang, "Image classification for content-based indexing," *IEEE Transactions on Image Processing*, vol. 10, no. 1, pp. 117–130, 2001.

[10] E. Chang, K. Goh, G. Sychay, and G. Wu, "Cbsa: content-based soft annotation for multimodal image retrieval using bayes point machines," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 1, pp. 26–38, 2003.

[11] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 12, pp. 1349–1380, 2000.

[12] J. Smith and S. Chang, "Visualseek: a fully automated content-based image query system," in *Proceedings of the fourth ACM international conference on Multimedia*. ACM, 1997, p. 98.

[13] A. Yavlinsky, *Behold: a content based image search engine for the World Wide Web*. Citeseer, 2006.

[14] S. A. Chatzichristofis, Y. S. Boutalis, and M. Lux, "Selection of the proper compact composite descriptor for improving content based image retrieval," in *SPPRA*, 2009, pp. 134–140.

[15] J. Wang, J. Li, and G. Wiederhold, "Simplicity: Semantics-sensitive integrated matching for picture libraries," *IEEE Transactions on pattern analysis and machine intelligence*, pp. 947–963, 2001.

[16] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Proc. CVPR*, vol. 5. Citeseer, 2006, pp. 2161–2168.

[17] S. A. Chatzichristofis, K Zagoris, Y. S. Boutalis and N. Papamarkos, "Accurate image retrieval based on compact composite descriptors and relevance feedback information," *International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI)*, 2010 (To appear).

[18] S. Chatzichristofis and Y. Boutalis, "Cedd: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval," vol. 5008. Springer, 2008, p. 312.

[19] Chatzichristofis, S.A. and Boutalis, Y.S., "Fcth: Fuzzy color and texture histogram-a low level feature for accurate image retrieval," in *Proceedings of the 9th International Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS*, 2008, pp. 191–196.

[20] B. Manjunath, P. Salembier, and T. Sikora, *Introduction to MPEG-7: multimedia content description interface*. John Wiley & Sons Inc, 2002.

[21] B. Manjunath, J. Ohm, V. Vasudevan, A. Yamada *et al.*, "Color and texture descriptors," *IEEE Transactions on circuits and systems for video technology*, vol. 11, no. 6, pp. 703–715, 2001.

[22] G. Schaefer and M. Stich, "Ucid-an uncompressed colour image database," *Storage and Retrieval Methods and Applications for Multimedia*, vol. 5307, pp. 472–480, 2004.