14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

USER-CENTRIC EVOLUTION OF MULTIMEDIA CONTENT DESCRIPTION

Anastasios Doulamis and Nikolaos Doulamis
Department of Electrical & Computer Engineering
National Technical University of Athens
Zografou, Athens, Greece
Email: ndoulam@cs.ntua.gr

ABSTRACT

Humans usually interprets the same content is a different way mainly due to their subjective perception. For this reason, generic multimedia content description schemes, which remain constant regardless of the user's preferences, are not appropriate for a reliable and efficient content description. In this paper, a user-centric multimedia content description scheme is proposed. In particular, initially simple content multimedia algorithms are applied and in the following (and based on user's profile) particular descriptor algorithms are analyzed in more details in a progressive framework. This results in a multi-layer oriented scheme. To estimate, which descriptors are important for a particular user, a user profile estimator is applied by exploiting user's interaction. In particular, visual descriptor is organized into different categories, the energy of which determines the respective degree of importance.

1. INTRODUCTION

The explosion in multimedia information due to the development of low cost devices for capturing and encoding multimedia data and the use of media-rich applications has resulted in an appreciation for the value of multimedia content, and a realization of the challenges in managing that content [1, 2]. New multimedia applications such as content based retrieval, multimedia organization, browsing and navigation of audiovisual content, semantic content description, web ontologies and web mining require new tools and algorithms for efficient and reliable description of multimedia content.

Humans perceive and characterize content using high-level concepts, such as the amount of action, romance, comedy or emotional factors, which are not related in a straightforward way to visual attributes of pixels that compose images [3]. Furthermore, there is vagueness as far as multimedia content is concerned mainly due to the *subjective perception* of humans, which usually interpret the same semantic content in a different way, at different times. A piece of music can involve different feelings in different people; a picture is "worth a thousand words" [4]. Therefore, generic multimedia content description schemes, in the sense that they remain constant regardless of the user's preferences, are not appropriate for a reliable and efficient multimedia content management.

To address this situation, relevance feedback algorithms have been incorporated in a content-based retrieval scheme to update the system response to the current user's information needs. In the relevance feedback approach, the human is considered as a part of the retrieval process resulting in an interactive multimedia system.

Recently relevance feedback algorithms have been extended from text-based information retrieval [5] to Content-Based Image Retrieval (CBIR) systems. In particular, in [6], a probabilistic framework was reported using a Bayesian formulation scheme. In [7], a relevance feedback algorithm is introduced using as metric the weighted Euclidean distance and a heuristic scheme, adopted to perform the weight updating based on the variation of the elements of the feature vector of selected relevant data. This approach however, is an ad hoc method and as mentioned in the conclusions of this paper, there is a need for an optimal learning strategy. The first approaches towards this direction have been reported in [8, 9, 10]. In particular, in [8] and [9] a weighted Euclidean distance is used as in [7] and the weight updating strategy is performed by minimizing the Euclidean distance metric over all selected samples. In [10] instead, the generalized Euclidean distance is used as similarity measure to take into account the interconnection of different feature elements with each other. Feature element interconnection is also examined in [11]. However, the works of [8] and [9] yield unstable performance in case of negative examples, while "smoothing" the system's response when many positive relevant media data are selected. In addition, the work of [10] involves the inversion of the covariance matrix of the selected samples. It is clear, however, that the covariance matrix is not invertible if the number of selected samples is smaller than the size of image feature vectors, which is a common case in real situations. To confront this difficulty, the authors of [10] propose a solution based on the pseudo-inverse of the covariance matrix. Although in theory, such an approach eliminates the singularity problems, in practice, the retrieval performance is not so satisfactory [12]. To reduce the aforementioned difficulties, a "hierarchical model" has been proposed in [12] for decomposing the feature vectors into vectors of smaller size. Additionally, the algorithm introduces a dynamic switch for the weight updating to decrease the effect of singularity. However, this scheme presents difficulties in case of CBIR systems of large size feature vectors where the hierarchical model cannot be applied. A different approach is proposed in [13], where the relevance feedback problem is addressed using discriminant analysis. In particular, the algorithm proposes biased discriminant analysis to address the symmetry between positive and negative examples under small training samples, enhanced by a kernel version to facilitate non-linearities. A generalized relevance feedback scheme is presented in [14].

However, the aforementioned schemes update only the similarity measure used for ranking multimedia objects. Instead, the algorithms *applied* for multimedia content description remain constant. As, however, mentioned above there is a need for a user-centric multimedia description algorithm. This means that different descriptors (e.g. features) are extracted dynamically according to the user's preferences.

To implement such a personalized multimedia description algorithm, a user profile estimator is first required. The user's profile estimator is based on an on-line learning strategy, which exploits user's interaction. In particular, the system on line estimates the user's preferences and then activates new relevant to the user's wishes content description methods.

In the proposed architecture, the user-centric multimedia content description is implemented in a multi-layer oriented scheme. This

means that initially simple content multimedia algorithms are applied and in the following (based on user's profile) particular descriptor algorithms are analyzed in more details in a progressive framework. In this paper emphasis will be given for images and video content.

This paper is organized as follows: Section 2 presents the proposed user-centric multimedia description (in particular the user profile estimator, the way of system reconfiguration, the similarity metric for data ranking and the proposed on-line learning strategy). Section 3 presents the simulation results of the paper.

2. USER-CENTRIC MULTIMEDIA DESCRIPTION

In this section, we describe the user's profile estimator algorithm.

2.1 User Profile Estimation

Let us denote as $\mathbf{f}_i$ the feature vector of the i[th] object (e.g., an image) in the database and as $\mathbf{f}_q$ the feature vector of the query object (e.g., an image). Feature vectors $\mathbf{f}_i$ and $\mathbf{f}_q$ refers to a specific descriptor set as have been obtained by the proposed algorithm in the previous time instance (feedback iteration). Let us consider that the *current visual descriptors* are classified into *L* different categories, say $C_1, C_2, \cdots, C_L$.

Thus, C_i refers to a set of descriptors of specific type, such as color, motion, shape, and texture. For example, the color descriptors category (class) contains the color descriptors such as the color layout, dominant color, color space and so on.

Each category includes visual descriptors of common properties. For example, at the initial iteration, each category corresponds to color, shape, texture, motion visual descriptors. In the following iteration and under the assumption that the color descriptors are considered most important for the multimedia content description, the categories refers to different *color descriptors,* such as the dominant color, color layout, etc.

For each feature element (visual descriptor), we assign a degree of significance to the user's preferences, say λ_i. All λ_i which are associated to a particular category, say C_j, are included in a vector denoted as $\boldsymbol{\lambda}_{cj}$.

In order to perform the user-centric multimedia content description, we need to estimate the energy of the vector $\boldsymbol{\lambda}_{cj}$ in respect to the total energy of all categories

$$E_{\boldsymbol{\lambda}_{cj}} = \frac{\sum_{i \in Cj} \lambda_i^2}{\left\| C_j \right\|} \tag{1}$$

where $\|\cdot\|$ denote the norm of the set C_j, and λ_i all the parameters belonging to the category C_j. Then, the algorithm estimates all the categories $C_{\hat{j}}$ whose respective energy contributes more (defined by a *constant factor* α) to the total energy, i.e.,

$$\hat{j} : \sum_{j} E_{\boldsymbol{\lambda}_{c\hat{j}}} = \alpha \cdot E \tag{2}$$

Equation (2) means that, we select all these descriptor categories (e.g., color descriptors, shape descriptors) whose the energy (e.g., the importance) contribute a significant partition (defined by the factor a) to the total energy. Thus, the constant factor a indicates the portion of energy that the selected descriptor categories contribute to the total energy. In our algorithm, we select *a* to be equal to 80%. Index j refers to the indices of the selected descriptors categories.

Using equation (2), we estimate the categories that almost approximate the total descriptor energy, or in other words we estimate those categories which correspond to the most important descriptors with respect to the current user's information needs and preferences. This means that the descriptors of the selected categories are the most important to the user's profile and therefore, these descriptors should be analyzed in more details in the following retrieval iteration. On the contrary the descriptors corresponding to the categories of low energy (less significant categories) are ignored. In this way, a "multi scale" processing of visual data is accomplished oriented to the current preferences of users. Table I presents the main steps of the proposed algorithm for a relevance feedback iteration.

Table I: Description of the proposed scheme

1. Select a set of relevant/irrelevant images
2. Estimate the λ_i using equations (8) and (9)
3. Estimate the sum of all λ_i associated with features of a common group, i.e., color features
4. Estimate the energy $E_{\boldsymbol{\lambda}_{cj}}$ for a group, say C_j
5. Estimate the groups of the highest energy among all groups, using equation (2)
6. Extend the features of the group by considering additional features

In our implementation three category types are used; the color, motion, and texture category. In the initial visual processing only global-based descriptors are considered. In particular, for the color category, the color image histogram of the three RGB (Red, Green, and Blue) components is taken into account. For the motion category, the histogram of the motion vectors is exploited to indicate the motion activity of a video frame. Finally, as far as the texture category is concerned, the non-homogeneous texture descriptors are used (edge histogram) as in the MPEG-7 standard [16].

In the following retrieval steps and according to the user's selection, some (or all) descriptor categories are analyzed in more details. More specifically, the color information is enhanced using *dominant color descriptor* information, which expresses the principal color in a region (object-based descriptor). In the following retrieval iteration levels, different color spaces are used such as the HSV (Hue, Saturation and Value) format, the HMMD (Hue Min Max Difference) as well as the color distribution in an arbitrary shape region. Therefore, a *multi-layer mobile agent* reconfiguration is achieved.

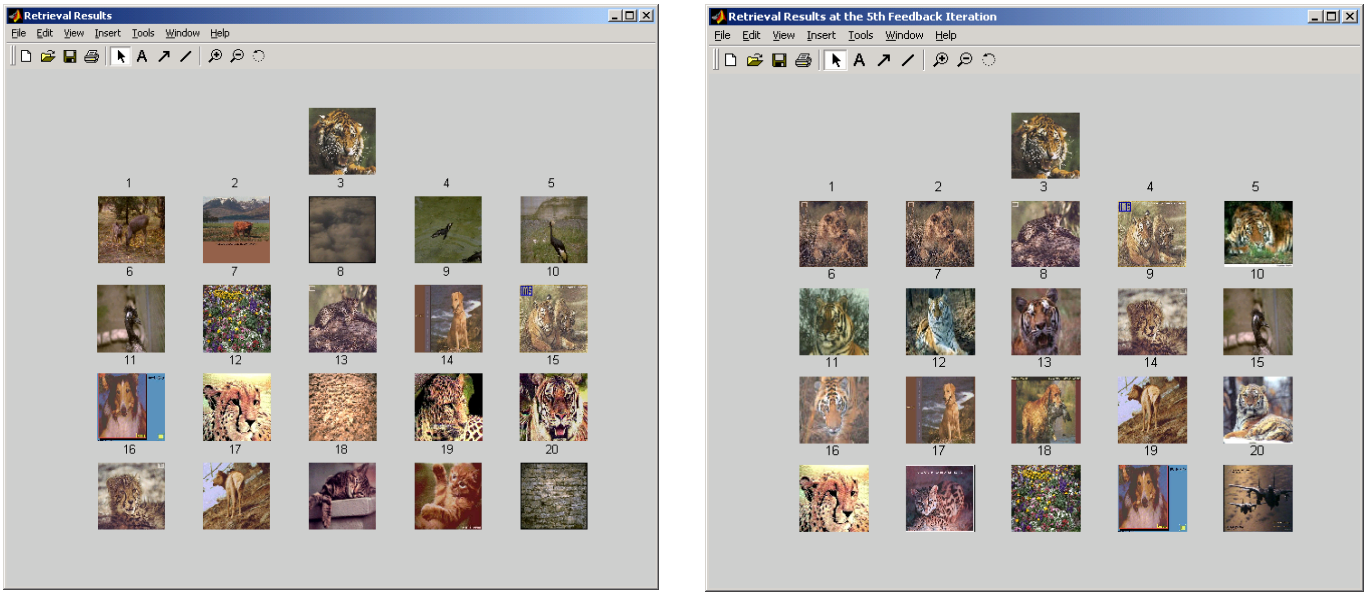
The motion category is enhanced, in the following retrieval iterations, by exploiting information of the camera motion and the motion distribution in arbitrary shape region as obtained using a motion segmentation algorithm. For the texture category, the homogeneous texture descriptors are used. The homogeneous descriptor aims at representing directionality, coarseness, and regularity of patterns in a video frame.

2.2 Reconfiguration

To implement such a dynamic scenario, reconfigurable architectures are required. In this paper, system reconfiguration is

performed using a mobile agent scheme. Mobile agents have the ability of migrating autonomously from node to node, while performing various tasks. The code on-demand approach supports dynamic configuration of agent properties. Upon instantiating a mobile agent, its code, an initial configuration and its programming interfaces are transferred to the target site.

More specifically, when a mobile agent returns the retrieval results, an on-line learning strategy is activated to evaluate the energy of the different descriptors categories, i.e., the importance of the weights λ_i. This is performed in an interactive framework, by allowing to the user to classify the results as relevant /irrelevant. Then, a new mobile agent is activated which, however, is bearing new codes corresponding to the new visual descriptors.



(a) (b)

Figure 1. Retrieval results (The query image is in the middle of the canvas). (a) Without reconfiguration. (b) With reconfiguration.

2.3 Multimedia Ranking

Having estimated the visual descriptors used for multimedia content description, a similarity measure is used for ranking multimedia data. In our implementation, ranking is performed based on Euclidean distance

$$d_g(\mathbf{f}_q,\mathbf{f}_i)=(\mathbf{f}_q-\mathbf{f}_i)^T\cdot(\mathbf{f}_q-\mathbf{f}_i) \tag{3}$$

where, however, $\mathbf{f}_q$ and $\mathbf{f}_i$ are the visual descriptors of the query and the database image as they have been selected by the aforementioned scheme.

2.4 Learning Strategy

In this section, we describe the algorithm used for estimating the weights (degree of significance) λ_i. More specifically, we parameterize the aforementioned similarity measure between two samples as follows

$$d_g(\mathbf{f}_q,\mathbf{f}_i)=(\mathbf{f}_q-\mathbf{f}_i)^T\cdot\mathbf{\Lambda}\cdot(\mathbf{f}_q-\mathbf{f}_i) \tag{4}$$

where $\mathbf{\Lambda}$ is a diagonal matrix containing the parameters of λ_i.

Then, using a set of relevant /irrelevant images, as selected by the user, the parameters λ_i of $\mathbf{\Lambda}$ are estimated by minimizing the similarity measure over a set of selected relevant /irrelevant samples.

$$J(\mathbf{\Lambda})=\sum_{i=1}^{m}d_{\mathbf{w}}(\mathbf{f}_q,\mathbf{y}_i)=\sum_{i=1}^{m}(\mathbf{f}_q-\mathbf{y}_i)^T\cdot\mathbf{\Lambda}\cdot(\mathbf{f}_q-\mathbf{y}_i) \tag{5}$$

Vectors $\mathbf{y}_i$ correspond to the feature vectors of the selected relevant (or irrelevant) samples.

However, it is easy to prove that minimization of (5) without imposing any constraint on parameters λ_i results in the trivial solution of all zeros, i.e., $\lambda_i=0,\forall i$. To overcome this problem, we constrain parameters λ_i as

$$\|\mathbf{\Lambda}\|_F=\sum_i\lambda_i^2=1 \tag{6}$$

where $\|\mathbf{\Lambda}\|_F$ denotes the Forbenius norm of matrix **Λ,** defined as the sum of the squares of the matrix elements. As a result, the optimal parameters $\hat{\lambda}_i$ or equivalently the optimal matrix $\hat{\mathbf{\Lambda}}$ is estimated by following constraint minimization problem

$$\hat{\mathbf{\Lambda}}=\arg\min_{\|\mathbf{\Lambda}\|_F=1}J(\mathbf{\Lambda})=\arg\min_{\|\mathbf{\Lambda}\|_F=1}\left\{\sum_{i=1}^{m}(\mathbf{f}_q-\mathbf{y}_i)^T\cdot\mathbf{\Lambda}\cdot(\mathbf{f}_q-\mathbf{y}_i)\right\} \tag{7}$$

Minimization of equation (7) results in

$$\hat{\lambda}_i=A_i\left(\sum_{k=1}^{N}A_k^2\right)^{-1/2},\quad i=1,\ldots,N \tag{8}$$

where

$$A_k=\sum_{i=1}^{m}(f_{q,k}-y_{i,k})^2 \tag{9}$$

$f_{q,k}$ is the kth element of feature vector $\mathbf{f}_q$, while $y_{i,k}$ is the kth element of the ith selected feature vector $\mathbf{y}_i$.

3. Simulation Results

In this section, we describe a typical scenario of the proposed user-centric multimedia description scenario. Initially, the user submits an image query. The image query is analyzed and the initial visual descriptors are extracted. Then, a mobile agent is activated to fetch images of similar content characteristics (as it is modeled by the initial extracted visual descriptors).

For the evaluation of the results a large image database has been used comprising of 15.000 images. The database is described in more details in [14].

The user evaluates the retrieved results and the on-line learning strategy of section 2 is applied to estimate the descriptor categories, which are closer to the user's information needs. Visual descriptors of the important categories are analyzed in more details in the following iteration.

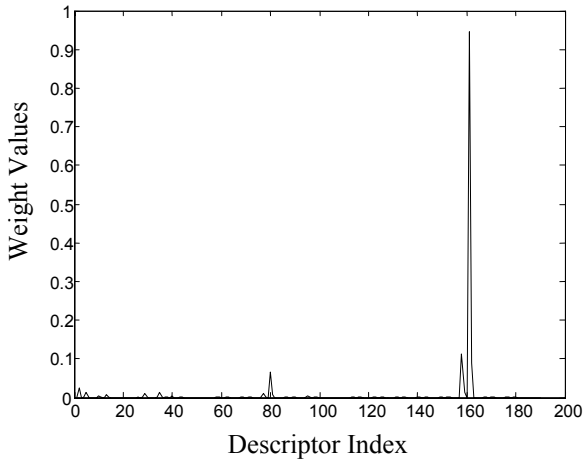


Figure 2. The weight values after a feedback iteration. As is observed only some weights are important (high weight values) in this particular example.

Figure 1(a) presents an example of a query image submitted to the system. The results of the initial retrieval iteration are presented in Figure 1(a). In the following the user-centric multimedia description is applied.

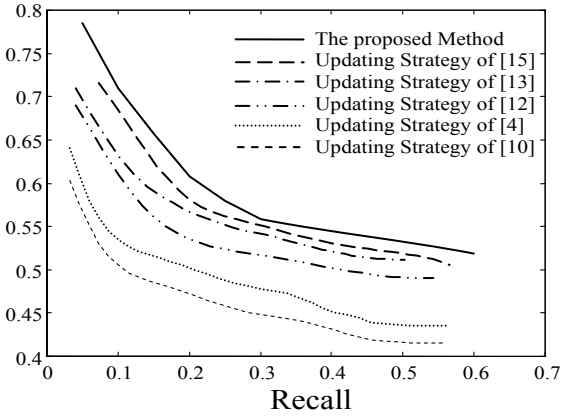


Figure 3. The precision recall curve of the proposed user-centric multimedia description along with other dynamic updating mechanism.

Figure 2 presents the weight values after the learning strategy. As is observed, some weights present greater values. In this particular example, almost all the energy (importance) descriptors result from the color descriptors. For this region additional color components are taken into consideration for the next retrieval iteration such as the dominant color and the color distribution in arbitrary shape regions. The final retrieval results (after the new multimedia description) are shown in Figure 1(b). As is observed, the final retrieval results are much closer to the actual user's information needs than the results obtained from the initial retrieval iteration.

Figure 3 presents the precision-recall curve of the proposed method along with other dynamic updating strategies. As is observed, the proposed reconfigurable scheme outperforms the examined ones (higher precision at a given recall). The outperformance is observed for all recall values. As a result, the presented user-centric multimedia content description schemes better fits the content of media data to the current user's information needs and preferences. This is due to the fact that in the proposed approach adaptation is achieved using a user-centric approach instead of just updating the descriptors weights (degree of importance) as the conventional approaches do.

Figure 4 presents the precision versus the number of feedback iterations, where similar conclusions are drawn. As can be seen, the precision accuracy increases with respect to the number of feedback iterations. However, the proposed user-centric multimedia content description outperforms the other methods.

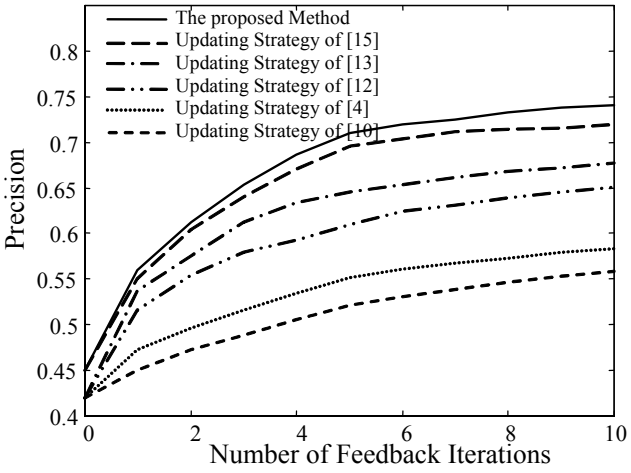


Figure 4. Precision versus the number of feedback iterations of the proposed scheme along with other dynamic updating mechanism.

Table II presents a comparison of the proposed method with several updating strategies. The evaluation has been performed using the Average Normalized Modified Retrieval Rank (ANMRR) as proposed by the MPEG-7 standard [18]. Low values of the ANMRR correspond to accurate retrieval with the retrieved data ranked in first places of the mining process. On the other hand, as the ANMRR values increase, the retrieval performance deteriorates along with the ranking of the data. As is observed form Table II, the proposed

methods presents the best results among the compared ones, since in the presented system the code used for data retrieval is modified instead of the similarity measure weights as happens in the other approaches.

Table II. The Performance 6f the Proposed Reconfigurable Architecture Compared to Several Relevance Feedback Schemes in Terms of the Average Normalized Modified Retrieval Rank (ANMRR) Criterion.

Different Algorithms	***ANMRR***
The proposed Algorithm	0.09
The Method of [15]	0.11
The Method of [13]	0.12
The Method of [12]	0.14
The Method of [4]	0.17
The Method of [10]	0.19

4. CONCLUSIONS

In this paper, we propose a user-centric multimedia content description scheme. The goal of this technique is to update the performance of a multimedia retrieval system to the actual current user's information needs and preferences. The proposed algorithm instead of applying traditional relevant feedback schemes, able to update the degree of importance of each descriptor to the similarity measure, extends at each iteration the set of descriptors applied using more detailed set of descriptors for the classes that are considered more relevant. Thus, in the presented approach, we modify the set of descriptors used to represent the visual content than by exploring more detailed description in domains, (e.g., color, shape, motion, etc) that are considered of high importance to the user's current actual information needs and preferences. Experimental results and comparisons with other approaches indicate the outperformance of the proposed scheme compared to other approaches used for relevance feedback in the literature.

5. REFERENCES

[1] Oren Etzioni, "The World-Wide Web: quagmire or gold mine?," *Communications of the ACM*, Vol. 39, No. 11, pp. 65-68, 1996.

[2] Special Issue on Content-Based Image Retrieval Systems, IEEE Computer Magazine, Vol. 28, No. 9, 1995. Guest Editors: Venkat N. Gudivada and Jijay V. Raghavan.

[3] N. Vasconcelos and A. Lippman, "Statistical Models of Video Structure for Content Analysis and Characterization ," *IEEE Trans. on IP,* Vol. 9, No. 1, pp. 3-19, January 2000.

[4] A. Doulamis, N. Doulamis and T. Varvarigou, "Efficient Content-based Image Retrieval using Fuzzy Organization and Optimal Relevance Feedback," *Internal Journal of Image and Graphics*, Special Issue on *Multimedia Data Storage and Management,* Vol.14, No. 1, pp. 150-166, January 2003.

[5] J. Rocchio, Relevance Feedback in Information Retrieval: The SMART Retrieval System: Experiments in Automatic Document Processing, Prentice Hall, 1971.

[6] I. Cox, M. L. Miller, S. M. Omohundro and P. N. Yianilos, "Pichunter: Bayesian Relevance Feedback for Image Retrieval," in Proc. *of Conf. Pattern Recognition*, Vol. 3, pp. 362-369, 1996.

[7] Y. Rui, T. S. Huang, M. Ortega and S. Mehrotra, "Relevance Feedback: A Power Tool for Interactive Content-Based Image Retrieval," *IEEE Trans. Circuits. Systems for Video Technology*, Vol. 8, No. 5, pp. 644-655, Sept. 1998.

[8] Y. Avrithis, A. D. Doulamis, N. D. Doulamis and S. D. Kollias, "An Adaptive Approach to Video Indexing and Retrieval," Proc. of *International Workshop on Very Low Bitrate Video Coding* (VLBV), pp. 69-72,Urbana IL, October 1998.

[9] A. D. Doulamis, Y. S. Avrithis, N. D. Doulamis and S. D. Kollias, "Interactive Content-Based Retrieval in Video Databases Using Fuzzy Classification and Relevance Feedback," Proc. of the *IEEE International Conference on Multimedia Computing and Systems*, Vol. 2, pp. 954-958, Florence, Italy, June 1999.

[10] Y. Ishikawa, R. Subramanya and C. Faloutsos, "Mindreader: Query Databases through Multiple Examples," in Proc. of the 24th *VLDB conference*, New York, USA, 1998.

[11] Y. Choi, D. Kim, and R. Krishnapuram, "Relevance Feedback for Content-based Image Retrieval using Choquet Integral," in Proc. *IEEE Inter. Conf. on Multi. & Expo*, pp. 1207-1210, New York, Aug. 2000.

[12] Y. Rui, and T.S. Huang, "Optimizing Learning in Image Retrieval," Proceeding of *IEEE int. Conf. on Computer Vision and Pattern Recognition,* Jun. 2000.

[13] Xiang Sean Zhou, T. S. Huang, "Small Sample Learning during Multimedia Retrieval using BiasMap", in Proc. *IEEE Conf. Computer Vision and Pattern Recognition*, Hawaii, Dec. 2001.

[14] A. Doulamis and N. Doulamis, "Generalized Non-Linear Relevance Feedback for Interactive Content-Based And Organization," *IEEE Trans. on Circuits and Systems for Video Technology*, special issue in audiovisual analysis for interactive multimedia services, Vol. 14, No. 5, pp. 656-671, May 2004.

[15] N. Doulamis and A. Doulamis "Fuzzy Histograms and Optimal Relevance Feedback for Interactive Content-based Image Retrieval," *IEEE Trans on Image Processing* (under second review for final acceptance.)

[16] A. Papoulis, Probability, Random Variables, and Stochastic Processes. New York: McGraw Hill, 1984.

[17] ISO/IEC JTC 1/SC 29/WG 11/N3964,N3966**,** "Multimedia Description Schemes (MDS) Group", March 2001, Singapore.

[18] "MPEG-7 Visual part of eXperimentation Model Version 2.0," MPEG-7 Output Document ISO/MPEG, Dec 1999.