Multimed Tools Appl (2006) 30: 255–272
DOI 10.1007/s11042-006-0028-0

Object-based MPEG-2 video indexing and retrieval in a collaborative environment

Vasileios Mezaris · Ioannis Kompatsiaris · Michael Gerassimos Strintzis

Published online: 27 July 2006

Abstract In this paper, an object-based video retrieval methodology for search in large, heterogeneous video collections is presented. The proposed approach employs a real-time, compressed-domain, unsupervised algorithm for the segmentation of image sequences to spatiotemporal objects. For the resulting objects, MPEG-7 compliant low-level descriptors describing their color, shape, position and motion characteristics are extracted. These are automatically associated using a fuzzy C-means algorithm with appropriate intermediate-level descriptors, which are part of a simple vocabulary termed *object ontology*. Combined with a relevance feedback mechanism, this scheme allows the qualitative definition of the high-level concepts the user queries for (*semantic objects*, each represented by a *keyword*) and relations between them, facilitating the retrieval of relevant video segments. Furthermore, it allows the collaborative construction of a knowledge base by accumulating the information contributed to the system during feedback by different users. Thus, it enables faster and more accurate retrieval of commonly requested keywords or semantic objects. Experimental results in the context of a collaborative environment demonstrate the efficiency of the proposed video indexing and retrieval scheme.

Keywords Compressed-domain video segmentation · Object-based video indexing · Ontologies · Relevance feedback · Support vector machines · Collaborative knowledge base construction

1 Introduction

The ever increasing generation of digital video content has recently established the need for the development of human-centered tools for the efficient access and retrieval from video

V. Mezaris · I. Kompatsiaris · M. G. Strintzis (✉)
Information Processing Laboratory, Electrical and Computer Engineering Department,
Aristotle University of Thessaloniki, Thessaloniki 54124, Greece
e-mail: strintzi@eng.auth.gr

V. Mezaris · I. Kompatsiaris · M. G. Strintzis
Informatics and Telematics Institute (ITI)/Centre for Research
and Technology Hellas (CERTH), Thessaloniki 57001, Greece

collections. Several such retrieval schemes have been proposed, employing descriptors which range from low-level features to higher-level semantic concepts [39]. Low-level features are machine-oriented and can be automatically extracted (e.g., MPEG-7 compliant descriptors [10]), whereas high-level concepts require manual annotation of the medium or are restricted to specific domains. In all cases, preprocessing of video data is necessary as the basis on which indices are extracted. The preprocessing is of *coarse granularity* if it involves processing of video frames as a whole, whereas it is of *fine granularity* if it involves detection of objects within a video frame [2] using segmentation tools [17, 32].

To allow efficient indexing of large video databases, an algorithm for the real-time, unsupervised spatiotemporal segmentation of video sequences in the MPEG-2 compressed domain is employed in this work [23, 25]. The choice of a compressed-domain segmentation algorithm is justified by the need for its real-time operation. This is particularly important in a retrieval application, due to the large volume of video that has to be processed. The employed algorithm performs both moving object extraction, by exploiting the motion information contained in the compressed stream, and background segmentation using DC color information. The latter is also useful in retrieval applications, where instead of querying for a compound background, this allows querying for the background's constituent objects, such as sky, sea, mountain, etc. The spatiotemporal segmentation algorithm is applied to shots; shot detection is performed using the method of [19], chosen for its computational simplicity.

Employing such a segmentation algorithm, this paper aims to bridge the gap between the low-level features extracted from the spatiotemporal objects and the high-level concepts used for querying. Solutions to this problem are usually restricted to domain-specific applications [36, 38], where exact mapping of low-level features to objects using *a priori* knowledge is feasible. In contrast to that, the proposed scheme attempts to address the problem of retrieval in generic video collections, where no possibility of employing *a priori* domain-specific knowledge exists. In such generic collections, the *query-by-example* paradigm [15, 30] is usually employed. This is based on the assumption that the user has access to a clip which represents what the user seeks, which is not very realistic [31] and for this reason, other query strategies have recently been proposed, such as the *query-by-sketch* paradigm presented in [9]. In [28, 31] the problem of bridging the gap between low level representation and high level semantics is formulated as a probabilistic pattern recognition problem. In [7, 11, 18] hybrid methods extending the query-by-example strategy are developed.

In the proposed indexing and retrieval scheme, instead of adopting the query-by-example strategy, the spatiotemporal segmentation algorithm is combined with a simple *ontology* [8] and a *relevance feedback* mechanism based on support vector machines [13, 37]. This scheme (figure 1) allows for MPEG-7 compliant low-level indexing features to be extracted for the spatiotemporal objects and subsequently be associated with higher-level descriptors that humans are more familiar with. A fuzzy C-means algorithm is used to create an automatically determined number of fuzzy partitions of the low-level feature space. The higher-level descriptors are used to restrict the search to a set of potentially relevant spatiotemporal objects, while query results are finally produced by ranking the potentially relevant spatiotemporal objects using a machine learning technique and additional information in the form of user feedback.

In a collaborative environment, the accumulation of the information contributed to the system in the form of feedback by different users is employed to create a knowledge base useful for improving the efficiency of the retrieval. User collaboration has been found beneficial for a number of applications, ranging from knowledge base authoring [16] to

medical image analysis and visualization [1]. In this work, user collaboration and knowledge base construction are integrated in a scheme for video retrieval.

The novel contributions of this work include the generation of the object ontology by creating automatically determined fuzzy partitions of the low-level feature space, using a fuzzy C-means algorithm. They also include a novel proposed scheme for creating a knowledge base, using the feedback supplied by different users. The latter is seen to provide highly efficient retrieval.

The paper is organized as follows: in Section 2 the employed compressed domain segmentation algorithm is briefly presented. In Section 3, the extraction of low- and intermediate-level descriptors for the spatiotemporal objects and the use of ontologies are discussed. The proposed scheme for video retrieval in a collaborative environment allowing the accumulation of knowledge is discussed in Section 4. Section 5 contains experimental evaluation of the developed methodology, and finally, conclusions are drawn in Section 6.

2 Spatiotemporal video segmentation

The extraction of spatiotemporal moving objects is the key-challenge of any video segmentation algorithm. The information used to this end by the employed segmentation algorithm is extracted from MPEG-2 [27] sequences during the decoding process. Specifically, motion vectors are extracted from the P-frames and are used for foreground/background segmentation and for the subsequent identification of different foreground objects. Color information is extracted for the purpose of background segmentation and is restricted to the DC coefficients of the macroblocks of I-frames. These coefficients correspond to the Y, Cb and Cr components of the MPEG color space.

The employed algorithm for moving object extraction is based on exploiting the extracted macroblock motion information and consists of three main steps:

- Step 1. Using the bilinear motion model, iterative macroblock rejection is performed frame-wise to detect macroblocks with motion vectors deviating from the single rigid plane assumption. As a result, certain macroblocks of the current frame are activated (marked as possibly belonging to the foreground).
- Step 2. The temporal consistency of the output of iterative rejection over the last few frames is examined, to detect activated macroblocks of the current frame that cannot be tracked back to activated macroblocks for a few previous frames. These are excluded from further processing (deactivated). This process is based on temporal tracking of activated macroblocks using their motion vectors.
- Step 3. Macroblocks still activated after step 2 are clustered to connected regions and these are in turn assigned to either preexisting or newly appearing spatiotemporal objects, based on the motion vectors of their constituent macroblocks. Spatial and temporal constraints are also applied to prevent the creation of spatiotemporal objects inconsistent with human expectation (e.g., single-macroblock objects or objects with unrealistically small temporal duration).

After the extraction of moving objects, background segmentation is performed on the basis of color homogeneity. K-means clustering of the DC color coefficients of the I-frames is performed, followed by temporal tracking of the resulting regions in P-frames using motion information. This results in the creation of a number of background spatiotemporal objects.

A more detailed description of the employed compressed-domain segmentation algorithm can be found in [23, 25].

3 Low-level and intermediate-level descriptors

3.1 Overview

The employed segmentation algorithm is suitable for introducing object-based functionalities to video indexing and retrieval applications, due to the formation of both foreground and background spatiotemporal objects, for which object-based descriptors in the context of the MPEG-7 Visual standard [34] can be extracted. Examples of such standardized descriptors include the *dominant color descriptor*, the *scalable color descriptor*, *contour-based* and *region-based* shape descriptors, *motion trajectory* and *parametric motion* descriptors. The use of such object-based descriptors permits the processing of more expressive queries and also more efficient indexing and retrieval, compared to key-frame based indexing.

With the exception of a few MPEG-7 descriptors, such as *Motion Activity*, which are fairly high-level, most standardized descriptors are low-level arithmetic ones, chosen so as to ensure their usefulness in a wide range of possible applications. These descriptors, however, are not suitable for direct manipulation by the user of an indexing and retrieval scheme, e.g., for defining the color of a desired object. When examining the specific application of object-based video indexing, it is possible to alleviate this problem by translating certain low-level arithmetic values to intermediate-level descriptors qualitatively describing the object attributes; the latter are preferable, since humans are more familiar with manipulating qualitative descriptors than arithmetic values.

Extending the approach in [24], the values of the intermediate-level descriptors used for this qualitative description form a simple vocabulary, the *object ontology*. Ontologies are tools for structuring knowledge, defined as the specification of a representational vocabulary for a shared domain of discourse which may include definitions of classes, relations, functions and other objects. In the proposed scheme, ontologies are used to facilitate the mapping of low-level descriptor values to higher-level semantics. More specifically, an object ontology is employed to enable the user to form a simple qualitative description of the desired objects and their relationships in the shot; in parallel, a qualitative description of each spatiotemporal object in the database is automatically estimated using the object ontology, as will be discussed in Section 3.3.

Under the proposed scheme, a query is initiated by the user qualitatively describing the semantic objects and their relations in the desired shot. By comparing the user-supplied qualitative description with the one automatically estimated for each spatiotemporal object, clearly irrelevant ones can be discarded; the remaining, potentially relevant ones are presented to the user at random order. The user then evaluates a subset of them, marking relevant ones simply by checking the appropriate “relevant” box. By submitting this relevance feedback, one or two support vector machines are trained and subsequently rank according to relevance all potentially relevant spatiotemporal objects, using their low-level descriptor values; the shots containing these objects are then presented to the user, ordered by rank. This relevance feedback process can then be repeated, to further enhance the output of the query. As discussed in Section 4.2, this approach is further extended so as to allow the collaborative construction of a knowledge base useful for improving the efficiency of the retrieval. The architecture of the indexing scheme and the resulting query procedure are graphically illustrated in figures 1 and 2.

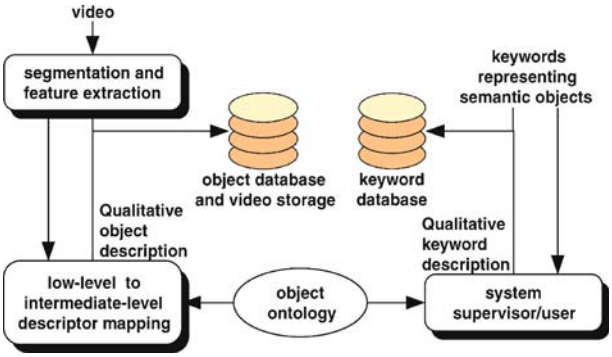

Fig. 1 Indexing system overview. Low-level and intermediate-level descriptor values for the spatiotemporal objects are stored in the object database; intermediate-level descriptor values for the user-defined keywords are stored in the keyword database

3.2 MPEG-7 descriptors

As soon as a sequence of segmentation masks is produced for each video shot, a set of descriptor values useful in querying the database are calculated for each spatiotemporal object. Standardized MPEG-7 descriptors are used, to allow for flexibility in exchanging indexing information with other MPEG-7 compliant applications. The different MPEG-7 descriptors used in this work are summarized in Table 1.

As can be seen from Table 1, each object property need not be associated with a single descriptor. For example, in the case of object motion, two different motion trajectories are calculated for each object, as the result of the use of two different coordinate systems (values "Local" and "Integrated" of *Spatial 2D Coordinates* descriptor). In the latter case, the use of a fixed, with respect to the camera, reference point for the coordinate system allows the categorization (e.g., fast or slow, direction) of foreground object motion even in the presence of a moving camera. Simultaneously, using "Local" coordinates facilitates the extraction of the qualitative space-localization of objects in the frame.

Two MPEG-7 "color" descriptors are also used; unlike motion descriptors, they both apply to all objects. This duality serves the purpose of satisfying the diverse requirements set by the general architecture of the retrieval scheme: low-level descriptors should be easy to map to intermediate-level qualitative descriptors (e.g., names of basic colors) and still permit accurate retrieval. A few most-dominant colors of the *Dominant Color* descriptor are

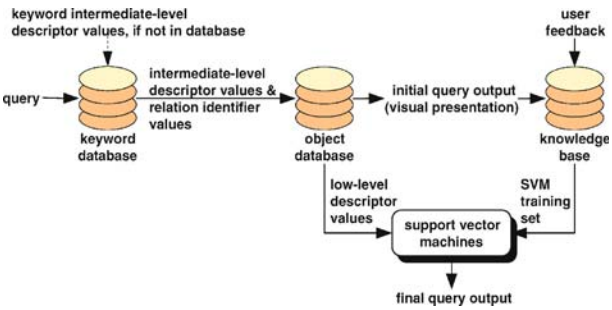

Fig. 2 Query process overview

Table 1 Set of used MPEG-7 descriptors

Descriptor	Video entity described
Motion Activity	Shot
Dominant Color	Spatiotemporal object
GoF/GoP Color	Spatiotemporal object
Contour Shape	Spatiotemporal object
Motion Trajectory using "Local" coordinates	Spatiotemporal object
Motion Trajectory using "Integrated" coordinates	Spatiotemporal object

most appropriate for associating with color-names, whereas when using the low-level descriptors directly, color histograms (*GoF/GoP Color*) demonstrate better retrieval performance [21]; they also have the advantage of being compatible with the L2 norm used as part of the employed relevance feedback mechanism.

3.3 Object ontology

In this work, ontologies [22, 33] are employed to allow the user to query a video collection using semantically meaningful concepts (semantic objects), without the need for performing manual annotation of visual information. A simple *object ontology* is used to enable the user to describe semantic objects, like "tiger", and relations between semantic objects, using a set of *intermediate-level descriptors* and *relation identifiers*. The simplicity of the employed object ontology permits its applicability to generic video collections without requiring the correspondence between spatiotemporal objects and relevant descriptors to be defined manually. This object ontology can be expanded so as to include additional descriptors corresponding either to low-level properties (e.g., texture) or to higher-level semantics which, in domain-specific applications, could be inferred either from the visual information itself or from associated information (e.g., subtitles).

Similarly to [6], an ontology is defined as follows.

Definition An *object ontology* is a structure (figure 3)

$$\mathcal{O}: = (\mathcal{D}, \leq_{\mathcal{D}}, \mathcal{R}, \sigma, \leq_{\mathcal{R}})$$

consisting of: (i) Two disjoint sets $\mathcal{D}$ and $\mathcal{R}$ whose elements d and r are called, respectively, intermediate level descriptors (e.g., "motion", "x speed," etc.) and relation identifiers (e.g., "spatial relation"). To simplify the terminology, relation identifiers will often be called *relations* in the sequel. The elements of set $\mathcal{D}$ are often called *concept identifiers* or *concepts* in the literature. (ii) A partial order $\leq_{\mathcal{D}}$ on $\mathcal{D}$, called concept hierarchy or taxonomy (e.g., "x speed" is a subconcept of "motion"). (iii) A function $\sigma: \mathcal{R} \rightarrow \mathcal{D}^{+}$ called *signature*; $\sigma(r) = (\sigma_{1,r}, \sigma_{2,r}, \sigma_{\Sigma,r})$, $\sigma_{i,r} \in \mathcal{D}$ and $|\sigma(r)| \equiv \Sigma$ denotes the number of elements of $\mathcal{D}$ on which $\sigma(r)$ depends. (iv) A partial order $\leq_{\mathcal{R}}$ on $\mathcal{R}$, called relation hierarchy, where $r_1 \leq_{\mathcal{R}} r_2$ implies $|\sigma(r_1)| = |\sigma(r_2)|$ and $\sigma_{i,r_1} \leq_{\mathcal{D}} \sigma_{i,r_2}$ for each $1 \leq i \leq |\sigma(r_1)|$.

For example, the signature of relation r "spatial relation" is by definition $\sigma(r) =$ ("position", "position"), indicating that it relates a position to a position; $|\sigma(r)| = 2$ denotes that r involves two elements of set $\mathcal{D}$. Both the intermediate-level "position" descriptor

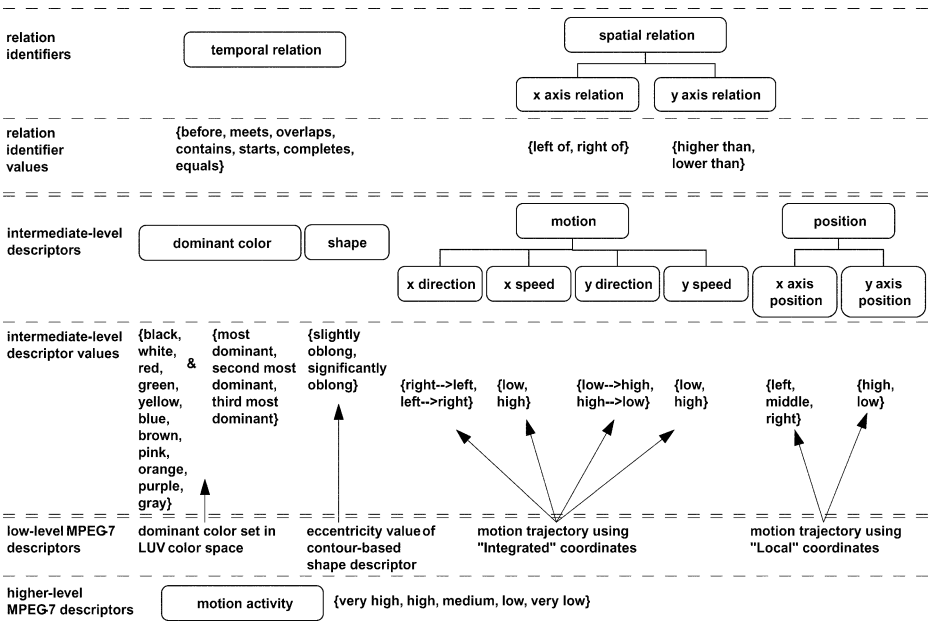

Fig. 3 Object ontology. The intermediate-level descriptors and the MPEG-7 Motion Activity descriptor are the elements of set $\mathcal{D}$, whereas the relation identifiers are the elements of set $\mathcal{R}$. The correspondence between low-level MPEG-7 descriptors and intermediate-level descriptors is shown

values and the underlying low-level descriptor values can be employed by the "spatial relation" relation.

The object ontology is presented in figure 3, where the possible intermediate level descriptors and descriptor values are shown. Each value of these intermediate-level descriptors is mapped to an appropriate range of values of the corresponding low-level, arithmetic descriptor. With the exception of color (e.g., "black") and direction (e.g., "low→high") descriptor values, the value ranges for every low-level descriptor are chosen by applying a fuzzy C-means algorithm (FCM) [4, 5] to the values of the corresponding low-level descriptor. This makes possible the efficient creation of fuzzy clusters represented by the intermediate-level descriptor values, while introducing a degree of fuzziness to the latter; for example, both "x axis position: left" and "x axis position: middle" values may be used to describe a single object. Furthermore, it enables the selection of the appropriate number of values for each intermediate-level descriptor.

More specifically, the employed FCM algorithm minimizes the functional

$$J = \sum_{k=1}^{K} \sum_{i=1}^{N} u_{i,k}^{m} e_{i,k}^{2} \tag{1}$$

where $u_{i,k}$ is the fuzzy membership of data point x_i to partition $\mathcal{P}$, m is a fuzzification parameter (typically $m = 2$), and $e_{i,k}$ is the Euclidean distance of x_i from the centroid $\bar{x}_k$. The latter is calculated as:

$$\bar{x}_k = \frac{\sum_{i=1}^{N} u_{i,k}^{m} x_i}{\sum_{i=1}^{N} u_{i,k}^{m}} \tag{2}$$

For functional J to be minimized, the fuzzy memberships are calculated as:

$$u_{i,k} = \frac{1}{\sum_{q=1}^{K} \left(\frac{e_{i,k}}{e_{i,q}}\right)^{\frac{2}{m-1}}} \tag{3}$$

The selection of the number of clusters K (and, thus, the number of values of the corresponding intermediate-level descriptor) is performed by applying FCM clustering for a number of possible values of K, typically, $K = 2, \ldots, 7$ (at least two values are necessary for each intermediate-level descriptor, while more than seven were found not to be useful to users). The value of K for which the normalized entropy $H_n(\mathcal{U}_K, K)$ is minimized is then chosen, where $\mathcal{U}_K = \{u_{i,k}\}$,

$$H_n(\mathcal{U}_K, K) = \frac{H(\mathcal{U}_K, K)}{1 - \frac{K}{N}}, \tag{4}$$

$$H(\mathcal{U}_K, K) = -\frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{N} u_{i,k} \log(u_{i,k}) \tag{5}$$

For example, for the object speed in the horizontal axis (x speed), the normalized entropy is minimized for $K = 2$ fuzzy clusters, while for the object position in the same axis (x axis position), it is minimized for $K = 3$ fuzzy clusters. The resulting fuzzy cluster memberships for the spatiotemporal objects in the collection and the association of value ranges of the low-level arithmetic descriptors with the intermediate-level descriptor values are illustrated in figure 4.

Regarding color, a correspondence between the 11 basic colors [3], used as color descriptor values, and the values of the HSV color space is heuristically defined. More accurate correspondences based on the psychovisual findings of e.g., [3] and others are possible, as in [20, 26]; however this investigation is beyond the scope of the present work. Regarding the direction of motion, the mapping between values for the descriptors "x direction", "y direction" and the MPEG-7 *Motion Trajectory* descriptor is based on the sign of the cumulative displacement of the foreground spatiotemporal objects.

Concerning the relation identifiers of the object ontology, these include the temporal relations defined in [12], along with simple spatial relationship descriptors defining the desired position of an object with respect to the position of another object.

4 Video retrieval in a collaborative environment

4.1 Query process

A query is formulated using the object ontology to provide a qualitative description of the sought object or objects (using the intermediate-level descriptors) and the relations between them. Existing descriptions of objects, i.e., descriptions which were previously imported to the system by the same or other users, can also be directly reused. As soon as a query is formulated, the intermediate-level descriptor values associated with each desired semantic object/keyword are compared to those of each spatiotemporal object contained in the

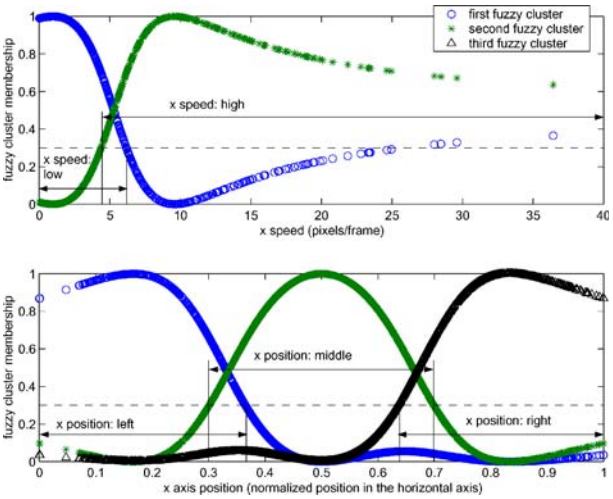

Fig. 4 Fuzzy cluster memberships for the spatiotemporal objects in the collection and association of low-level values with intermediate-level descriptor values, for the object speed and position in the horizontal axis ("x speed" and "x axis position")

database. Descriptors for which no values have been associated with the desired semantic object are ignored; for each remaining descriptor, spatiotemporal objects not sharing at least one descriptor value with those assigned to the desired semantic object are deemed irrelevant. In the case of dual-keyword queries, the above process is performed for each desired semantic object separately and only shots containing at least two distinct potentially relevant spatiotemporal objects, one for each keyword, are returned; if desired spatial or temporal relationships between the semantic objects have been defined, compliance with these constraints is checked using the corresponding low-level descriptors, in order to further reduce the number of potentially relevant shots returned to the user.

After narrowing down the search to a set of potentially relevant spatiotemporal objects, relevance feedback is employed to produce a qualitative evaluation of the degree of relevance of each spatiotemporal object. The employed mechanism is based on a method proposed in [13], where it is used for image retrieval using global image properties under the query-by-example scheme. This method combines support vector machines (SVM) [37] with a constrained similarity measure (CSM) [13]. Support vector machines employ the user-supplied feedback (training samples) to learn the boundary separating the two classes (positive and negative samples, respectively). Each sample (in our case, spatiotemporal object) is represented by its low-level descriptor vector $\mathbf{F}$. Given a set of training vectors belonging to two distinct classes, $(\mathbf{F}_1, y_1), \ldots, (\mathbf{F}_I, y_I)$, where $y_i \in \{-1, 1\}$ denotes class membership, the SVM attempts to find the optimal separating hyperplane. Using kernel mapping [37], non-linear discrimination of samples can be realized. In the latter case, the SVM classifier and its decision boundary are given by:

$$f(\mathbf{F}) = sign\left(\sum_{i=1}^{I} \overline{\alpha}_i y_i K(\mathbf{F}_i, \mathbf{F}) + \overline{b}\right), \tag{6}$$

$$\sum_{i=1}^{I} \overline{\alpha}_i y_i K(\mathbf{F}_i, \mathbf{F}) + \overline{b} = 0, \tag{7}$$

where $K(\mathbf{F}_i, \mathbf{F})$ is the employed kernel function, $\overline{\alpha}_i$ are the Lagrange multipliers and $\overline{b}$ is the solution to the optimization problem [13].

Following the decision boundary estimation, the CSM is employed to provide a ranking; in [13], the CSM employs the Euclidean distance from the key-image used for initiating the query for images inside the boundary (images classified as relevant) and the distance from the boundary for those classified as irrelevant. Under the proposed scheme, no key-image is used for query initiation; the CSM is therefore modified so as to assign to each spatiotemporal object classified as relevant the minimum of the Euclidean distances between it and all positive training samples (i.e., spatiotemporal objects marked as relevant by the user during relevance feedback).

In this work, the Gaussian radial basis function is used as a kernel function by the SVM, to allow for non-linear discrimination of the samples. The low-level descriptor vector **F** is composed of the 256 values of the histogram (*GoF/GoP Color* descriptor) along with the *eccentricity* value of the *Contour Shape* descriptor and either the position or the speed in the x- and y-axis, depending on whether the examined spatiotemporal object belongs to the background or the foreground, respectively. In the case of dual-keyword queries, two different SVMs are independently trained and the shot rank is calculated as the sum of the two ranks. This relevance feedback process can be repeated as many times as necessary.

4.2 Collaborative knowledge-base construction

The combination of ontology-based selection of potentially relevant spatiotemporal objects with relevance feedback overcomes the limitations of query-by-example approaches, with the construction of the training set used by the SVMs remaining a crucial factor for achieving efficient and accurate video retrieval. The proposed system attempts to address this by allowing users to annotate and index on-line video shots. Descriptions and feedback are then shared with other users, thus allowing the construction of a knowledge base. More specifically, in the case of queries for already described semantic objects/keywords, the feedback supplied by different users is combined to enable more efficient retrieval and diminish the need for each individual user to supply extensive feedback.

One possibility would be to use previous feedback to directly present to the user the video segments that where deemed relevant by all previous users. This, however, would be suitable only for a static (i.e., non-expanding) video collection and only under the assumption that the majority of the relevant video segments in the collection have already been identified. The latter is unknown to the users of any retrieval scheme. The assumption of a static collection is also very limiting. For this reason, a different approach is followed.

In the proposed approach, the feedback supplied by different users is used to collaboratively construct a knowledge base, associating with each keyword the spatiotemporal objects evaluated during the feedback stage of a query for this keyword. Let $\mathcal{B} = \{b_i\}$, $i = 1, ..., M$ denote this set of spatiotemporal objects for a given keyword, and let $pos(b_i)$ and $neg(b_i)$ denote the number of times b_i was deemed by the various users a positive sample and a negative one, respectively (clearly, $pos(b_i) + neg(b_i) > 0$, $\forall b_i \in \mathcal{B}$). Then, a training set for the SVM can be efficiently constructed by evaluating at any time, for each $b_i \in \mathcal{B}$, the ratio

$$R(b_i) = \frac{pos(b_i)}{pos(b_i) + neg(b_i)} \tag{8}$$

By including to the training set only samples $b_i \in \mathcal{B}$ for which $R(b_i) \geq pos_{th} > 0.5$ or $R(b_i) \leq neg_{th} < 0.5$, it is possible to exclude from the training set spatiotemporal objects, for the relevance of which there was significant ambiguity among the different users. In our experiments, pos_{th} and neg_{th} were set to 0.8 and 0.2, respectively.

5 Experimental results

The proposed methodology was tested on a collection of $1,213$ video shots corresponding to approximately 2 h of video. The collection comprised sports videos (e.g., Formula1, football, car racing), parts of nature-related documentaries, parts of digitized movies, and news videos belonging to the development data of the 2004 NIST TRECVID evaluation [35], that included various indoor and outdoor scenes.

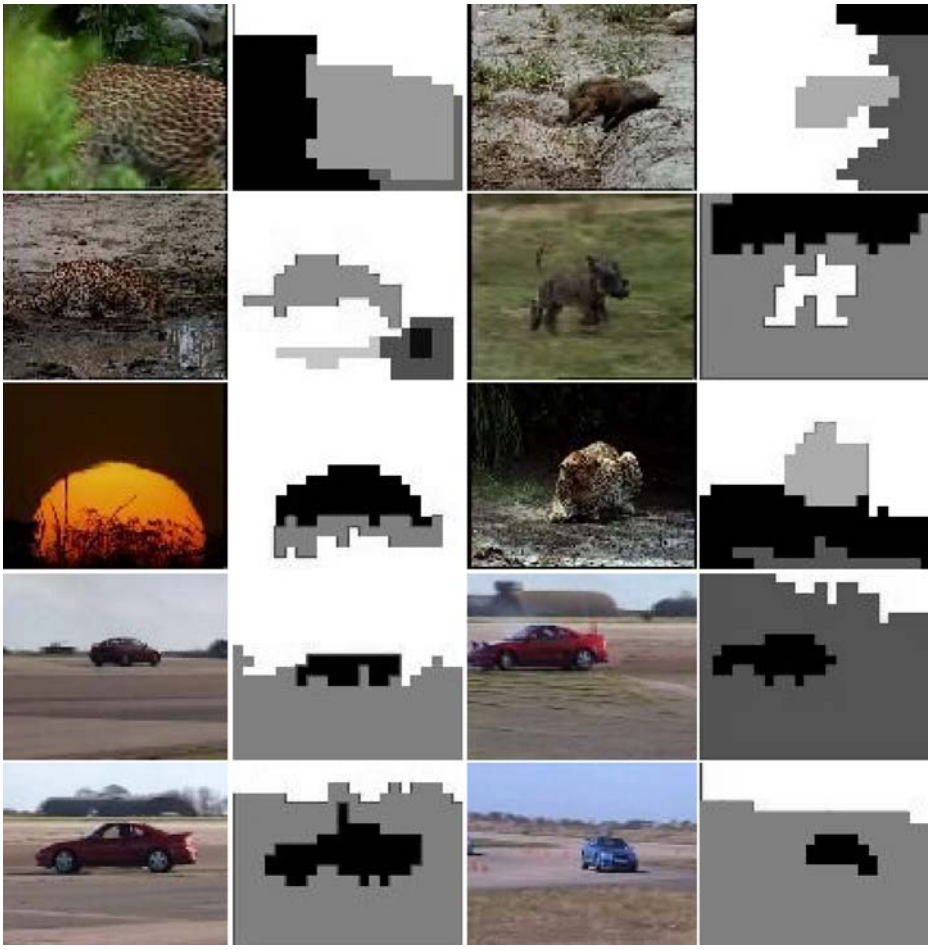

Fig. 5 Sample results of compressed-domain spatiotemporal segmentation of videos of the collection

The application of the segmentation algorithm of Section 2 to this collection of 1,213 video shots resulted in the creation of 6,584 spatiotemporal objects. Sample results of the application of the segmentation algorithm to videos of the collection are shown in figure 5. MPEG-7 low-level descriptors were calculated for each of the created objects, as described in Section 3.2. Following that, the association between these low-level descriptors and the intermediate-level descriptors defined by the object ontology was established; this was done, as discussed in Section 3.3, by applying the FCM algorithm to the collection of values of each low-level descriptor. This resulted to the estimation of the appropriate number of descriptor values for each intermediate level descriptor and the low-level-descriptor lower and upper boundaries corresponding to each such intermediate-level descriptor value. Appropriate textual labels such as "high", "medium", etc. were manually assigned to the latter. Since a large number of heterogeneous spatiotemporal objects was used for the initial boundary calculation, this process need not be repeated after future insertions of video clips to the database.

Subsequently, the object ontology was employed to qualitatively describe, using the available intermediate-level descriptors, high-level concepts, i.e., semantic objects. Since the purpose of the first phase of each query is to employ these definitions to reduce the data set by excluding obviously irrelevant spatiotemporal objects, the definitions of semantic objects need not be particularly restrictive; this is convenient from the users' point of view, since the user cannot be expected to have perfect knowledge of the color, shape and motion characteristics of the object sought in the database [29]. Four such definitions, namely for the "RedCar", "face", "Cheetah" and "BlueCar" keywords, are illustrated in figure 6. Subsequently, these descriptions were employed for querying. This resulted in initial query results being produced by excluding the majority of spatiotemporal objects in the database, which were found to be clearly irrelevant.

One or more pages of 15 randomly selected, potentially relevant spatiotemporal objects were then presented to the user for manual evaluation; the user checked the "relevant" check-box for those that were actually relevant. This information was used for the creation of a knowledge base, as discussed in Section 4.2, indicating for each keyword both positive and negative samples suitable for training the SVM-based mechanism. Consequently, when a query was executed for the first time, only the feedback provided at that time was employed for training the SVM, whereas in subsequent queries using the same keyword or keywords, the training set was determined using the knowledge base. The average time required for the SVM training and the subsequent object ranking was 0.18 s, on a Pentium IV PC. Results for

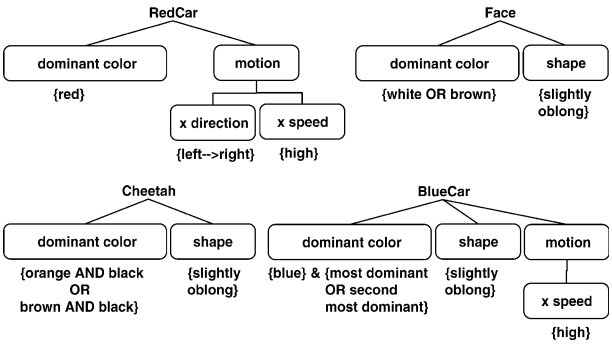

Fig. 6 Exemplary definitions of semantic objects using the object ontology, employed in retrieval experiments

two queries, both when submitted for the first time (i.e., when the knowledge base contained no relevant information) and after their tenth submission (i.e., employing the knowledge base constructed using the feedback of the nine previous users) are presented in figures 7 and 8. Detailed results using precision-recall diagrams, illustrating the average performance of the proposed method for four different queries, are presented in figure 9. This figure also

Fig. 7 Results for a "cheetah" query: (a) after one feedback (no previous knowledge employed) (b) using previous knowledge (query previously submitted by nine users)

reports the fraction of the database that is relevant to each of these queries, based on the *generality* measure of [14]. The diagrams of figure 9 demonstrate the usefulness of the intermediate-level descriptors, the efficiency of the proposed scheme in the absence of prior knowledge, and the significant improvement attained by additionally using a knowledge base collaboratively constructed by the users.

Fig. 8 Results for a "face" query: (a) after one feedback (no previous knowledge employed) (b) using previous knowledge (query previously submitted by nine users)

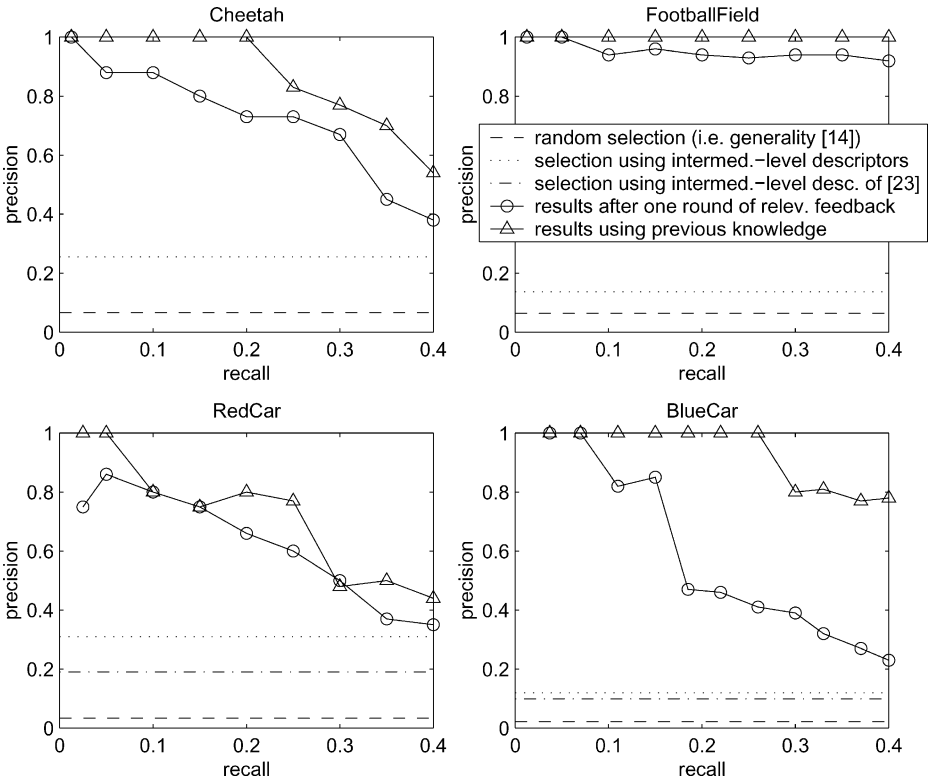

Fig. 9 Precision-recall diagrams for four queries. Results are shown both after one round of relevance feedback (i.e., without any previous knowledge) and using the previous knowledge supplied during query execution by nine users. Random selection results indicate the fraction of the database that is relevant to the given query, while random selection results using the intermediate level descriptors indicate the effectiveness of the latter in restricting the search to potentially relevant objects. The employed method (FCM) for associating certain intermediate-level and low-level descriptor values is shown to outperform the corresponding method of [23], when such intermediate-level descriptor values are employed (i.e., in the RedCar and BlueCar queries)

6 Conclusions

A methodology was presented in this paper for the flexible and efficient retrieval of video in a collaborative environment, combining a number of techniques. These include a real-time spatiotemporal segmentation algorithm, a simple object ontology, support vector machines and a collaboratively constructed knowledge base. The resulting methodology is applicable to generic, non-static video collections. It overcomes the restrictions of conventional methods, such as the need for the availability of a clip similar to what the user looks for. It is very suitable for web applications, which enable the required collaboration between a community of users. Experiments demonstrated the efficiency of the proposed approach in formulating descriptive queries and retrieving relevant visual information.

Acknowledgments This work was supported by the EU projects SCHEMA "Network of Excellence in Content-Based Semantic Scene Analysis and Information Retrieval" (IST-2001-32795), aceMedia "Integrating knowledge, semantics and content for user centred intelligent media services" (FP6-001765), and the

Hungary - Greece bilateral cooperation project CIRCE: Content-Based and Semantic Multimedia Indexing and Retrieval in Collaborative Environments. The assistance of COST211 quat and COST292 is also gratefully acknowledged.

References

1. Alberola C, Cardenes R, Martin M, Martin M, Rodriguez-Florido M, Ruiz-Alzola J (2000) diSNei: a collaborative environment for medical images analysis and visualization. In: Proc. Third Int. Conference on Medical Robotics, Imaging and Computer Assisted Surgery, Pittsburgh, Pennsylvania, pp 814–823
2. Al-Khatib W, Day Y, Ghafoor A, Berra P (1999) Semantic modeling and knowledge representation in multimedia databases. IEEE Trans Knowl Data Eng 11(1):64–80
3. Berlin B, Kay P (1969) Basic color terms: their universality and evolution. University of California, Berkeley
4. Bezdek J (1981) Pattern recognition with fuzzy objective function algorithms. Plenum, New York
5. Bezdek J, Keller J, Krishnapuram R, Pal N (1999) Fuzzy models and algorithms for pattern recognition and image processing. Kluwer, Norwell, Massachusetts
6. Bozsak E, Ehrig M, Handschuh S, Hotho A, Maedche A, Motik B, Oberle D, Schmitz C, Staab S, Stojanovic L, Stojanovic N, Studer R, Stumme G, Sure Y, Tane J, Volz R, Zacharias V (2002) KAON—towards a large scale Semantic Web. In: Proc. Third Int. Conf. on E-Commerce and Web Technologies (EC-Web 2002). Aix-en-Provence, France
7. Chan S, Qing L, Wu Y, Zhuang Y (2002) Accommodating hybrid retrieval in a comprehensive video database management system. IEEE Trans Multimedia 4(2):146–159
8. Chandrasekaran B, Josephso J, Benjamins V (1999) What are ontologies, and why do we need them? IEEE Intell Syst 14(1):20–26
9. Chang S-F, Chen W, Meng H, Sundaram H, Zhong D (1998) A fully automated content-based video search engine supporting spatiotemporal queries. IEEE Trans Circuits Syst Video Technol 8(5):602–615
10. Chang S-F, Sikora T, Puri A (2001) Overview of the MPEG-7 standard. IEEE Trans Circuits Syst Video Technol, special issue on MPEG-7 11(6):688–695
11. Chen W, Chang S-F (2001) VISMap: an interactive image/video retrieval system using visualization and concept maps. In: Proc. IEEE Int. Conf. on Image Processing, Vol. 3, pp 588–591
12. Day Y, Dagtas S, Iino M, Khokhar A, Ghafoor A (1995) Spatio-temporal modeling of video data for on-line object-oriented query processing. In: Proc. Int. Conf. on Multimedia Computing and Systems, pp 98–105
13. Guo G-D, Jain A, Ma W-Y, Zhang H-J (2002) Learning similarity measure for natural image retrieval with relevance feedback. IEEE Trans Neural Netw 13(4):811–820
14. Huijsmans D, Sebe N (2001) Extended performance graphs for cluster retrieval. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2001), Vol. 1. Kauai, Hawaii, pp 26–31
15. Izquierdo E, Casas J, Leonardi R, Migliorati P, O'Connor N, Kompatsiaris I, Strintzis MG (2003) Advanced content-based semantic scene analysis and information retrieval: the schema project. In: Proc. Workshop on Image Analysis for Multimedia Interactive Services, London, UK
16. Karp P, Chaudhri V, Paley S (1999) A collaborative environment for authoring large knowledge bases. Journal of Intelligent Information Systems 13(3):155–194
17. Kim C, Hwang J-N (2002) Fast and automatic video object segmentation and tracking for content-based applications. IEEE Trans Circuits Syst Video Technol 12(2):122–129
18. Kiranyaz S, Caglar K, Guldogan E, Guldogan O, Gabbouj M (2003) MUVIS: a content-based multimedia indexing and retrieval framework. In: Proc. Seventh Int. Symposium on Signal Processing and its Applications, ISSPA 2003, Paris, France, pp 1–8
19. Kobla V, Doermann D, Lin K (1996) Archiving, indexing, and retrieval of video in the compressed domain. In: Proc. SPIE Conf. on Multimedia Storage and Archiving Systems, vol. 2916, pp 78–89
20. Lammens J (1994) A computational model of color perception and color naming. Ph.D. thesis, University of Buffalo
21. Manjunath B, Ohm J-R, Vasudevan V, Yamada A (2001) Color and texture descriptors. IEEE Trans Circuits Syst Video Technol, special issue on MPEG-7 11(6):703–715
22. Martin P, Eklund P (2000) Knowledge retrieval and the World Wide Web. IEEE Intell Syst 15(3):18–25
23. Mezaris V, Kompatsiaris I, Kokkinou E, Strintzis MG (2003) Real-time compressed-domain spatiotemporal video segmentation. In: Proc. Third Int. Workshop on Content-Based Multimedia Indexing (CBMI03)

24. Mezaris V, Kompatsiaris I, Strintzis MG (2003) An Ontology Approach to Object-based Image Retrieval. In: Proc. IEEE Int. Conf. on Image Processing (ICIP03), Barcelona, Spain
25. Mezaris V, Kompatsiaris I, Boulgouris N, Strintzis MG (2004) Real-time compressed-domain spatiotemporal segmentation and ontologies for video indexing and retrieval. IEEE Trans Circuits Syst Video Technol 14(5):606–621
26. Mojsilovic A (2002) A method for color naming and description of color composition in images. In: Proc. IEEE Int. Conf. on Image Processing (ICIP02), New York, Rochester
27. MPEG-2 (1996) Generic coding of moving pictures and associated audio information. Technical report, ISO/IEC 13818
28. Naphade M, Huang T (2001) A probabilistic framework for semantic video indexing, filtering, and retrieval. IEEE Trans Multimedia 3(1):141–151
29. Naphade M, Huang T (2002) Extracting semantics from audio-visual content: the final frontier in multimedia retrieval. IEEE Trans Neural Netw 13(4):793–810
30. Naphade M, Yeung M, Yeo B (2000) A novel scheme for fast and efficient video sequence matching using compact signatures. In: Proc. SPIE Storage and Retrieval for Multimedia Databases, Vol. 3972, pp 564–572
31. Naphade MR, Kozintsev I, Huang T (2002) A factor graph framework for semantic video indexing. IEEE Trans Circuits Syst Video Technol 12(1):40–52
32. O'Connor N, Sav S, Adamek T, Mezaris V, Kompatsiaris I, Lui T, Izquierdo E, Bennstrom C, Casas J (2003) Region and object segmentation algorithms in the qimera segmentation platform. In: Proc. Third Int. Workshop on Content-Based Multimedia Indexing (CBMI03)
33. Schreiber A, Dubbeldam B, Wielemaker J, Wielinga B (2001) Ontology-based photo annotation. IEEE Intell Syst 16(3):66–74
34. Sikora T (2001) The MPEG-7 visual standard for content description—an overview. IEEE Trans Circuits Syst Video Technol, special issue on MPEG-7 11(6):696–702
35. TREC Video Track http://www-nlpir.nist.gov/projects/tv2004/
36. Tsechpenakis G, Akrivas G, Andreou G, Stamou G, Kollias S (2002) Knowledge-assisted video analysis and object detection. In: Proc. European Symp. on Intelligent Technologies, Hybrid Systems and their implementation on Smart Adaptive Systems (Eunite02). Algarve, Portugal
37. Vapnik V (1998) Statistical learning theory. Wiley, New York
38. Visser R, Sebe N, Lew M (2002) Detecting automobiles and people for semantic video retrieval. In: Proc. 16th Int. Conf. on Pattern Recognition, Vol. 2, pp 733–736
39. Yoshitaka A, Ichikawa T (1999) A survey on content-based retrieval for multimedia databases. IEEE Trans Knowl Data Eng 11(1):81–93

Vasileios Mezaris received the Diploma degree and the Ph.D. degree in Electrical and Computer Engineering from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2001 and 2005, respectively. He is a Postdoctoral Research Fellow with the Informatics and Telematics Institute / Centre for Research and Technology Hellas, Thessaloniki, Greece. Prior to this, he was a Postgraduate Research Fellow with the Informatics and Telematics Institute / Centre for Research and Technology Hellas, Thessaloniki, Greece, and a Teaching Assistant with the Electrical and Computer Engineering Department of the Aristotle University of Thessaloniki, Greece. His research interests include still image segmentation, video segmentation and object tracking, multimedia standards, knowledge-assisted multimedia analysis, knowledge extraction from multimedia, content-based and semantic indexing and retrieval. Dr. Mezaris is a member of the IEEE and the Technical Chamber of Greece.

Ioannis Kompatsiaris received the Diploma degree in electrical engineering and the Ph.D. degree in 3-D model based image sequence coding from Aristotle University of Thessaloniki (AUTH), Thessaloniki, Greece in 1996 and 2001, respectively. He is a Senior Researcher (Researcher D') with the Informatics and Telematics Institute, Thessaloniki. Prior to his current position, he was a Leading Researcher on 2-D and 3-D Imaging at AUTH. His research interests include 2-D and 3-D monoscopic and multiview image sequence analysis and coding, semantic annotation of multimedia content, multimedia information retrieval and knowledge discovery, MPEG-4 and MPEG-7 standards. His involvement with those research areas has led to the co-authoring of 2 book chapters, 14 papers in refereed journals and more than 50 papers in international conferences. He has served as a regular reviewer for a number of international journals and conferences. Since 1996, he has been involved in more than 15 projects in Greece, funded by the EC, and the Greek Ministry of Research and Technology. I. Kompatsiaris is an IEEE member, a member of the IEE Visual Information Engineering Technical Advisory Panel and a member of the Technical Chamber of Greece.

Michael Gerassimos Strintzis received the Diploma degree in electrical engineering from the National Technical University of Athens, Athens, Greece, in 1967, and the M.A. and Ph.D. degrees in electrical engineering from Princeton University, Princeton, NJ, in 1969 and 1970, respectively. He then joined the Electrical Engineering Department at the University of Pittsburgh, Pittsburgh, PA., where he served as Assistant Professor (1970-1976) and Associate Professor (1976-1980). Since 1980, he has been Professor of electrical and computer engineering at the University of Thessaloniki, Thessaloniki, Greece, and, since 1999, Director of the Informatics and Telematics Research Institute, Thessaloniki. His current research interests include 2-D and 3-D image coding, image processing, biomedical signal and image processing, and DVD and Internet data authentication and copy protection. Dr. Strintzis served as Associate Editor for the IEEE Transactions on Circuits and Systems for Video Technology from 1999 to 2004. In 1984, he was awarded one of the Centennial Medals of the IEEE. He is a Fellow of the IEEE.