Search and Retrieval of Rich Media Objects Supporting Multiple Multimodal Queries

Petros Daras, *Member, IEEE*, Stavroula Manolopoulou, and Apostolos Axenopoulos

***Abstract*—In this paper, a novel framework for rich-media object retrieval is described. The searchable items are media representations consisting of multiple modalities, such as 2-D images, 3-D objects and audio files, which share a common semantic concept. The proposed method utilizes the low-level descriptors of each separate modality to construct a new low-dimensional feature space, where all media objects can be mapped irrespective of their constituting modalities. While most of the existing state-of-the-art approaches support queries of one single modality at a time, the proposed one allows querying with multiple modalities simultaneously, through efficient multimodal query formulation, and retrieves multimodal results of any available type. Finally, a multimedia indexing scheme is adopted to tackle the problem of large scale media retrieval. The present framework proposes significant advances over existing methods and can be easily extended to involve as many heterogeneous modalities as possible. Experiments performed on two multimodal datasets demonstrate the effectiveness of the proposed method in multimodal search and retrieval.**

***Index Terms*—Manifold learning, multimedia description, multimedia indexing, multimodal search and retrieval.**

I. Introduction

THE amount of multimedia content, which is available in the Internet, is increasing at an incredible pace. This is not surprising, since media creation, even by nonprofessional users, has been enhanced through the widespread availability of digital recording devices, improved modeling tools, advanced scanning mechanisms as well as display and rendering devices. This increasing amount of multimedia data intensifies the need for effective search through the various online media databases.

Moving beyond traditional text-based retrieval approaches, a lot of research has been conducted on developing methods for content-based multimedia retrieval. The latter are based on the extraction of low-level features (e.g., color, texture, shape, etc.) automatically from content. While there are numerous content-based techniques that achieve retrieval of one single modality, such as 3-D objects [1]–[4], [39], images [5]–[7], video [8]–[10], or audio [11]–[13], only few are able to retrieve multiple modalities simultaneously.

Manuscript received February 15, 2011; revised July 29, 2011 and October 11, 2011; accepted December 08, 2011. Date of publication December 22, 2011; date of current version May 11, 2012. This work was supported by the EC-funded project I-SEARCH. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Shin'ichi Satoh.

The authors are with the Informatics and Telematics Institute, Centre for Research and Technology Hellas, Thessaloniki, GR-57001, Greece (e-mail: daras@iti.gr; manolop@iti.gr; axenop@iti.gr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2011.2181343

Cross-media retrieval, which has been introduced in the latest years, comprises all content-based multimedia search methods that take as input a query of one modality to retrieve results of another modality. That is, given as query the image of a dog (2-D), to be able to retrieve similar 3-D (dogs) objects. Moving beyond cross-media retrieval, multimodal retrieval allows users to enter multimodal queries and retrieve multiple types of media simultaneously. Both cross-modal and multimodal retrieval provide a significant step towards content-based multimedia retrieval, since users will be able to search and retrieve content of any type using a single unified retrieval framework and not a specialized system for each separate media type. Moreover, through multimodal retrieval, users will be able to enter multiple queries simultaneously, thus, retrieving more relevant results. However, this is a highly complicated process, since it requires successful modeling of the low-level feature associations among the different modalities.

A. Background and Related Work

In content-based multimedia retrieval, low-level descriptors, irrespective of the media type, are usually represented as high-dimensional vectors in the Euclidean space. Then, similarity matching between these vectors is performed by applying, usually, classical Euclidean metrics on their descriptor vectors. In most cases though, low-level descriptors follow a nonlinear manifold structure, which makes Euclidean metrics inappropriate. By properly unfolding this manifold structure, a more representative feature space of lower dimension is achieved. *Manifold learning* approaches have been already followed by many content-based search methods [14]–[16], which deal with one single modality, and significantly improve their retrieval performance. This concept can be easily extended to address cross-modal or multimodal retrieval problems, where descriptors from different modalities are mapped to the same low-dimensional manifold and reduce the multimodal similarity matching problem to a simple distance metric on this manifold. The most representative attempts in this field are given in the sequel. It should be noted that the following methods deal with cross-media and not multimodal retrieval, which is the main contribution of the present work.

Yang *et al.* [17] proposed a cross-media retrieval method, which connects various semantically-similar media types, aiming to retrieve results of different modalities compared to a given query. They introduced a structure called multimedia document (MMD) to define a set of multimedia objects (images, audio, and text) that carry the same semantics. The paper presented the concept of MMD distance, which intuitively merges the dissimilarities between different modalities as a

weighted sum of their mono-modal distances, based on each individual precision-recall performance. Using this MMD distance measure, all-to-all distances were computed, which were then used as input to a multidimensional scaling method to create a multimedia correlation space (MMCS), where every MMD is represented as a data point. In this space, a ranking algorithm was applied, called ranking with local regression and global alignment (LRGA), which learns a Laplacian matrix from data ranking. This algorithm uses a local linear regression model for each data point and then it globally aligns all of them through a unified objective function.

Although the above approach manages to associate different media types in the retrieval procedure, it has several weaknesses. Firstly, the MMD distance measure effectively merges multimedia information; however, it might significantly minimize the contribution of a specific modality, if its average retrieval accuracy (usually measured by precision/recall) is low. Furthermore, the computation of MMD distances implies the creation of a pre-calculated all-to-all distance matrix. This is not an efficient procedure when dealing with very large multimedia databases, since the size of the distance matrix becomes prohibitive as the database size increases. Another important issue is that the retrieval procedure proposed in [17] supports only mono-modal queries (cross-media retrieval). When a query object does not belong to the database, the closest database objects of the same modality are initially retrieved and their host MMDs are regarded as queries henceforth.

Similarly in [18], Zhang *et al.* investigated the intra- and inter-media correlations to build a map from heterogeneous multi-modal feature spaces, called "multimedia bags", into a semantic subspace created using Laplacian eigenmaps, called multi-modality Laplacian eigenmaps semantic subspace (MLESS). The different modalities supported are text, image, and audio. In the retrieval phase, queries can be either multimedia bags, if they belong to the database, or mono-modal media instances, otherwise. In the second case, the mono-modal neighbors of the query are found and the query is mapped into the center of their neighborhood in MLESS.

Furthermore, Wu *et al.* [19] proposed another cross-media retrieval method, which constructs an isomorphic subspace based on canonical correlation analysis (CCA) and thus called CCA subspace, to learn multi-modal correlations of media objects. A general distance function is defined in the CCA subspace using polar coordinates. When the query belongs to the database, the k-nearest neighbors of every modality are found and all of them are presented as results. These results can be further improved through one or more relevance feedback iterations. When a query does not belong to the database, k-nearest neighbors of the same modality are retrieved and their average coordinates in CCA subspace form a new query in CCA. The latter is used as input to retrieve cross-media results. Here, cross-media results depend highly on mono-modal neighbors, as well as on the user's judgement to mark the relevant ones.

A method for cross-modal association called cross-modal factor analysis (CFA) was introduced in [20]. The method achieves significant dimensionality reduction, while it effectively identifies the correlations between two different modalities. The method is tested in cross-media retrieval and demonstrates superior performance than similar approaches, such as canonical correlation analysis [22] and latent semantic indexing [21]. In [23], authors introduced a cross-media retrieval method based on mining the co-existence information of the heterogeneous media objects and users' relevance feedbacks, while in [24], they extended their work by proposing a structure for cross-media indexing over large multi-modal media databases. In [25], an approach for cross-media information aggregation was presented, which adopts online newspaper articles and TV newscasts as information sources to deliver a service made up of items including both contributions. In order to achieve information aggregation, the method is based on the concept of semantic relevance and on a novel asymmetric aggregation function. Finally, in [26], authors used kernel canonical correlation to build a kernel space where global inter-media correlation is analyzed. Correlations among text, image, and audio are analyzed to understand their underlying semantics. The method achieves significant retrieval accuracy for queries that do not belong to the database; however, it supports queries of only a single modality at a time.

Apart from cross-media retrieval, manifold ranking has been also used to improve the retrieval performance of methods that deal with mono-modal data [27]. In [14], Ohbuchi *et al.* proposed a framework for similarity comparison of shape features extracted from 3-D models. The overall scheme is divided into two phases: the learning phase and the retrieval phase. During the first one, shape features are extracted from the 3-D models and after subsampling them, unsupervised learning results in a lower dimensional manifold. Then, an approximation of the manifold is created using a neural network and features of all 3-D models are mapped to the lower dimensional space. During the retrieval phase, the query's features are projected onto the approximated manifold and distances with all 3-D models are computed so as to identify the closest matches.

Another dimensionality reduction algorithm similar to LE is the locally linear embedding (LLE) [15]. It performs unsupervised learning as well, by computing low-dimensional neighborhood preserving embeddings of high-dimensional data. In LLE, the data points in the high-dimensional space are represented as a linear combination of their nearest neighbors, thereby assuming that the manifold is locally linear. In the low-dimensional space it attempts to retain the weights of the linear combinations.

He *et al.* [16] introduced a manifold-ranking-based image retrieval (MRBIR) scheme, which measures the relevance between the query and the database images by exploring the relationship of all the data points in the feature space. Firstly, MRBIR forms a weighted graph regarding the data points as nodes. A positive ranking score is assigned to each query, while zero to the remaining points, and then all points spread their scores to the nearby points based on the weights of the graph. This step is repeated until a global stable state is reached. Finally, all points except for the query have their own scores according to which they are ranked. A modified version of the previous approach, the "modified manifold ranking (MMR)" algorithm [4], was proposed for the improvement of the 3-D shape retrieval performance. The significant points of this modified algorithm is that: 1) it creates a graph by connecting edges be-

tween the models and their nearest neighbors, 2) it assigns a weight to every edge based on that rank, and 3) it labels not only the query but the nearest neighbor as well. Inspired by the previous manifold ranking application on mono-modal data, the authors of [28] tried to extend this graph-based semi-supervised learning to multi-modal data. In order to achieve multi-modality, they created an independent graph for each kind of feature from one modality and the learning task was formulated as inferring from the constrains in every graph as well as supervision information, if available.

In this paper, a novel framework for multimodal retrieval is proposed. The framework enables search and retrieval of several media types, namely 3-D objects, images, and sounds, using as query any of the above types or combinations of them. This is achieved by mapping the low-level descriptors of the different modalities into the same low-dimensional feature space. By moving to this new feature space, multimedia data are not treated as separate media items but as rich media representations. The method is novel in the sense that queries may consist of multiple modalities simultaneously and retrieve results of multiple modalities as well. Another innovative feature of the proposed method is that it can be applied even to very large multimedia databases, by exploiting an appropriate large-scale indexing scheme. Finally, the method can be easily extended in order to address a wider variety of media types and application paradigms. Specific innovative steps are proposed throughout the whole framework and analyzed in detail. Experiments performed on two multimodal datasets prove the superiority and the efficiency of the proposed framework even for cross-modal retrieval.

The rest of the paper is organized as follows: In Section II, an overview of the proposed framework is available. In Section III, the creation of the multimodal feature space, using Laplacian eigenmaps, is analyzed. Large-scale indexing, which is used to make the proposed method applicable to large databases, is described in Section IV. Section V presents the multimodal search and retrieval process, while Section VI analyzes the experimental results. Finally, conclusions are drawn in Section VII.

II. Method Overview and Innovations

When dealing with multimodal search and retrieval, it is much more convenient to enclose multiple media types, which share the same semantics, into a media container, and label the entire container with the semantic concept, instead of labeling each media instance separately. This approach has been already followed in both [17] and [18], where authors introduced new structures to organize data based on their semantic correlations, namely multimedia documents (MMDs) and multimedia bags, respectively. Following the same concept, we adopted the term MMD to refer to rich multimedia representations. An example of an MMD is given in Fig. 1. This describes the physical entity "My_Dog" and consists of the 3-D representation (VRML model), multiple 2-D views (jpeg images) of the dog, as well as its sound (wav file of the barking sound). In the current work, 3-D objects, 2-D images, and sounds are considered as the constituting modalities of MMDs.

The proposed framework is depicted in the block diagrams presented in Figs. 2 and 3. The whole framework is separated in two stages: 1) creation of the multimodal feature space (Fig. 2) and 2) multimodal search and retrieval (Fig. 3).

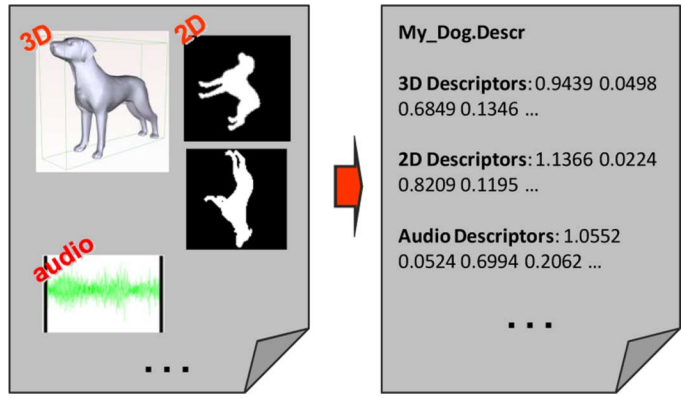

Fig. 1. Example of MMD, which describes the physical entity "My_Dog".

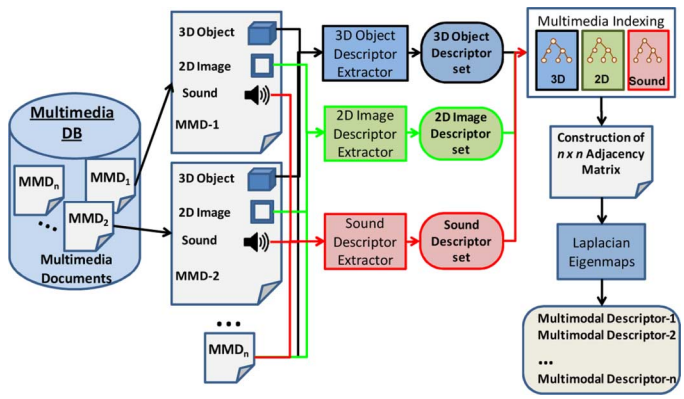

Fig. 2. Creation of the multimodal feature space.

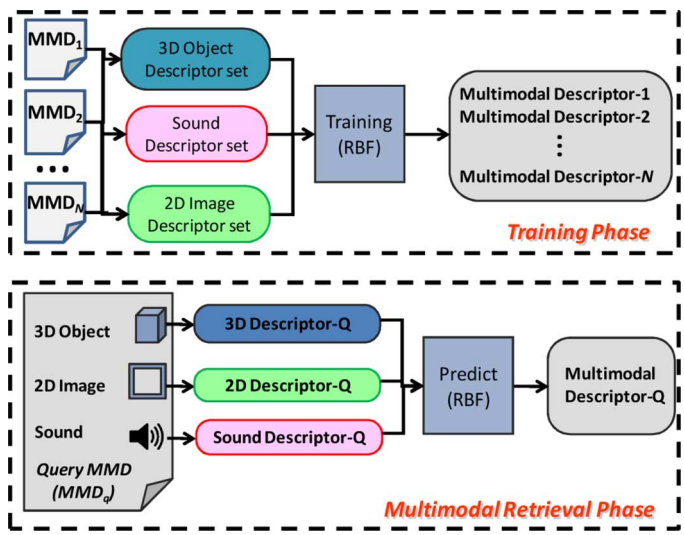

Fig. 3. Multimodal search and retrieval. Training phase: the RBF is trained using as input the descriptor vectors of the database MMDs' constituting modalities and output the corresponding multimodal descriptors. Multimodal retrieval phase: the RBF takes as input the descriptors of $\mathrm{MMD_Q}$ and predicts its multimodal descriptor vector. The latter is matched with the multimodal descriptor vectors of the database.

Given a database of multimedia items, they are organized as MMDs consisting of multiple modalities. During the first stage, the low-level descriptors of all constituting modalities are

mapped to a new low-dimensional feature space. In this feature space, semantically similar MMDs, irrespective of their constituting modalities, are described by multimodal descriptor vectors close to each other in the Euclidean space.

The second stage involves the multimodal search and retrieval procedure. During the training phase, a radial basis function (RBF) network is constructed using a predefined dataset of MMDs. The RBF network is a function that maps the initial descriptors of an MMD to the new low-dimensional multimodal space. During the multimodal retrieval phase, a query MMD, which does not belong to the database, is transformed to a multimodal descriptor vector, using the RBF function. Without the use of RBF, the manifold learning procedure described in the previous stage should be repeated each time a new query is inserted. This poses a computational burden, which is undesirable for online retrieval tasks, especially when database's size increases. This inefficiency can be avoided, thanks to the RBF function. The resulting multimodal descriptor vector is directly matched with the multimodal descriptors of the database MMDs and the most similar MMDs are retrieved.

The proposed method introduces the following innovative features:

Construction of the Adjacency Matrix: In order to identify close similarities between MMDs, an adjacency matrix is constructed. The nonzero elements of the adjacency matrix, which correspond to pairs of neighboring MMDs, are not weighted with respect to multimedia distance measures (as in [17]) but are all assigned the same value (equal to 1). This modification eliminates the need to compute a complex distance metric among MMDs, which would require merging of heterogeneous descriptor vectors of different modalities into one single equation. Instead, the single modalities are ranked separately using their specific distance metric and the first neighbors of each modality are assigned the same nonzero value to construct the adjacency matrix. Moreover, in the proposed approach, all modalities are equally contributing to the creation of the adjacency matrix, by providing the same number of neighbors per modality. This was proven in the end to be more efficient than the approach in [17], where the weights of the combined distance measure are dependent on the average precision of each modality's retrieval performance, because modalities with low-discriminative descriptors were underestimated.

Use of Large-Scale Indexing for the Adjacency Matrix Creation: The creation of the adjacency matrix requires calculating the all-to-all distances among all multimedia objects of the database. To avoid extensive computational time consumption, these distances are usually stored in large distance matrices. However, when the database size increases dramatically, the size of the all-to-all distance matrices becomes prohibitive. In this paper, a large-scale indexing method has been adopted and extended to the multimodal case, so as to accelerate the process of retrieving the nearest neighbors, without the need to store large distance matrices.

Support of Multimodal Queries: In cross-media retrieval, the query of one single modality is used to retrieve items of another modality, while in multimodal retrieval, two or more modalities are used simultaneously as query to retrieve items of multiple modalities. Most of the existing multimedia retrieval methods that deal with more than one media types are cross-media approaches, i.e., they do not support multimodal queries. In this paper, multimodal querying is fully supported, allowing users to enter as query an entire MMD. It should be clearly stressed here that cross-modal search is by no means underestimated. In some cases, querying with multiple modalities would pose an additional processing burden, while using as query a single modality could be more convenient to the user. In general, the querying behavior varies from user to user, so an ideal framework should support both mono-modal and multimodal queries. The proposed framework supports both aforementioned options; thus, it provides a unified solution for diverse types of users.

Provide Efficient Multimodal Query Formulation Using an RBF Network: When a new MMD is used as query, its descriptors need to be mapped to the new multimodal feature space. This is achieved by using an appropriately selected RBF network, which maps the input descriptors of the query MMD to the new multimodal feature space. In this case, the query can be directly matched with the database MMDs. It is worth mentioning that the RBF network achieves mapping of new MMDs, even if one or more modalities are missing. Such an approach has not been reported so far in the area of multimodal search and retrieval. Most of the existing methods use as query a single-modality media item and initially retrieve a ranked list of the media items of the same modality. The latter are used as queries, to further retrieve results of other modalities from the database. However, in this case, the multimodal nature of the query is not fully exploited.

However, the main novelty of the proposed work is that all the above features are combined in order to provide a complete framework for multimodal search and retrieval. This framework can be used in all types of multimodal datasets irrespective of their constituting modalities and the corresponding low-level descriptors. Theoretically, the number of different media types that can be supported simultaneously by the proposed framework is unlimited. The method can scale even to large-scale datasets and it can still retrieve accurate results even in cases where one or more modalities are missing from several multimedia documents of the dataset. To the best of our knowledge, it is the first time that such an approach is presented in the literature.

III. Creation of Multimodal Feature Space

In this section, the creation of the multimodal feature space is analyzed, where all MMDs, irrespective of their constituting modalities, are represented as d-dimensional vectors in a new feature space. In this feature space, semantically similar MMDs will lie close to each other with respect to a common distance metric. The methodology, which is usually followed, is known as manifold learning, where it is assumed that the multimodal data lie on a nonlinear low-dimensional manifold. The majority of manifold learning approaches is based on the computation of the k-nearest neighbors among all items of the dataset in order to create an adjacency matrix. In our case, the items of the dataset are MMDs. The k-nearest neighbor computation for an MMD is not a trivial process, since it requires merging descriptors of heterogeneous modalities into one unified distance metric. To avoid merging of heterogeneous distance metrics, an alternative

approach is introduced in this paper. The method is based on Laplacian eigenmaps (LE) but, in our case, the creation of the adjacency matrix is modified as follows: when items i, j are neighbors, the item W_{ij} of the adjacency matrix is assigned the value 1 instead of the actual distance between i and j. Since the items of the adjacency matrix are MMDs, the neighborhood criterion is determined as follows:

Lemma 1: "Two MMDs, i and j are neighbors if and only if at least one pair of their constituting items of the same modality are neighbors. If the two MMDs do not have items of common modality they are not considered as neighbors. Neighborhood among single-modality items is determined by ranking these items with respect to their mono-modal distance. Then, the k-first items are selected for each single-modality item."

A. Creation of a Multimodal Adjacency Matrix

In this step, the creation of a multimodal adjacency matrix is described in detail. Given a multimedia database of N MMDs and p different modalities, the goal is to compute the k-nearest neighbors for every MMD_i, $1 \leq i \leq N$. For simplicity, we assume that each MMD_i consists of exactly one item per modality, although it is possible to have more than one items of the same modality in MMD_i or even missing modalities.

In order to compute the k-nearest neighbors of MMD_i, the nearest neighbors of each separate modality need to be determined. Let a media item within MMD_i of the mth modality $(1 \leq m \leq p)$ be represented by the descriptor vector $\mathbf{x}_i^m$. For the mth modality, a distance measure is defined as $d^m(\mathbf{x}_i^m, \mathbf{x}_j^m)$ to calculate the mono-modal similarities. The k^m-nearest neighbors of $\mathbf{x}_i^m$ are retrieved by ranking all the media items of the mth modality $(\mathbf{x}_j^m)$ within the database, with respect to their mono-modal distances d^m. The ranked list of k^m-nearest neighbors of $\mathbf{x}_i^m$ is defined as

$$\mathrm{NeighList}_i^m = \mathrm{index}_1^m, \mathrm{index}_2^m, \ldots, \mathrm{index}_{k_m}^m \qquad (1)$$

where index_1^m is the index of the MMD which corresponds to the media item of the mth modality, ranked as the first nearest neighbor of $\mathbf{x}_i^m$. $\mathrm{index}_2^m, \ldots, \mathrm{index}_{k_m}^m$ are the indices of the MMDs corresponding to the $2^{\mathrm{nd}}, \ldots, k^m$th ranked items, respectively. Similarly, p ranked lists of k^m-nearest neighbors for each modality are extracted.

The final k-nearest neighbors of $\mathrm{MMD}_i (k \geq k^m)$ are computed by taking equal number of first neighbors from each list $\mathrm{NeighList}_i^m, 1 \leq m \leq p$, i.e., k/p neighbors, with $(k/p) = k^m$. In case an MMD_j appears in the k/p neighbors of more than one lists $\mathrm{NeighList}_i^m$, this MMD_j is counted only once. The remaining positions in the k-nearest neighbors list are then filled with the next closest MMDs.

In the general case that an MMD consists of less than p modalities, more nearest neighbors are kept from each modality, in order to keep the number k of the neighboring MMDs the same. As an example, let $k = 6$ be the number of k-nearest neighbors of MMD_i. If MMD_i consists of $p = 2$ modalities, we need $(k/p) = 3$ nearest neighbors from each modality. If MMD_i consists of $p = 1$ modality, we need $(k/p) = 6$ nearest neighbors, all from the same modality.

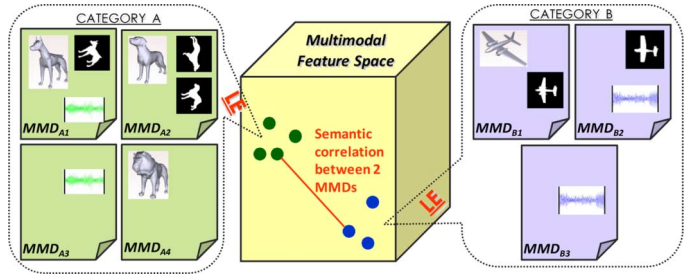

Fig. 4. In the new multimodal feature space created by LE, semantically similar MMDs are placed close to each other, while MMDs of different semantic categories are far from each other.

Finally, an $N \times k$ matrix, $\mathbf{NN}_{CO}$, is created, where each row i represents the k-nearest neighbors of MMD_i.

B. Laplacian Eigenmaps

The $\mathbf{NN}_{CO}$ matrix is used as input by the LE algorithm, where, in our case, the adjacency matrix is modified by putting only ones (instead of distances) to its nonzero elements. The steps of the algorithm are given below:

Step 1) Construct the graph G, by connecting nodes (i.e., MMDs) i and j with an edge, if j is among k-nearest neighbors of i.

Step 2) Produce the $N \times N$ adjacency matrix, $\mathbf{W}$, of G:

$$W_{ij} = \begin{cases} 1, & \text{if MMD j belongs to k nearest neighbors of MMD } i \\ 0, & \text{otherwise.} \end{cases} \qquad (2)$$

Step 3) Create an $N \times N$ diagonal matrix $\mathbf{H}$: $H_{ii} = \sum_j W_{ij}$.

Step 4) Create an $N \times N$ Laplacian matrix $\mathbf{L} = \mathbf{H} - \mathbf{W}$.

Step 5) Solve the generalized eigenproblem $\mathbf{LY} = \boldsymbol{\lambda}\mathbf{HY}$ to find the eigenvalues $\boldsymbol{\lambda}$ and the eigenvectors $\mathbf{Y}$ of $\mathbf{L}$.

Step 6) Sort eigenvalues in an ascending order and keep the l eigenvectors that correspond to the l-first eigenvalues (excluding the first one).

The l selected eigenvectors correspond to l-dimensions of the new multimodal feature space, where all database MMDs are mapped to low-dimensional points (Fig. 4). In this feature space, semantically similar MMDs are placed close to each other, while MMDs of different semantic categories are far from each other.

The method described above shares common features with the manifold learning approaches presented in [17], [18], and [27]. The main differences and contributions of the proposed work are given below. In [17], an all-to-all distance matrix among all MMDs is required as input to multidimensional scaling (MDS). To merge heterogeneous distances between MMDs, the authors in [17] propose a weighting based on the average top-p precision in terms of content-based retrieval of each modality. This introduces a bias comparing with our approach, since classification information of the MMD dataset, which is required to compute precision, may not be available. Moreover, the contribution of modalities with low retrieval precision is underestimated. By using only zeros and ones, in our case, merging of heterogeneous distances is avoided and all modalities contribute equally to the creation of the adjacency

matrix. The method proposed in [18] does not take into account multimodal distances to create the adjacency matrix but it uses mono-modal and cross-modal distances to create multiple adjacency sub-matrices (of all possible pairs). In this case, the computational complexity of the algorithm is dramatically increased, when new modalities are introduced, while, in our case, the complexity of the LE-based method is not affected. A similar adjacency matrix is produced in [27] to be used as input to a manifold ranking approach. The nonzero elements of the matrix are exponential functions of the pairwise distances between the objects. It must be noted, however, that the method in [27] is used only for mono-modal retrieval (either image or text) and not for cross-modal or multimodal retrieval. An extension to multimodal retrieval is not a trivial task; it requires merging distances of heterogeneous modalities, which would suffer from the same weaknesses as in [17]. Finally, it is worth mentioning that the LE method is applied for the first time in this paper to address multimodal retrieval.

IV. Large-Scale Indexing

The nonzero elements W_{ij} of the adjacency matrix $\mathbf{W}$ described above indicate that the jth object is neighbor of the ith object, with respect to a specific distance metric. It is obvious that for the creation of the $N \times N$ adjacency matrix, an $N \times N$ distance matrix is required, which stores the pair-wise distances among all database's MMDs. However, when it comes to really large multimedia datasets, both calculation and storage of all-to-all distance matrices becomes prohibitive. Consequently, the distance matrix does not provide an efficient solution in real-life problems, where multimedia databases store thousands (or even millions) of media items.

On the other hand, multimedia indexing is a widely used method to speed up the nearest-neighbor search in large databases. Through indexing, only a subset of the most relevant data for a given query is returned, without the need to compute one-to-all distances of the query with all database objects. Based on its clear advantages in media retrieval, large-scale indexing has been adopted in the present work to avoid computation of large distance matrices. The indexing algorithm that was extended and used in our multimodal retrieval method has been introduced in [29] and is based on inverted files. The main idea of the method is that when two objects are very similar (close to each other in a metric space), their view of the surrounding world is similar as well. Thus, instead of using the distance between two objects, their similarity can be approximated by comparing their ordering of similarity according to some reference points. This particular technique is also implemented by the use of inverted files. A brief overview of the algorithm is given in the sequel for the sake of completeness.

Let $\boldsymbol{S} = \{o_1, o_2, \ldots, o_M\}$ be a set of M media objects and d a distance function between objects of $\boldsymbol{S}$. Let $\mathrm{RO} \subset S$ be a set of reference objects chosen from $\boldsymbol{S}$. An object $o_i \in S$ can be represented as the ordering $\bar{o}_i$ of the reference objects RO according to their distance d from o_i, as follows: $\bar{o}_i \in O^{\mathrm{RO}}_{d,o_i}$, where O^{RO}_{d,o_i} is the ordered list containing all objects of RO, ordered according to their distance d from o_i. The position in O^{RO}_{d,o_i} of a reference object $ro_j \in \mathrm{RO}$ is denoted as $O^{\mathrm{RO}}_{d,o_i}(ro_j)$. The distance between two objects in the transformed domain is given by $\bar{d}(\bar{o}_1, \bar{o}_2) = \mathrm{SFD}(O^{\mathrm{RO}}_{d,o_1}, O^{\mathrm{RO}}_{d,o_2})$, where SFD is the *Spearman footrule distance*, which is used as a measure to compare ordered lists:

$$\mathrm{SFD}\left(O^{\mathrm{RO}}_{d,o_1}, O^{\mathrm{RO}}_{d,o_2}\right) = \sum_{ro \in \mathrm{RO}} \left| O^{\mathrm{RO}}_{d,o_1}(\mathrm{ro}) - O^{\mathrm{RO}}_{d,o_2}(\mathrm{ro}) \right|. \quad (3)$$

The distance between the two objects in the transformed domain can be used to perform approximate similarity search, instead of using the classical distance metric d.

Let us suppose that we have a query q, which is used to retrieve relevant objects o_i from $\boldsymbol{S}, i = 1, 2, \ldots M$. An exhaustive approach would be to compute the pairwise distances $d(q, o_i)$ of the query descriptor vector with the descriptors of all objects o_i of the dataset $\boldsymbol{S}$. The approximate ordering of $\boldsymbol{S}$ with respect to q can be obtained by computing the distance $\bar{d}(\bar{q}, \bar{o}_i), \forall o \in \boldsymbol{S}$. This distance can be easily computed by representing (indexing) the transformed objects with inverted files, as follows:

Entries of the inverted file are the objects of RO. The posting list associated with an entry $ro_j \in \mathrm{RO}$ is a list of pairs $(o_i, O^{\mathrm{RO}}_{o_i}(\mathrm{ro}_j)), o_i \in \mathbf{S}$, that is a list where each object o_i of the dataset S is associated with the position of the reference object ro_j in $\bar{o}_i$. In other words, each reference object is associated with a list of pairs each referring an object of the dataset and the position of the reference object in the transformed objects representation. The inverted file will have the following structure:

$$\begin{aligned} &\mathrm{ro}_1 \rightarrow \left(\left(o_1, O^{\mathrm{RO}}_{o_1}(ro_1)\right), \ldots, \left(o_M, O^{\mathrm{RO}}_{o_M}(ro_1)\right)\right) \\ &\ldots \\ &\mathrm{ro}_n \rightarrow \left(\left(o_1, O^{\mathrm{RO}}_{o_1}(ro_n)\right), \ldots, \left(o_M, O^{\mathrm{RO}}_{o_M}(ro_n)\right)\right) \end{aligned} \quad (4)$$

where M is the size of the dataset $\boldsymbol{S}$ and n is the size of the set of reference objects RO. An algorithm for computing the distances $\bar{d}$ of the query q with all objects o of $\boldsymbol{S}$ is given in Fig. 5. At the end of the algorithm, all objects are associated with an accumulator a_o that contains their distance $\bar{d}$ from the query q.

By using the above indexing structure, search within the dataset $\boldsymbol{S}$ is much faster than using the classical distance metric d to calculate dissimilarity between descriptor vectors. The search and retrieval time depends on the size of the dataset

INPUT:
  query: q,
  reference objects: RO
  posting lists associated with reference objects
OUTPUT:
  The set of accumulators A
ALGORITHM
  Set $A \leftarrow \{\}$
  For each $ro_j \in RO$
    Let pl be the posting list associated with ro_j
    For each $(o_i, O^{RO}_{d,o_i}(ro_j)) \in pl$
      If $a_o \notin A$
        Set a_o=0
        Set $A \leftarrow A \cup \{a_o\}$
      Set $a_o = a_o + \left| O^{RO}_{d,q}(ro_j) - O^{RO}_{d,o_i}(ro_j) \right|$

Fig. 5. Searching algorithm using inverted files.

of reference objects RO. According to [29], the following inequality must hold so that the retrieval performance is not affected:

$$\text{Size(RO)} \geq 2 \cdot \sqrt{\text{Size}(S)}. \quad (5)$$

Working in a similar way, we adopted the aforementioned multimedia indexing scheme in order to create the adjacency matrix of the manifold learning approach presented in this paper. The indexing algorithm is applied for each modality separately; thus, the dataset $\boldsymbol{S}$ is the set of media items o of the same mth modality and d is the distance metric $d^m(\mathbf{x}_i^m, \mathbf{x}_j^m)$ that computes the dissimilarity between the mono-modal descriptors $\mathbf{x}^m$ of the mth modality. Then, a ranked list of k_m-nearest neighbors is returned, similar to the one of (1).

The above indexing scheme is also applicable to the methods presented in [18] and [27], since they take into account only the distances from neighboring objects to create an adjacency matrix. On the other hand, the method in [17] requires computation of all-to-all distances to apply MDS; therefore, large-scale indexing is not applicable in this case. In the Experimental Results section, the performance of the proposed LE-based manifold learning method, with and without using large-scale indexing, is presented.

V. Multimodal Search and Retrieval

In state-of-the-art cross-media retrieval systems, the user enters a query of a single modality to retrieve objects of different modalities, i.e., use an audio file to retrieve relevant images, use an image query to retrieve relevant sounds, and so on. The framework proposed in this paper is able to support multimodal queries. As an example, an MMD can be used as query to retrieve semantically similar MMDs. The constituting modalities of the retrieved MMDs may be different from the query's modalities, which is a clear step forward in the field of multimodal retrieval.

By using as query an MMD that belongs to the database, the retrieval procedure is straightforward: the low-dimensional multimodal descriptor vector of the query MMD, which was computed using the proposed LE-based manifold learning method, is matched against the multimodal descriptor vectors of the rest MMDs of the database and the most relevant results are retrieved. The situation, though, is different when dealing with queries which do not belong to the database. An MMD that does not belong to the database is not included in the manifold learning process, and thus, its low-dimensional multimodal descriptor vector is not available. Therefore, it cannot be directly matched with the database MMDs.

A. Dealing With Multimodal Queries Which Do Not Belong to the Database

In the complex case where the query does not belong to the database, the only information that can be extracted is the initial mono-modal descriptors of its constituting media items. Instead of repeating the procedure described in Section IV, by adding the query to the initial dataset, a faster and more approximate solution is preferred in order to obtain its multimodal descriptor vector. Towards this direction, several machine learning techniques (such as neural networks, SVMs, etc.) can be adopted to train a sample dataset taking as input the initial descriptor vectors and producing the final low-dimensional vectors. Such an approach was presented in [14], where an RBF network was applied to map the initial low-level descriptors of 3-D objects to a new feature space of lower dimension. However, in [14], authors deal with one single modality. The situation is more complex when two or more modalities need to be trained simultaneously, as is the case in the present work.

In this paper, the RBF [31] was eventually chosen. The implementation of this method was obtained from the Weka [32] library. The reason for choosing RBF instead of similar machine learning techniques (such as SVM) was the fact that RBF was the only method that supported missing input data. Since our multimodal datasets consist of MMDs with one or more modalities missing, no other method could be applied for training.

An ideal RBF network, in our case, would take as input the mono-modal descriptors of the query MMD's constituting modalities and return an l-dimensional multimodal descriptor vector to be matched with the multimodal descriptors of the database MMDs. If D_m is the dimension of the mth modality's descriptor vector ($m = 1, \ldots, p$) and p is the number of the MMDs' constituting modalities, the total number D of inputs to the RBF is given by

$$D = \sum_{m=1}^{p} D_m. \quad (6)$$

Although current implementations of RBF functions support multiple inputs, even when the number D of inputs is large, they do not support multiple outputs. This is due to the fact that RBFs are mainly classifiers; thus, they can only return a class label in the output. To deal with this limitation, the method for predicting the multimodal vector of a query MMD has been modified as follows:

Let $\boldsymbol{M} = \{\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_N\}$ be a set of N multimodal descriptor vectors, where each vector $\mathbf{m}_i$ consists of l descriptors ($\mathbf{m}_i = [\mathbf{m}_i(1), \mathbf{m}_i(2), \ldots, \mathbf{m}_i(l)]^T$). By applying a K-means clustering algorithm, the multimodal descriptor vectors of $\boldsymbol{M}$ can be grouped into Kl-dimensional clusters ($K < N$). Let also $\boldsymbol{C} = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_K\}$ be the set of cluster centers produced by the K-means clustering algorithm, where $\mathbf{c}_i$ is also an l-dimensional vector, and $\mathbf{CL} = \{L_1, L_2, \ldots, L_K\}$ is the set of cluster labels associated with each cluster center. A multimodal descriptor vector $\mathbf{m}_i \in \boldsymbol{M}$ is assigned the label L_j associated with the cluster center $\mathbf{c}_j$ closer to $\mathbf{m}_i$. This can be written mathematically as

$$\arg(L) = \arg\left(\min_j \left(\sqrt{\sum_{u=1}^{l} (\mathbf{m}_i(u) - \mathbf{c}_j(u))^2}\right)\right). \quad (7)$$

During the training procedure, each MMD of the training set is used as training sample to the RBF network as follows: the input to the RBF is the set of D mono-modal descriptors of the MMD, while the output is the cluster label L_j assigned to the MMD's multimodal descriptor vector $\mathbf{m}_i$.

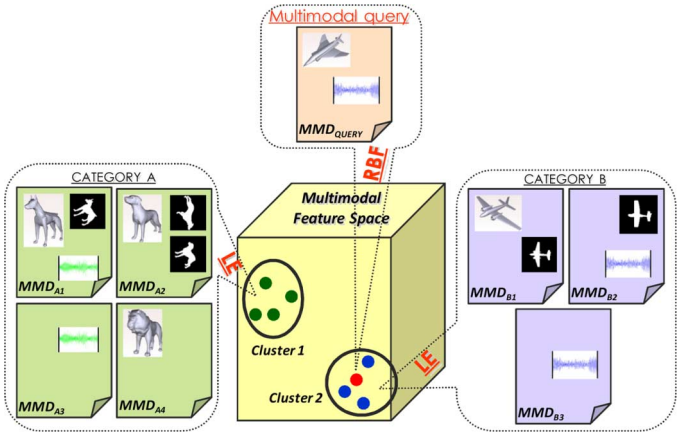

Fig. 6. Multimodal retrieval using a query MMD which does not belong to the database.

During the multimodal retrieval phase, the query MMD_Q enters the RBF network using its D mono-modal descriptors as input. Then, the RBF function predicts a cluster label L_Q from the set $\mathbf{CL}$. The cluster center $\mathbf{c}_Q$ that is associated with L_Q is used as multimodal descriptor vector to retrieve similar MMDs from the database.

An interesting property of RBF networks is that they support missing inputs. Thus, the performance of an RBF is not significantly affected when multiple inputs are empty. Therefore, even if one or more modalities are missing from several MMDs of the database and the corresponding D_m RBF inputs are empty, the network can still train the RBF function successfully. Similarly, during retrieval, several modalities from the query MMD can be missing, which again does not affect the prediction accuracy of the RBF.

In Fig. 6, the procedure of multimodal retrieval using a query MMD which does not belong to the database is illustrated. The MMDs of the database are represented as vectors in the multimodal feature space. Then, clusters are created, which group semantically similar MMDs together. When an MMD out of the database is used as query, the RBF predicts the center of the cluster that is closer to this MMD. The cluster center is then used as the multimodal descriptor of the query in order to retrieve similar MMDs from the database. Note that two MMDs that consist of different modalities can belong to the same cluster, which demonstrates the ability of the proposed method to model semantic relationships in the multimodal space.

Comparing with similar methods presented in [17] and [18], the method proposed in this paper is novel in the sense that it supports multimodal queries. In [17], when the query does not belong to the database, only one single modality at a time can be used as query. Similarly, the method in [18] supports querying with one single modality at a time. The RBF framework, on the other hand, enables querying with multiple modalities simultaneously. It must be noted that the present framework based on RBF is used for the first time in multimodal retrieval, which is a clear advantage of the proposed method. Moreover, to the best of our knowledge, there is no method reported so far that supports the option to enter multiple query modalities simultaneously.

VI. Experimental Results

For the experimental evaluation of the proposed method, three multimodal datasets were compiled by us, since, to the best of our knowledge, no benchmark dataset for multimodal retrieval is available. The first dataset consists of 264 MMDs, which were created using 159 3-D objects and 312 2-D images. The 3-D objects constitute a subset of the ITI 3-D object database [40], which has been used for experimental testing of 3-D object retrieval methods, while the 2-D images are snapshots of the corresponding 3-D objects. The classification scheme of ITI database has been adopted in the first multimodal dataset to classify the 264 MMDs into 12 categories. For the creation of the second dataset, 266 3-D objects, 370 2-D images, and 283 sounds resulted in a total number of 495 MMDs. The 3-D objects are a subset of the SHREC 2011 Generic Shape Benchmark [41] (10 out of the 50 categories), which has been used in SHREC 2011 contest for experimental testing of 3-D object retrieval methods, while the 2-D images are snapshots of the corresponding 3-D objects. The 3-D objects and 2-D images were classified into 10 categories using the SHREC 2011 classification scheme. The 283 sounds were collected from publicly available websites of the Internet and were manually attached to specific MMDs. Finally, the third dataset comprises 2334 MMDs classified into 50 semantic categories. We used 1550 real images and 1557 3-D objects to create the third dataset. The 3-D objects are derived from both the SHREC 2011 Generic Shape Benchmark and the Princeton Shape Benchmark [42] datasets, while the 2-D real images were collected from publicly available websites of the Internet and were manually attached to specific 3-D objects. The classification schemes of SHREC 2011 and Princeton Shape Benchmark were used to classify the 2334 MMDs into 50 semantic categories.

The 3-D object descriptors for all three datasets were extracted using the combined Depth-Silhouette-Radialized Extent (DSR) descriptor [35]. The 2-D image descriptors for the first two datasets were extracted using 2-D Polar-Fourier coefficients, 2-D Zernike moments and 2-D Krawtchouk moments [3], while for the images of the third dataset the CEDD descriptor [36] was used, which constructs a vector of 144 3-bit values based on color and edge histogram. The reason for choosing different low-level image descriptors was that, in the first two datasets, images are actually snapshots of the corresponding 3-D objects (no background or color information is available), while the third dataset consists of real images gathered from the web. Thus, for the first two datasets, shape descriptors are appropriate, while for the third dataset, background and color information should be taken into account, which led to the choice of the CEDD descriptor. Finally, the audio descriptors of the second dataset are extracted using the algorithm presented in [34].

The datasets used in our experiments can be downloaded from the following web links:

http://3d-test.iti.gr:8080/3d-test/Download/Multimodal_Database_1.zip
http://3d-test.iti.gr:8080/3d-test/Download/Multimodal_Database_2.zip
http://3d-test.iti.gr:8080/3d-test/Download/Multimodal_Database_3.zip

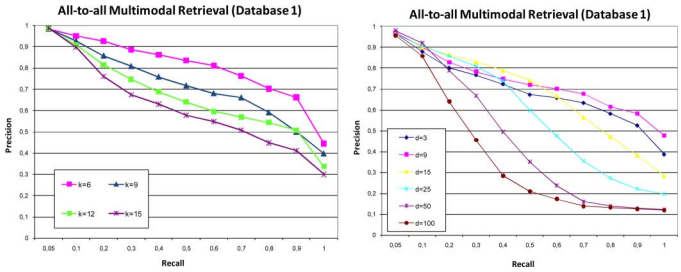

Fig. 7. Parameter selection of values k and l in Database 1.

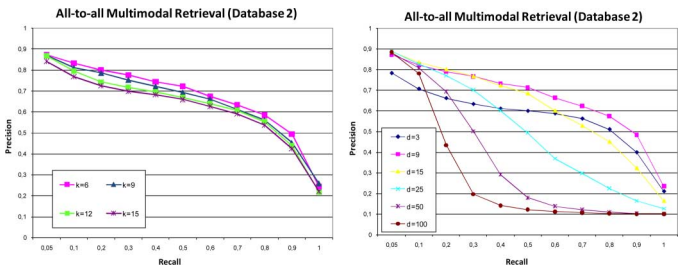

Fig. 8. Parameter selection of values k and l in Database 2.

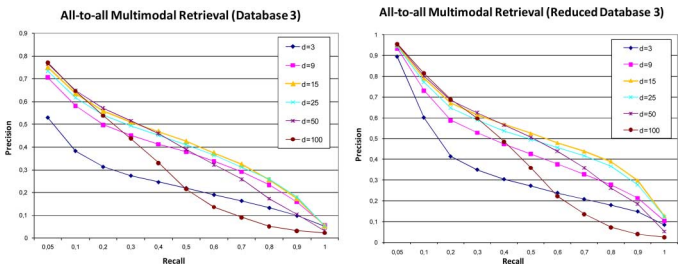

Fig. 9. Parameter selection of values l (a) in Database 3 and (b) in a subset of Database 3 where almost half of the MMDs have been removed.

In the first series of experiments, each MMD from the database was used as query to retrieve similar MMDs. In the second, new MMDs not belonging to the database were used as queries. The retrieval performance was evaluated in terms of "*precision-recall*", where precision is the proportion of the retrieved models that are relevant to the query and recall is the proportion of relevant models in the entire database that are retrieved. The classification information of the three datasets was used as a ground truth, in order to distinguish the relevant MMDs from the irrelevant ones. When the retrieved result belongs to the same category with the query, it is marked as relevant; otherwise it is marked as irrelevant.

A. Parameter Selection for the Laplacian Eigenmaps Algorithm

The performance of the proposed manifold learning method based on Laplacian eigenmaps is affected by the number k of nearest neighbors of MMD_i required for the creation of the adjacency matrix and the number l of dimensions of the new multimodal feature space. In Fig. 7–9, the precision-recall diagrams for different values of k and l, in all three datasets, are presented. In order to generate these diagrams, each MMD of a dataset is used as query to retrieve MMDs from the same dataset (all-to-all multimodal retrieval). After calculating the individual precision-recall for each query, the average precision-recall is extracted for the entire dataset. Concerning the value k of nearest neighbors, the best performance is achieved for $k = 6$ in all three datasets.

The aim of applying nonlinear dimensionality reduction, which is achieved by using the LE method, is to keep the distances of neighboring points close enough, while at the same time to stretch the distances of non-neighboring points [37]. This improves the discriminative power of the distance metric (Euclidean distance). Regarding the dimensionality of the new mapped multimodal descriptor vectors, a value of $l = 9$ proved to be the optimal choice for the first two datasets. By reducing the dimensionality to values $l \leq 3$, the retrieval accuracy is decreased, which implies that less than four dimensions are not adequate to provide a complete data representation for the two datasets. By increasing the dimensionality to values $l > 15$, the accuracy starts to decrease again. This is due to the fact that by adding more dimensions, the effect of distance stretching becomes weaker and the discriminative power of Euclidean metrics in this space is reduced.

Similar behavior is observed in the third dataset; however, in this case, the optimal performance is achieved for $l = 15$. Although the optimal dimensionality here is different from the one observed for the previous two datasets ($l = 9$), this was expected taking into account the different nature of low-level mono-modal descriptors (e.g., the real image descriptors of the third dataset are totally different from the image descriptors of Database 1 and Database 2). In order to prove that this parameter is not affected by the size of the dataset, we repeated the experiments on a reduced dataset, which was created from Database 3 by keeping only half of its MMDs. The results for both Database 3 (2334 MMDs) and reduced Database 3 (1177 MMDs) are presented in Fig. 9, where it is obvious that the optimal dimensionality in both cases is $l = 15$.

B. Comparison of the Proposed Approach With Other Manifold Learning Methods

The proposed method was compared with the following manifold learning approaches: LRGA [17], MMR [4], and LLE [15]. In order to implement these methods, an $N \times N$ dissimilarity matrix among all MMDs of each experimental dataset was initially created. Each cell of the matrix is a weighted sum of the mono-modal distances, for each pair of MMDs, where the weights represent the average precision values for each modality, according to the method in [17]. The implementation of the LRGA algorithm was available at the authors' website, while both MMR and LLE methods were implemented by us.

A comparison of our method with the competitive ones is presented in Fig. 10. The precision-recall diagrams correspond to all-to-all multimodal retrieval in all three datasets. In the first two datasets, the proposed method is slightly better than LRGA and MMR, while it clearly outperforms the LLE method. In the third dataset, the proposed approach outperforms all other competitive methods.

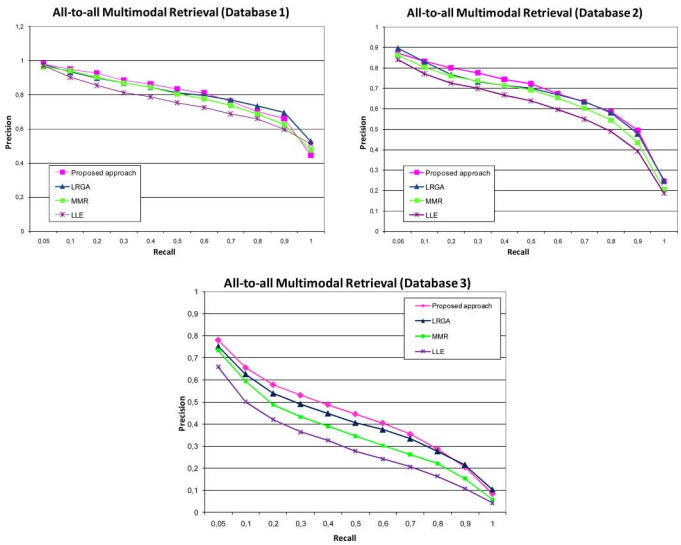

Fig. 10. Comparison of the proposed method (all-to-all multimodal retrieval) against LRGA, MMR, and LLE, in all three databases.

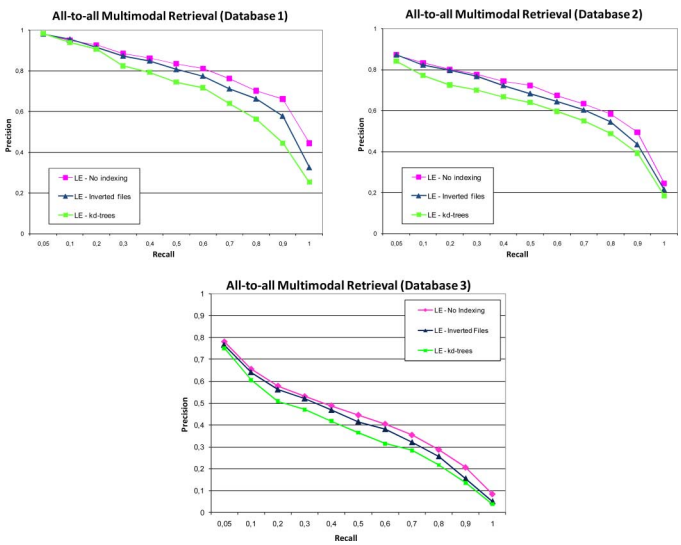

Fig. 11. Comparison of the proposed method (all-to-all multimodal retrieval), in all three databases, using two multimedia indexing schemes.

C. Performance of the Proposed Approach Using Large-Scale Indexing

In Fig. 11, the performance of the proposed method using two different indexing algorithms is presented for all three datasets. The first one is the inverted-file-based indexing method [29], while the second one is a method based on *kd*-trees [30]. The method based on inverted files was eventually selected since it achieves higher retrieval accuracy than *kd*-trees. However, the most interesting observation is that the performance of the proposed method, when no indexing is used, is not significantly affected when the inverted-file-based indexing method is adopted. This is of high importance taking into account that in large multimedia databases, indexing can drastically reduce the computational cost and storage requirements.

Another advantage of the proposed framework against the LRGA method, which achieved the best performance among the other two competitive methods, is that LRGA requires creation of a large dissimilarity matrix in order to apply multidimensional scaling. Since our method requires computation of only the nearest neighbors of every MMD, it can be easily combined with the proposed large-scale indexing scheme. This can significantly reduce computation times and storage requirements in very large databases.

D. Performance of the Proposed Approach for Queries That Do Not Belong to the Database

One of the main innovative features of the proposed method is that it supports multimodal queries, that is, an MMD, which does not belong to the database can be used as query with all its constituting modalities simultaneously. This is very important, since searching with multiple queries simultaneously can retrieve more relevant results than using one query at a time. The proposed method uses an RBF network to predict the multimodal descriptor vector of a query $\mathrm{MMD_Q}$ that does not belong to the database. In order to perform this experiment, for each dataset, a set of query MMDs, which do not belong to the dataset, was used. More specifically, 12, 10, and 100 query MMDs were used for the first, second, and third dataset, respectively. In the first database, the query MMDs consist of two modalities (2-D images and 3-D objects), in the second, they consist of three modalities (2-D, images, 3-D objects and sounds), while in the third, they consist of two modalities (2-D real images and 3-D objects). In some of the queries, one or more modalities are missing, in order to check the ability of RBF to support missing inputs. The numbers of clusters (Section V) were found experimentally to be $K = 15$ for the first two datasets and $K = 47$ for the third dataset.

The proposed approach was compared with the method presented in [17] for queries that do not belong to the database. More specifically, the method in [17] uses a query of a single modality to produce an initial ranked list. Then, it uses the LRGA algorithm to retrieve results from the database, using as input the first retrieved results of the previous mono-modal ranked list. In order to have a common comparison basis, the MMDs of our three multimodal query sets were split to single modality queries. These were used as queries to both the LRGA-based method and the proposed method. The performance is given in Fig. 12. Although in the previous experiment, where the query object belongs to the dataset, the proposed method was only slightly better than the one based on LRGA, in the case where the query does not belong to the dataset the proposed method clearly outperforms the LRGA-based method (Fig. 12) in all three datasets. This is of great importance if we consider that the latter corresponds to a real-life case, where queries usually do not belong to the target dataset. Moreover, if instead of single modality queries, the entire MMDs are used, the retrieval accuracy is further improved, which highlights the improvement that is achieved by using multimodal queries instead of mono-modal queries.

E. Computational Issues

In terms of computational efficiency of the proposed method, we will focus on the following: 1) offline processing time, which

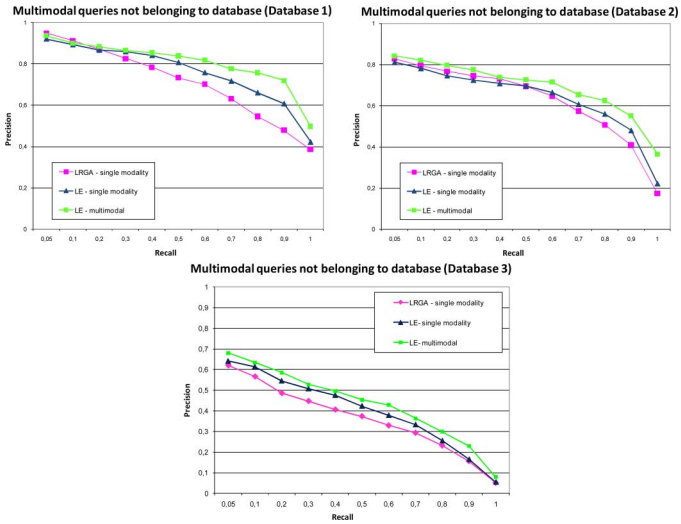

Fig. 12. Comparison of the proposed method against LRGA for queries that do not belong to the database.

TABLE I
COMPARISON OF THE PROPOSED METHOD TO THE THREE COMPETING ONES IN TERMS OF OFFLINE AND ONLINE PROCESSING TIMES

.	Offline processing (pre-processing) (msec)			Online processing (search one-to-all) (msec)		
Method	Dataset 1	Dataset 2	Dataset 3	Dataset 1	Dataset 2	Dataset 3
Proposed Method	4623	8012	131383	0.6	2.3	40.5
LRGA	438000	3717441	237945224	4.5	20.6	193.64
MMR	6312	10266	477765	39.4	88.7	886.9
LLE	6265	8626	108218	7.5	16.6	44.6

TABLE II
COMPUTATIONAL EFFICIENCY OF MULTIMODAL INDEXING IN TERMS OF COMPUTATION TIME AND STORAGE REQUIREMENTS

.	Without Multimedia Indexing			Using Multimedia Indexing		
	Dataset 1	Dataset 2	Dataset 3	Dataset 1	Dataset 2	Dataset 3
Time for computing all-to-all nearest neighbours (msec)	2485	7437	121734	759	1372	8689
Storage Requirements (KBytes)	409.6	1698.1	38440	16.2	102	976.4

involves the procedure of multimodal feature space creation; 2) online processing time, which is the time for one-to-all matching of a query MMD to the MMD dataset, and 3) the computational complexity of the multimedia indexing scheme. In Table I, the proposed framework is compared to the other three competing methods (LRGA [17], MMR [4], and LLE [15]) in terms of computation times. It is worth mentioning that, in all cases, the times for extraction of the low-level descriptors for each modality were not included, since they are the same for all cases. In the offline processing stage, the most time consuming method is the LRGA method [17]. The reason is that LRGA utilizes MDS during pre-processing, which has a complexity of $O(N^3)$ [38]. Therefore, for very large datasets, the offline processing time becomes prohibitive. In the online processing stage, the MMR method [4] is the most time consuming since it involves an iterative procedure. The proposed method achieves low computation times in both cases. The times were obtained using a PC with a dual-core 2.4-GHz processor and 8 GB of RAM.

In Table II, the contribution of large-scale indexing to computation time and storage requirements is presented. According to Table II, the time for computing all-to-all nearest neighbors (exhaustive search) in the first dataset, is 2485 ms without indexing, and 759 ms with the use of indexing. Thus, computing all-to-all nearest neighbors becomes 3 times faster when indexing is applied. The time cost improvement is more obvious as we move from dataset 1 to dataset 2 (bigger), since the all-to-all neighbor computation becomes 5.4 times faster (from 7437 ms to 1372 ms). Finally, when indexing is applied on the third dataset, which is the biggest among our test datasets, the all-to-all neighbor computation becomes 14 times faster (from 121 734 ms to 8689 ms). This clearly proves that the time cost improvement is always increasing with the increase of database sizes.

However, the efficiency of the proposed indexing method with respect to computational cost and storage requirements cannot be demonstrated using a limited multimedia dataset, such as the ones used in the current work. In order to illustrate the capabilities of indexing, a theoretical example follows. Let us assume a database of $M = 50\,000$ 3-D objects. Low-level descriptors of these objects are extracted using STT [33], i.e., an l-dimensional descriptor vector ($l \approx 2000$) is extracted for every 3-D object. For a query q, similarity search within this database, without using indexing, involves $M \times l = 10^8$ calculations, if a simple distance metric (such as L2 distance) is used. If the proposed indexing method is adopted, a total of $l \times S_{\mathrm{RO}} + S_{\mathrm{RO}} \times M'$ calculations is required, where $S_{\mathrm{RO}} = 450$ is the number of reference objects and M' is the number of objects in each inverted file (4), which are actually accessed. According to [29], not all of the $M = 50\,000$ objects of the inverted file need to be accessed for a given query q. Only a number of $M' = 100$ objects is enough to obtain the accurate results. Therefore, the total number of calculations, using the proposed indexing method is reduced to 945 000. This reduction in computational cost is more distinct as the database size increases. Concerning the storage requirements, it must be noted that the use of indexing obviates the need to store a pre-calculated distance matrix. The size of the distance matrix is proportional to $M \times M$, where M^2 is the number of database objects. On the other hand, the storage requirements of the proposed indexing method is proportional to $S_{\mathrm{RO}} \times M = 2 \cdot M^{3/2}$, which is more compact than the distance matrix.

The proposed framework is available for testing at the following link: http://www2.isearch-project.eu:8080/isearch/search/index.php. The first dataset has been used for this demo. The user is able to insert as query an MMD that either belongs or does not belong to the dataset and retrieve relevant MMDs. The framework supports queries of multiple modalities simultaneously.

VII. CONCLUSION

In this paper, a novel framework for multimodal search and retrieval was presented. The framework achieves retrieval of multiple media types using as queries multiple modalities simultaneously. The multiple media types are organized into rich media representations, called MMDs. The proposed multimodal retrieval framework is appropriate for search and retrieval of MMDs using as query even an entire MMD. The method creates a new low-dimensional feature space, using Laplacian eigenmaps, where all MMDs can be mapped irrespective of their constituting modalities. Then, multimodal retrieval of MMDs is achieved by simply computing the pairwise distances

among their low-dimensional multimodal descriptor vectors. When the query MMD does not belong to the database, an RBF network was trained to map the MMD's mono-modal descriptors to the new low-dimensional feature space. Finally, the proposed framework can be applied even to very large multimedia databases, by exploiting an appropriate large-scale indexing scheme.

Experiments performed on two multimodal datasets demonstrated the superiority and the efficiency of the proposed method in multimodal search and retrieval. Another interesting conclusion is that when multiple query modalities are used simultaneously, higher retrieval accuracy is achieved than using each modality separately. Finally, the method can be easily extended in order to address a wider variety of media types and application paradigms.

REFERENCES

[1] A. Mademlis, P. Daras, A. Axenopoulos, D. Tzovaras, and M. G. Strintzis, "Combining topological and geometrical features for global and partial 3D shape retrieval," *IEEE Trans. Multimedia*, vol. 10, no. 5, pp. 819–831, Aug. 2008.

[2] A. Mademlis, P. Daras, D. Tzovaras, and M. G. Strintzis, "Ellipsoidal harmonics for 3D shape description and retrieval," *IEEE Trans. Multimedia*, vol. 11, no. 8, pp. 1422–1433, Dec. 2009.

[3] P. Daras and A. Axenopoulos, "A compact multi-view descriptor for 3D object retrieval," in *Proc. IEEE 7th Int. Workshop Content-Based Multimedia Indexing (CBMI 2009)*, Chania, Greece, Jun. 2009.

[4] T. P. Vanamali, A. Godil, H. Dutagaci, T. Furuya, Z. Lian, and R. Ohbuchi, "SHREC'10 track: Generic 3D warehouse," in *Proc. Eurographics/ACM SIGGRAPH Symp. 3D Object Retrieval*, 2010.

[5] M. Worring and T. Gevers, "Interactive retrieval of color images," *Int. J. Image Graph.*, vol. 1, no. 3, pp. 387–414, 2001.

[6] M. Kokare, P. K. Biswas, and B. N. Chatterji, "Texture image retrieval using new rotated complex wavelet filters," *IEEE Trans. Syst., Man, Cybern. B*, vol. 35, no. 6, pp. 1168–1178, Dec. 2005.

[7] E. Attalla and P. Siy, "Robust shape similarity retrieval based on contour segmentation polygonal multiresolution and elastic matching," *Pattern Recognit.*, vol. 38, no. 12, pp. 2229–2241, Dec. 2005.

[8] P. Geetha and V. Narayanan, "A survey of content-based video retrieval," *J. Comput. Sci.*, vol. 4, no. 6, pp. 474–486, 2008.

[9] J. Sivic, F. Schaffalitzky, and A. Zisserman, "Efficient object retrieval from videos," Ph.D. dissertation, Univ. Oxford, Oxford, U.K., 2006.

[10] A. Joly, O. Buisson, and C. Frélicot, "Content-based copy retrieval using distorsion-based probabilistic similarity search," *IEEE Trans. Multimedia*, vol. 9, no. 2, pp. 293–306, Feb. 2007.

[11] P. Wan and L. Lu, "Content-based audio retrieval: A comparative study of various features and similarity measures," *Proc. SPIE*, vol. 6015, p. 60151H, 2005.

[12] G. Tzanetakis, G. Essl, and P. Cook, "Audio analysis using the discrete wavelet transform," in *Proc. Conf. Acoustics and Music Theory*, Sep. 2001.

[13] G. Li and A. A. Khokhar, "Content-based indexing and retrieval of audio data using wavelets," in *Proc. ICME*, 2000.

[14] R. Ohbuchi and J. Kobayashi, "Unsupervised learning from a corpus for shape-based 3D model retrieval," in *Proc. ACM MIR*, Santa Barbara, CA, 2006.

[15] L. K. Saul and S. T. Roweis, "Think globally, fit locally: Unsupervised learning of low dimensional manifolds," *J. Mach. Learn. Res.*, 2003.

[16] J. R. He, M. J. Li, H. J. Zhang, H. H. Tong, and C. S. Zhang, "Manifold-ranking based image retrieval," in *Proc. ACM MM*, New York, 2004.

[17] Y. Yang, D. Xu, F. Nie, J. Luo, and Y. Zhuang, "Ranking with local regression and global alignment for cross media retrieval," in *Proc. ACM MM*, Beijing, China, 2009.

[18] H. Zhang and J. Weng, "Measuring multi-modality similarities via subspace learning for cross-media retrieval," *Advances in Multimedia Information Processing—PCM*, 2006.

[19] F. Wu, H. Zhang, and Y. Zhuang, "Learning semantic correlations for cross-media retrieval," in *Proc. IEEE Int. Conf. Image Processing*, 2006.

[20] D. Li, N. Dimitrova, M. Li, and I. K. Sethi, "Multimedia content processing through cross-modal association," in *Proc. 11th ACM Int. Conf. Multimedia (MM'03)*, 2003.

[21] M. Li, D. Li, N. Dimitrova, and I. K. Sethi, "Audio-visual talking face detection," in *Proc. Int. Conf. Multimedia and Expo (ICME)*, Baltimore, MD, Jul. 2003, pp. 473–476.

[22] P. L. Lai and C. Fyfe, "Canonical correlation analysis using artificial neural networks," in *Proc. Eur. Symp. Artificial Neural Networks (ESANN)*, 1998.

[23] Y.-T. Zhuang, Y. Yang, and F. Wu, "Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval," *IEEE Trans. Multimedia*, vol. 10, no. 2, pp. 221–229, Feb. 2008.

[24] Y. Zhuang, Q. Li, and L. Chen, "A unified indexing structure for efficient cross-media retrieval," *DASFAA 2009, LNCS 5463*, pp. 677–692, 2009.

[25] A. Messina and M. Montagnuolo, "A generalised cross-modal clustering method applied to multimedia news semantic indexing and retrieval," in *Proc. 18th Int. Conf. World Wide Web (WWW 2009)*, Madrid, Spain, Apr. 20–24, 2009.

[26] H. Zhang and F. Meng, "Multi-modal correlation modeling and ranking for retrieval," *PCM 2009, LNCS 5879*, pp. 637–646, 2009.

[27] D. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Scholkopf, "Ranking on data manifold," in *Proc. Conf. Advances in Neural Information Processing Systems (NIPS)*, 2003.

[28] H. Tong, J. HE, M. Li, C. Zhang, and W. Y. Ma, "Graph based multi-modality learning," in *Proc. ACM MM*, Singapore, 2005.

[29] C. Gennaro, G. Amato, P. Bolettieri, and P. Savino, "An approach to content-based image retrieval based on the Lucene search engine library," in *Proc. 14th Eur. Conf. Research and Advanced Technology for Digital Libraries (ECDL'10)*, Heidelberg, Germany: Springer-Verlag, 2010.

[30] M. Muja and D. G. Lowe, "Fast approximate nearest neighbors with automatic algorithm configuration," in *Proc. Int. Conf. Computer Vision Theory and Applications (VISSAPP'09)*, 2009.

[31] A. G. Bors, "Introduction of the radial basis function (RBF) networks," in *Proc. Online Symp. Electronics Engineers*, Feb. 13, 2001, vol. 1, DSP Algorithms: Multimedia, no. 1, pp. 1–7. [Online]. Available: http://www.osee.net.

[32] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.

[33] D. Zarpalas, P. Daras, A. Axenopoulos, D. Tzovaras, and M. G. Strintzis, "3D model search and retrieval using the spherical trace transform," *EURASIP J. Adv. Signal Process.*, vol. 2007, Jul. 2006, Article ID 23912, 14 pp., doi: 10.1155/2007/23912.

[34] G. Wichern, J. Xue, H. Thornburg, B. Mechteley, and A. Spanias, "Segmentation, indexing, and retrieval for environmental and natural sounds," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 688–707, Mar. 2010.

[35] D. Vranic, "3D model retrieval," Ph.D. dissertation, Univ. Leipzig, Leipzig, Germany, 2004.

[36] S. A. Chatzichristofis and Y. S. Boutalis, "CEDD: Color and edge directivity descriptor—A compact descriptor for image indexing and retrieval," in *Proc. 6th Int. Conf. Advanced Research on Computer Vision Systems (ICVS 2008), Proceedings: Lecture Notes in Computer Science (LNCS)*, Santorini, Greece, May 12–15, 2008, pp. 312–322.

[37] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 15, no. 6, pp. 1373–1396, 2003.

[38] I. Borg and P. Groenen, *Modern Multidimensional Scaling: Theory and Applications*. New York: Springer-Verlag, 2005.

[39] P. Daras, D. Zarpalas, D. Tzovaras, and M. G. Strintzis, "Shape matching using the 3D radon transform," in *Proc. 3D Data Processing, Visualization & Transmission (3DPVT 2004)*, Thessaloniki, Greece, Sep. 2004.

[40] P. Daras, D. Zarpalas, D. Tzovaras, and M. G. Strintzis, "Efficient 3-D model search and retrieval using generalized 3-D radon transforms," *IEEE Trans. Multimedia*, vol. 8, no. 1, pp. 101–114, Feb. 2006.

[41] H. Dutagaci, A. Godil, P. Daras, A. Axenopoulos, G. Litos, S. Manolopoulou, K. Goto, T. Yanagimachi, Y. Kurita, S. Kawamura, T. Furuya, R. Ohbuchi, B. Gong, J. Liu, and X. Tang, "SHREC'11 track: Generic shape retrieval," in *Proc. 4th Eurographics Workshop 3D Object Retrieval (3DOR 2011)*, Llandudno, U.K., Apr. 10, 2011.

[42] P. Shilane, P. Min, M. Kazhdan, and T. Funkhouser, "The Princeton shape benchmark," in *Proc. Shape Modeling Int. (SMI'04)*, Genova, Italy, Jun. 2004, pp. 167–178.

Petros Daras (M'07) was born in Athens, Greece, in 1974. He received the Diploma degree in electrical and computer engineering, the M.Sc. degree in medical informatics, and the Ph.D. degree in electrical and computer engineering, all from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 1999, 2002, and 2005, respectively.

He is a Researcher Grade C, at the Informatics and Telematics Institute (ITI) of the Centre for Research and Technology Hellas (CERTH). His main research interests include search, retrieval and recognition of 3-D objects, 3-D object processing, medical informatics applications, medical image processing, 3-D object watermarking, and bioinformatics. He serves as a reviewer/evaluator of European projects.

Dr. Daras is a key member of the IEEE MMTC 3DRPC IG.

Stavroula Manolopoulou was born in Larissa, Greece, in 1984. She received the B.Sc. degree in informatics and the M.Sc. degree in digital media, both from Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2007 and 2009, respectively.

She has been a Research Assistant at the Informatics and Telematics Institute (ITI) of the Centre for Research and Technology-Hellas (CERTH) since January 2010. Her main research interests include digital processing of medical images, biomedical signal processing, and multimedia content-based search and retrieval.

Apostolos Axenopoulos was born in Thessaloniki, Greece, in 1980. He received the Diploma degree in electrical and computer engineering and the M.S. degree in advanced computing systems from Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2003 and 2006, respectively. Currently, he is pursuing the Ph.D. degree in the Computer and Communication Engineering Department, University of Thessaly, Volos, Greece.

He has been an Associate Researcher at the Informatics and Telematics Institute (ITI) of the Centre for Research and Technology-Hellas (CERTH) since November 2003. His main research interests include 3-D object indexing, content-based search and retrieval, and bioinformatics.