

# A Unified Framework for Multimodal Retrieval

D. Rafailidis, S. Manolopoulou, P. Daras\*

*Information Technologies Institute, Centre for Research and Technology Hellas,  
Thessaloniki, 6th km Xarilaou - Themi, 57001, P.O.Box: 60361, Thessaloniki, Greece*

---

## Abstract

In this paper, a unified framework for multimodal content retrieval is presented. The proposed framework supports retrieval of rich media objects as unified sets of different modalities (image, audio, 3D, video and text), by efficiently combining all monomodal heterogeneous similarities to a global one according to an automatic weighting scheme. Then, a multimodal space is constructed, to capture the semantic correlations among multiple modalities. In contrast to existing techniques, the proposed method is also able to handle external multimodal queries, by embedding them to the already constructed multimodal space, following a space mapping procedure of a submanifold analysis. In our experiments with five real multimodal datasets, we show the superiority of the proposed approach against competitive methods.

### *Keywords:*

Multimedia description, multimodal search and retrieval, incremental manifold learning.

---

\*Corresponding author

*Email addresses:* [drafail@iti.gr](mailto:drafail@iti.gr) (D. Rafailidis), [manolop@iti.gr](mailto:manolop@iti.gr) (S. Manolopoulou), [daras@iti.gr](mailto:daras@iti.gr) (P. Daras)

## 1. Introduction

The continuously increasing amount of multimedia content on the Internet emerged the imperative need for searching in various online multimedia databases. The traditional text-based retrieval techniques failed to address the requirements for searching this massive media content, therefore, research has been focused on content-based multimedia retrieval methods. Searching for similar to a query content requires the automatic extraction of low-level features from media, e.g. in case of an image these would be color, texture, shape, etc. Thus, several content-based techniques have been developed in the past, performing retrieval of a single modality, such as 3D objects [11, 19], images [1, 31], video [12, 23] or audio [2, 30].

However, users who search for content are interested in finding semantically similar results to a query, regardless of its modality. Towards this aim, Yang et al. [37] proposed a method for connecting various semantically similar media of different modalities. In order to manage the case of having different modalities that carry the same semantics, the concept of Multimedia Document (MMD) was introduced. An example of a MMD is presented in Figure 1, which describes a physical entity of an airplane and consists of its 3D representation, real image and sound.

Recently, multimedia search engines have evolved, allowing combinations of queries of different modalities. Multimodal search allows users to enter multiple query types and retrieve multiple types of media simultaneously in the form of MMDs. An approach for multimodal search has been in-




AIRLINER_002	
<u>Media Item</u>	<u>Descriptor</u>
	0.9439 0.0498 0.6849 0.1346 ...
	1.1366 0.0224 0.8209 0.1195 ...
	0.9439 0.0498 0.6849 0.1346 ...
...	

Figure 1: Multimedia Document (MMD) example containing three modalities: 3D, image and audio.

roduced by the I-SEARCH <sup>1</sup> framework [3]. I-SEARCH is a real world application, which enables retrieval of several types of media (3D objects, 2D images, sound, video and text) using as query any of the above types or their combinations in the form of MMDs. I-SEARCH made a significant step towards content-based multimedia retrieval, where users can search and retrieve media of any modality using a single unified retrieval framework and not a specialized system for each separate modality. Moreover, users in I-SEARCH can enter multiple queries simultaneously and thus, retrieve more relevant results. However, handling media in the form of MMDs is a highly complicated process, since the successful modeling of the low-level feature associations among the different modalities is required, in order to perform multimodal retrieval.

---

<sup>1</sup>Available at <http://vcl.itl.gr/is>

## 2. Related Work and Contribution

### *2.1. Dimensionality Reduction Methods for Monomodal Retrieval*

In content-based retrieval methods, media are usually represented by low-level features in the form of high-dimensional descriptor vectors in which a distance metric (more often Euclidean-based) is applied to calculate similarity. However, since in most cases the extracted high-dimensional descriptor vectors face the problem of Dimensionality Curse [6], such metrics are inappropriate for efficient large scale retrieval. Therefore, nonlinear dimensionality reduction methods based on Manifold Learning [4, 5, 22] were proposed for mapping the high-dimensional descriptors to a more representative feature space of lower dimensions. Such methods have been widely applied in monomodal cases, like 3D [20], image [29], and audio [21] based retrieval, raising significantly the retrieval accuracy.

### *2.2. Multimodal Retrieval Methods*

Similar approaches were recently introduced for multimodal retrieval in order also to produce a low dimensional feature space, able to combine feature spaces of different modalities. The existing multimodal methods are divided into two broad categories where: (a) the cases of internal multimodal queries and external monomodal queries are handled; and (b) the generic category of multimodal retrieval, where also the case of external multimodal queries is handled. At this point we must specify that according to [9], a multimodal query is considered as a rich media object in the form of a MMD, constituting of medias of different modalities simultaneously, whereas a monomodal query is considered as a media of a certain modality.

A multimodal retrieval method was presented in [37], where a global MMD distance measure is calculated as a weighted distance of the monomodal distances. Then, the dimensionality reduction method of Multidimensional Scaling is applied, so as to construct a unified multimodal (MMD) space, where each MMD is represented as a point. Afterwards, a Laplacian matrix is constructed, using the Local Regression and Global Alignment (LRGA) technique, to generate the ranked lists of each query. A weakness of LRGA is that it supports multimodal queries when they exist in the database and only monomodal, otherwise. Additionally, efficient multimodal retrieval is not ensured, since the global MMD distance is highly dependent on the discriminative power of each separate monomodal descriptor.

Besides LRGA, several multimedia retrieval approaches were also proposed in the literature, capable of handling internal multimodal queries and external monomodal queries. For example, Yang et al. [36] proposed to generate a semi-semantic graph (MMDSSG), based on which a Cross Media Indexing Space (CMIS) is constructed. Then, for each query the optimal dimension of CMIS was determined and the multimodal retrieval was performed on a per-query basis. Additionally, relevance feedback methods were exploited to improve the retrieval performance. However, the case of external multimodal queries was not supported. Zhang et al. [39] applied the Laplacian Eigenmaps method to construct a semantic space of MMDs, called Multi-modality Laplacian Eigenmaps Semantic Subspace (MLESS), so as to map the monomodal query to the center of its monomodal neighbors in MLESS and retrieve MMD results. Additionally, in [33], Wu et al. proposed a multimodal retrieval method, following the Canonical Correla-

tion Analysis (CCA), in order to create an isomorphic subspace. Moreover, through one or more relevance feedback iterations, authors demonstrated that the retrieved results could be further improved for the case of internal queries. 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. Another multimodal retrieval method, called Cross-modal Factor Analysis (CFA) [18], identifies the correlations between two different modalities and performs a dimensionality reduction method to build the semantic space of MMDs. CFA proved to be superior against other similar approaches, such as the CCA method [17]. Alternatively, instead of constructing the semantic space of MMDs, several multimodal retrieval methods followed different strategies. For example, in the Kernel Canonical Correlation method [38], the correlations between the modalities are identified, so as to perform multimodal retrieval for the case of internal queries. However, all the aforementioned works do not support the case of external multimodal queries.

Additionally, in the work of [16], the retrieved results by each modality are combined, using reranking and late fusion methods, in order to perform multimodal retrieval. Towards this direction, several works extended multimodal retrieval methods for mobile phones like the works of [10, 34], [35]. Nevertheless, since the aforementioned methods avoid constructing the semantic space of MMDs, multimodal descriptor vectors cannot be generated, indexed and stored. Therefore, scalable content-based retrieval was not ensured.

In contrast to all the aforementioned multimodal methods, a promising

multimodal retrieval method has been recently proposed in [9], where the semantic space of MMDs is built based on the Manifold Learning method of Laplacian Eigenmaps [5]. In order to preserve the local neighborhood of each media into the low-dimensional MMD space, a multimodal adjacency matrix is constructed. In case of internal queries, the nearest neighbors per modality are computed and then, an equal number of each modality’s nearest neighbors are combined to form the multimodal adjacency matrix, in which ones and zeros declare that two MMDs are neighbors or not, respectively. Consequently, multimodal descriptor vectors are generated from the constructed MMD space. The generated multimodal descriptors are then indexed and stored into the multidimensional structure of [13]. Moreover, in order to handle the case of posing external multimodal queries (in contrast to the aforementioned multimodal retrieval methods, where only the case of monomodal external queries was handled), a clustering method is applied to organize the constructed MMD space into  $CL$  predefined clusters. Afterwards, an RBF network is trained to handle the missing modalities of an external MMD-query and thus, to predict the center of the cluster that is closer to the MMD-query. The cluster center is then used as the multimodal descriptor of the external query and semantically similar MMDs are retrieved from the database.

Despite the fact that the multimodal retrieval method of [9] seems to be promising, two important factors can be further elaborated: (a) the more efficient construction of the MMD space and (b) the more successful handling of the case of posing external queries. In particular, by following the “simple minded” way of combing equal number of each modality’s nearest neighbors

so as to form the multimodal adjacency matrix, the availability of modalities in the database is omitted, which impacts on constructing the MMD space inefficiently. This happens, because the Laplacian Eigenmaps method tries to preserve the local neighborhood of each media, and therefore more media from a certain modality have higher contribution to the construction of the MMD space. Moreover, in case of posing external queries, the predicted clusters by the RBF network are insufficient to capture the semantic correlations among the internal MMDs and an external MMD-query, since the semantic space of MMDs evolves over time along with the continuous increase of multimedia content. Therefore, a method is required for embedding the external MMD-query into the already constructed multimodal semantic space, instead of predicting the missing modalities of the external MMD-query, as it happens in [9]. This is of great importance if we consider that the case of external queries corresponds to a real-life case, where queries usually do not belong to the database.

### *2.3. Contribution*

In this paper, both aforementioned challenges are successfully handled since all monomodal heterogeneous similarities are combined to a global MMD similarity by applying an automatic weighting scheme, taking into account the availability of modalities per MMD in the database. Based on the proposed global MMD similarity a heat kernel is built, which is capable of preserving the local neighborhood of each media modality. Thus, the low-dimensional MMD space is generated efficiently by capturing the semantic correlations among MMDs. Additionally, the proposed method is able to handle external MMD-queries, by embedding them into the existing MMD



space, following the space mapping technique of a submanifold analysis [15]. Therefore, for each external query, a representative multimodal descriptor vector is generated, based on which accurate multimodal content retrieval is achieved. As we will experimentally show, the proposed method, especially in the case of posing external MMD-queries, is capable of achieving high retrieval accuracy, along with the increase of the number of available modalities and the size of the database. Last but not least, while all aforementioned multimodal methods address up to three modalities, to the best of our knowledge, this is the first work which deals with five modalities (text, 3D, image, video, sound) simultaneously. In our experiments with five real multimodal datasets of different scale, we show the superiority of the proposed approach, in terms of retrieval accuracy and time efficiency, against other state-of-the-art multimodal content retrieval methods.

### 3. The Proposed Method

The proposed method is divided into the following two steps, (a) construction of the multimodal semantic space of MMDs and (b) multimodal search and retrieval.

#### 3.1. Construction of the Multimodal Semantic Space

An overview of this process, is depicted in Figure 2. Given a multimedia database of  $N$  MMDs with up to  $M$  different modalities each, the final goal is to represent each  $MMD_i$ , with  $1 \leq i \leq N$ , by a multimodal descriptor vector  $\mathbf{y}_i$  in the Euclidean space  $\mathbb{R}^d$ , where  $d$  denotes the dimensions of the multimodal semantic space of MMDs. For each  $MMD_i$  its  $\mathbf{x}_i^m$  monomodal descriptors ( $1 \leq m \leq M$ ) are extracted, and then for each  $m$ -th modality the

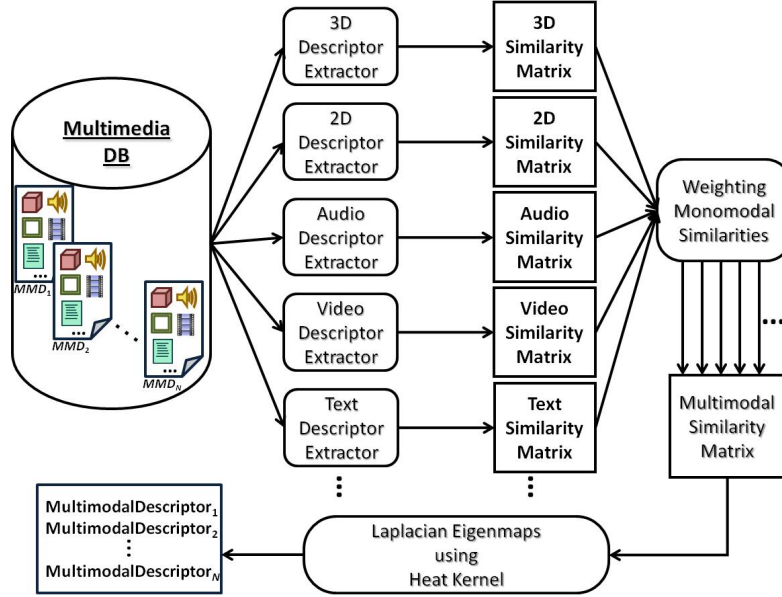


Figure 2: Constructing the multimodal semantic space of MMDs.

respective monomodal similarity matrix is calculated. Next, by following a new weighting scheme, the  $M$  different types of monomodal similarities are combined into a multimodal similarity. Then, a heat kernel is calculated, reflecting on the similarities between the  $N$  MMDs. Finally, according to the Laplacian Eigenmaps method and the calculated heat kernel, the multimodal semantic space of MMDs is constructed, and consequently, for each  $MMD_i$ , a multimodal descriptor  $\mathbf{y}_i$  is generated.

### 3.2. Calculation of the Multimodal Similarity Matrix

Initially, each monomodal similarity matrix  $\mathbf{S}_m$  is calculated, by considering (a) the availability of modalities per MMD in the database and (b) the different nature of the monomodal similarities as follows:

$$\mathbf{S}_m(i, j) = \begin{cases} 1 - \|\mathbf{x}_i^m - \mathbf{x}_j^m\|^{(m)} & , \text{ if } \exists \mathbf{x}_i^m \text{ and } \mathbf{x}_j^m \\ 0 & , \text{ otherwise} \end{cases} \quad (1)$$

where  $\|\mathbf{x}_i^m - \mathbf{x}_j^m\|^{(m)}$  ( $1 \leq i, j \leq N$ ) is the  $m$ -th monomodal distance between  $MMD_i$  and  $MMD_j$ , normalized to the range of  $[0,1]$  and 0s denote the absence of the  $m$ -th modality, in order to create the monomodal similarity matrices.

However, despite the fact that all monomodal matrices  $\mathbf{S}_m$  are normalized to the same interval, the similarity values differ significantly, because of the different nature of each monomodal distance  $\|\mathbf{x}_i^m - \mathbf{x}_j^m\|^{(m)}$  thus, the monomodal similarity matrices are not comparable. For this reason, an ‘‘alignment’’ transformation of the distributions of the similarity matrices of all  $M$  modalities is performed. In particular, in order to ‘‘shift’’ their similarity values towards the same point,  $\forall \mathbf{S}_m : 1 \leq m \leq M$ , a new monomodal similarity matrix  $\mathbf{S}'_m$  is calculated according to the  $Z$ -Score transformation as follows:

$$\mathbf{S}'_m(i, j) = \frac{\mathbf{S}_m(i, j) - \mu_{\mathbf{S}_m}}{\sigma_{\mathbf{S}_m}} \quad (2)$$

where  $\mu_{\mathbf{S}_m}$  and  $\sigma_{\mathbf{S}_m}$  are the mean value and the standard deviation of the normalized matrix  $\mathbf{S}_m$ , respectively. Next, each transformed matrix  $\mathbf{S}'_m$  is normalized to the interval of  $[0,1]$ . By doing so, all the transformed monomodal matrices  $\mathbf{S}'_m$  can be compared, since they share the same range of values and the same distribution.

Let us denote as  $gd_m$  the global density of  $\mathbf{S}'_m$  and as  $ld_m(i)$  the local density of each  $MMD_i$ , where  $gd_m = f_m/N^2$  and  $ld_m(i) = l_m(i)/N$ , where

$f_m$  is the total number of non-zero values in the transformed monomodal matrix  $\mathbf{S}'_m$  and  $l_m(i)$  is the total number of non-zero values in the  $i$ -th row of  $\mathbf{S}'_m$ . For each  $MMD_i$  the weight  $a_m(i)$  of its  $m$ -th modality is calculated according to:

$$a_m(i) = \frac{\frac{ld_m(i)}{gd_m}}{\sum_{p=1}^M \frac{ld_p(i)}{gd_p}} \quad (3)$$

For high values of local density  $ld_m(i)$  and low values of global density  $gd_m$ , the fraction  $\frac{ld_m(i)}{gd_m}$  becomes high in the  $m$ -th modality of  $MMD_i$ . This can be interpreted as follows: if  $MMD_i$  contains a media of the  $m$ -th modality, whereas the global density  $gd_m$  is low, when the rest of MMDs do not contain often a media of the  $m$ -th modality, then according to Equation (3) weight  $a_m(i)$  becomes high, so as to express that the  $m$ -th modality is more important for  $MMD_i$ , compared to the rest of MMDs. Additionally,  $a_m(i)$  is weighted by the sum  $\sum_{p=1}^M \frac{ld_p(i)}{gd_p}$  in order to express the importance of the  $m$ -modality for  $MMD_i$  compared to the rest of its modalities. Therefore, according to Equation (3), a high value of  $a_m(i)$  for  $MMD_i$  corresponds to a high weight of its  $m$ -th modality.

Then, in order to calculate the  $N \times N$  multimodal matrix  $\mathbf{S}_{mult}$ , the similarity between  $MMD_i$  and  $MMD_j$  is derived by:

$$\mathbf{S}_{mult}(i, j) = \frac{1}{M} \cdot \sum_{m=1}^M \frac{a_m(i) \cdot \mathbf{S}'_m(i, j) + a_m(j) \cdot \mathbf{S}'_m(i, j)}{a_m(i) + a_m(j)} \quad (4)$$

where the similarity  $\mathbf{S}'_m(i, j)$  between  $MMD_i$  and  $MMD_j$  is weighted by the respective weights  $a_m(i)$  and  $a_m(j)$ . Based on Equation (3) it holds

$a_m(i) \neq a_m(j)$ . This happens because despite the fact that the global density  $gd_m$  is preserved for all MMDs, the local densities  $ld_m(i)$  and  $ld_m(j)$  for  $MMD_i$  and  $MMD_j$  may be unequal, since the total number of non-zero values in the  $i$ -th and  $j$ -th row of  $\mathbf{S}'_m$  may differ, i.e. the  $l_m(i)$  and  $l_m(j)$  values, respectively, depending on the number of constituting modalities of  $MMD_i$  and  $MMD_j$ . Therefore, since  $a_m(i) \neq a_m(j)$ , the nominator of the fraction in the sum of Equation (4) ensures symmetry, while the dominator normalizes the fraction to the interval  $[0 \ 1]$ . By doing so, we ensure that the multimodal matrix  $\mathbf{S}_{mult}$  is symmetric, a prerequisite for the next step of the Laplacian Eigenmaps method.

### 3.3. Laplacian Eigenmaps using the Heat Kernel Approach

The Laplacian Eigenmaps algorithm using the heat kernel approach, is adapted to the multimodal framework as follows. Firstly, an adjacency graph  $G$  is constructed, where an edge  $\langle i, j \rangle$  is formed, with  $1 \leq i, j \leq N$ , if  $MMD_j$  is among the  $k$ -nearest neighbors of  $MMD_i$ , based on  $\mathbf{S}_{mult}$ . Next, the weights of edges  $\langle i, j \rangle$  in  $G$  are calculated based on the heat kernel approach, in order to form the  $N \times N$  adjacency matrix  $\mathbf{W}$  according to:

$$\mathbf{W}(i, j) = \begin{cases} e^{-\frac{1 - \mathbf{S}_{mult}(i, j)}{t}} & , \text{ if } \exists \langle i, j \rangle \in G \\ 0 & , \text{ otherwise} \end{cases} \quad (5)$$

where the heat kernel reflects on the similarity information of the symmetric multimodal matrix  $\mathbf{S}_{mult}$  between nodes-MMDs  $i$  and  $j$ , derived by the Gaussian kernel function and stored in the adjacency matrix  $\mathbf{W}$ . Also,  $t \in \mathbb{R}$  denotes the weight of  $\mathbf{S}_{mult}$  in the heat kernel, where for the extreme case

of  $t = \infty$  (which results in  $\mathbf{W}(i, j) = 1$ ), the heat kernel equals the “simple minded” approach [5, 15]. Alternatively, other types of kernel functions could be used such as linear and polynomial, thoroughly examined in [14] for machine learning methods.

Afterwards, we consider the problem of mapping the weighted graph  $G$  to a low-dimensional space, so that connected nodes-MMDs stay as close as possible. Let the  $N \times d$  matrix  $\mathbf{Y}$  be such a map, where the  $i$ -th row corresponds to the multimodal coordinates of  $MMD_i$ . Let, also  $\mathbf{H}$  be a diagonal weighting matrix, whose entries are column sums of  $\mathbf{W}$ , with  $\mathbf{H}(i, i) = \sum_j \mathbf{W}(j, i)$ . Then, the Laplacian matrix  $\mathbf{L}$ , with  $\mathbf{L} = \mathbf{H} - \mathbf{W}$ , is symmetric, positive and semidefinite, which can be considered as an operator on functions defined on nodes of  $G$ . Next, eigenvalues and eigenvectors are computed for solving the generalized eigenvector problem as follows:

$$\mathbf{L}\mathbf{Y} = \lambda\mathbf{H}\mathbf{Y} \tag{6}$$

Let the column vectors  $\mathbf{y}(0), \dots, \mathbf{y}(d)$  be the solutions of (6), ordered according to their eigenvalues,  $0 = \lambda_0 \leq \lambda_1 \leq \dots \leq \lambda_d$ . By excluding the eigenvector  $\mathbf{y}(0)$  and using the next  $d$  eigenvectors, each MMD is mapped to the  $d$ -dimensional Euclidean space:

$$MMD_i \rightarrow (y_i(1), y_i(2), \dots, y_i(d)) \tag{7}$$

Consequently, each MMD is mapped to a specific position into a unified multimodal space, where semantically similar MMDs lie close, forming neighborhoods, as presented in Figure 3.

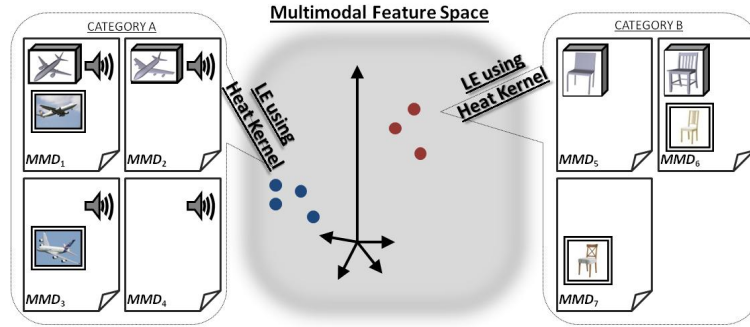


Figure 3: In the multimodal semantic space, each MMD is represented by a point, where semantically similar MMDs lie close to form neighborhoods.

### 3.4. Multimodal Search and Retrieval

By constructing the  $d$ -dimensional semantic space of MMDs, the proposed method supports multimodal search and retrieval for internal and/or external MMD-queries. In the case that the MMD-query  $Q$  belongs to the database, its multimodal descriptor vector  $\mathbf{y}_Q$  already exists, since it has already been mapped to the  $d$ -dimensional multimodal space according to (7). Thus,  $\mathbf{y}_Q$  is compared to multimodal descriptors  $\mathbf{y}_i$  ( $1 \leq i \leq N$ ) of the MMDs in the database, using the Euclidean distance, in order to retrieve the most similar MMDs to  $Q$ .

However, in the case of posing a MMD-query, which does not belong to the database, a different procedure is followed, since its low-dimensional multimodal descriptor vector is not available and thus a space mapping procedure is required to embed the external MMD-query into the already constructed space of MMDs. Let  $MMD_{N+1}$  be the MMD-query that does not belong to the database of the  $N$  MMDs. Initially, the  $k$ -nearest neighbors of  $MMD_{N+1}$  are found, by calculating the multimodal similarities between

$MMD_{N+1}$  and the  $N$  MMDs in the database, according to (1), (2), (3) and (4). Let  $\mathbf{X}_s = \{MMD_{S(1)}, \dots, MMD_{S(k)}, MMD_{N+1}\}$  be the set of MMDs, including (a) the  $k$ -nearest neighbors of  $MMD_{N+1}$  and (b) the external query  $MMD_{N+1}$ . According to [15], the MMD set  $\mathbf{X}_s$  can be considered as a submanifold, based on which the space mapping procedure for the Laplacian Eigenmaps method is developed as follows:

1. Laplacian Eigenmaps on the submanifold.

A full sub-adjacency matrix  $(k+1) \times (k+1)$ ,  $\mathbf{W}_S$ , of the submanifold is constructed following the heat kernel approach:

$$\mathbf{W}_S(i, j) = e^{-\frac{1 - \mathbf{S}_{mult}(i, j)}{t}} \quad (8)$$

with  $i, j$  denote  $MMD_i, MMD_j \in X_S$ ,  $t \in \mathbb{R}$  and each node in  $\mathbf{X}_S$  is connected to all other nodes. At this point, we must mention that in the case of following the “simple minded” approach, it would result in a full  $\mathbf{W}_S$  sub-adjacency matrix, having 1s in all  $(k+1) \times (k+1)$  cells, since each node in the  $\mathbf{X}_s$  submanifold is connected to all other nodes. Thus, by not considering the proposed global similarity weighting scheme of Equation (4), the eigen-decomposition of the full  $\mathbf{W}_S$  sub-adjacency matrix would not be feasible. Next, both  $(k+1) \times (k+1)$  matrices  $\mathbf{H}_S$  and  $\mathbf{L}_S$ , are computed according to:

$$\mathbf{H}_S(i, i) = \sum_j \mathbf{W}_S(j, i) \quad (9)$$

$$\mathbf{L}_S = \mathbf{H}_S - \mathbf{W}_S \quad (10)$$



with  $i, j = 1, \dots, k + 1$ . Eigenvalues and eigenvectors are then computed by solving the generalized eigenvector problem:

$$\mathbf{L}_S \mathbf{v} = \lambda_S \mathbf{H}_S \mathbf{v} \quad (11)$$

Let the column vectors  $\mathbf{v}(0), \dots, \mathbf{v}(d)$  be the solutions of (11), ordered according to their eigenvalues  $0 = \lambda_S^0 \leq \lambda_S^1 \leq \dots \leq \lambda_S^d$ . Low-dimensional coordinates for  $MMD_{S(1)}, \dots, MMD_{S(k)}, MMD_{N+1}$  on the submanifold are calculated according to:

$$MMD_i \rightarrow (\mathbf{v}_i(1), \dots, \mathbf{v}_i(d)), \forall MMD_i \in X_S \quad (12)$$

2. Calculating the multimodal descriptor vector  $\mathbf{y}_{N+1}$ .

The  $\mathbf{v}_{N+1}$  coordinates are transformed to the global  $\mathbf{y}_{N+1}$  coordinates, by preserving the relationships between  $MMD_{N+1}$  and its  $k$ -nearest neighbors  $MMD_{S(1)}, \dots, MMD_{S(k)}$ . Therefore, by applying the Laplacian Eigenmaps method on the submanifold, the global coordinates  $\mathbf{y}_{N+1}$  for  $MMD_{N+1}$  are computed according to:

$$MMD_{N+1} \rightarrow \mathbf{y}_{N+1} = \sum_{i=1}^k c_i \mathbf{y}_{S(i)} \quad (13)$$

where  $\mathbf{y}_{S(1)}, \dots, \mathbf{y}_{S(k)}$  are the low-dimensional coordinates in the multimodal space of  $MMD_{S(1)}, \dots, MMD_{S(k)}$  and  $c_i \in \mathbb{R}^k, \forall i : 1 \leq i \leq k$  are the constrained weights, which are calculated by minimizing the reconstruction error according to:

$$\min_{c_i} \left| \mathbf{v}_{k+1} - \sum_{i=1}^k c_i \mathbf{v}_i \right|^2 \quad (14)$$

with  $\sum_i c_i = 1$ .

Consequently, since (a) the external query  $MMD_{N+1}$  is mapped into existing multimodal semantic space of MMDs, as depicted in Figure 4 and (b) the multimodal descriptor vector  $\mathbf{y}_{N+1}$  of  $MMD_{N+1}$  is calculated according to (13),  $\mathbf{y}_{N+1}$  is compared to multimodal descriptors  $\mathbf{y}_i$  ( $1 \leq i \leq N$ ) of the MMDs in the database, using the Euclidean distance, in order to retrieve the most similar MMDs.

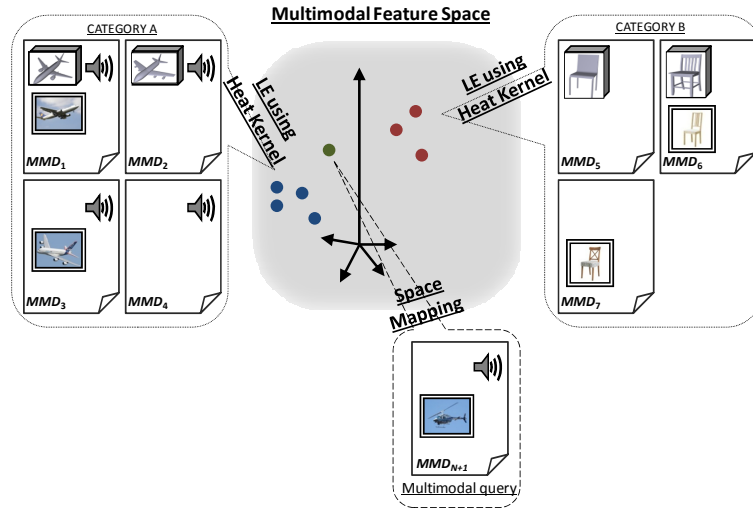


Figure 4: According to (13) each external query  $MMD_{N+1}$  is mapped to the multimodal semantic space.

## 4. Experimental Evaluation

### 4.1. Datasets

Experimental evaluation is performed in five real multimodal datasets, denoted by DS1<sup>2</sup>, DS2<sup>3</sup>, DS3<sup>4</sup>, DS4<sup>5</sup>, and DS5<sup>6</sup>. Further details are provided in Table 1.

Dataset	MMDs	Classes	3D	2D	Audio	Video	Text
DS1	264	12	✓	✓			
DS2	495	10	✓	✓	✓		
DS3	2 334	50	✓	✓			
DS4	2 779	50	✓	✓			✓
DS5	637	43	✓	✓	✓	✓	✓

Table 1: Evaluation Datasets

DS1, DS2 and DS3 are the evaluation datasets of the work in [9]. DS4 is a superset of DS3 by manually adding the text modality in the form of labels, relevant to the corresponding MMD’s content, since a MMD may contain more than one labels. Note that text was assigned to a subset of MMDs. Additionally, in DS4 the categorization of MMDs remains the same as in DS3. Finally, all media of different modalities of DS5 were crawled from the Internet, except for the text modality, which was assigned in the same way as in DS4.

---

<sup>2</sup><http://3d-test.iti.gr:8080/3d-test/Download/MultimodalDatabase1.zip>

<sup>3</sup><http://3d-test.iti.gr:8080/3d-test/Download/MultimodalDatabase2.zip>

<sup>4</sup><http://3d-test.iti.gr:8080/3d-test/Download/MultimodalDatabase3.zip>

<sup>5</sup><http://3d-test.iti.gr:8080/3d-test/Download/MultimodalDatabase4.zip>

<sup>6</sup><http://3d-test.iti.gr:8080/3d-test/Download/MultimodalDatabase5.zip>

In order to ensure the effectiveness of the proposed method, several media descriptors were extracted for each modality, due to the existence of different media in the evaluation datasets. In particular, for DS1, DS2, DS3 and DS4, the 3D object descriptors were extracted using the combined Depth-Silhouette-Radialized Extent (DSR) descriptor [28]. For DS5, 3D descriptors were extracted according to the Compact Multiview Descriptor (CMVD) [19]. For DS1 and DS2 2D image descriptors were extracted based on 2D Polar-Fourier coefficients, 2D Zernike moments and 2D Krawtchouk moments [8]. For DS3 and DS4 2D image descriptors were extracted according to the CEDD descriptor [7]. For DS5, the 2D color descriptors proposed in [24] were used. The reason for choosing different low-level image descriptors was that, in DS1 and DS2, images are actually snapshots of the corresponding 3D objects, where background and color information was not available, while 2D images in DS3, DS4 and DS5 are real images fetched from the Internet. Thus, for DS1 and DS2 the selected descriptors are based on shape, for DS3 and DS4 on background and color information, and for DS5 on color. The audio descriptors of DS2 and DS5 were extracted using the algorithm presented in [32]. For the text modality in DS4 and DS5, a lexicon was formed containing all the assigned labels. As a result, a text descriptor was formed as a vector of length equal to the lexicon’s size, filled with zeros and ones, to denote if the corresponding label was assigned to the respective MMD, respectively. Finally, for DS5, video descriptors were extracted using the color descriptors of [24] applied on the most representative keyframe of each video, following the work of [27].

#### *4.2. Experimental Organization*

The experiments were organized into four sets. In the first three sets of experiments each MMD of the database was posed as a query, in order to retrieve similar MMDs, where (a) the optimal parameters for the Laplacian Eigenmaps method were calculated, (b) the impact of nonlinear global (L-Isomap [26]) and local (Local Linear Embedding [25] and Laplacian Eigenmaps [5]) dimensionality reduction methods were evaluated, and (c) the proposed approach was compared against the state-of-the-art retrieval methods of LRGA [37] and SMMD [9]. In the last set of experiments, external MMD-queries were posed, so as to evaluate the performance of the proposed method for the real-life case, where a query does not belong to the database. Therefore, we present experimental results against the SMMD method, where also the case of external MMD-queries is handled, since to the best of our knowledge all the rest multimodal retrieval methods presented in the literature handle only the case of external monomodal queries. In all sets of experiments, the retrieval performance was evaluated in terms of precision-recall, where precision is the proportion of the retrieved MMDs that are relevant to the query and recall is the proportion of relevant MMDs in the entire database that are retrieved. In all set of experiments the average precision-recall is reported. The experiments were conducted in a Pentium 4 Quad Core machine with 3GHz, running Windows XP.

#### *4.3. Parameter Selection for Laplacian Eigenmaps*

Laplacian Eigenmaps (LE) is a Manifold Learning method that requires the specification of two input parameters: (a)  $k$ , the number of the nearest neighbors that are used to form the multimodal adjacency matrix, and (b)  $d$ ,

the number of dimensions of the low-dimensional multimodal MMD space. In Figures 5 and 6, the average precision-recall results are presented, by varying the  $k$  and  $d$  parameters in the ranges of  $\{3, 6, 9, 12\}$  and  $\{3, 9, 15, 25, 50, 100\}$ , respectively, following the parameter tuning of [9], for making fair comparisons on the evaluation datasets. Moreover, at this point we must mention that the values of the  $k$  and  $d$  parameters are limited to the aforementioned ranges, since the complexity of the proposed method is analogously increased along with the increase of the  $k$  and  $d$  parameters (for further details about the complexity of the proposed method refer to Section 4.7). The highest performance is achieved for  $k = 6$ , in all five datasets. For the  $d$  parameter, different values are proved to produce the optimal retrieval results. This can be explained by the fact that the  $d$  parameter, depends on the number of the different modalities and on the nature of their descriptor vectors. For large or small values of  $d$ , the retrieval accuracy is decreased, which is quite reasonable, concerning the way the LE method acts. In particular, the LE method tries to maintain the relationships among the  $k$  nearest neighbors, while at the same time unfolds the data points, so that they become more separable in the embedded space, using the Euclidean distance. Thus, larger values of  $d$  result in stretching the distances of the  $k$  nearest neighbors in the embedded space. On the contrary, lower values of  $d$  result in suppressing the distances in the embedded space, making the Euclidean distance incapable of discriminating the  $k$  nearest neighbors. Therefore, for DS1, DS2, DS3, DS4 and DS5 the optimal values of  $d$  were found to be equal to 9, 9, 15, 50 and 25, respectively.

Additionally, in Figure 7 we present the experimental results of comparing

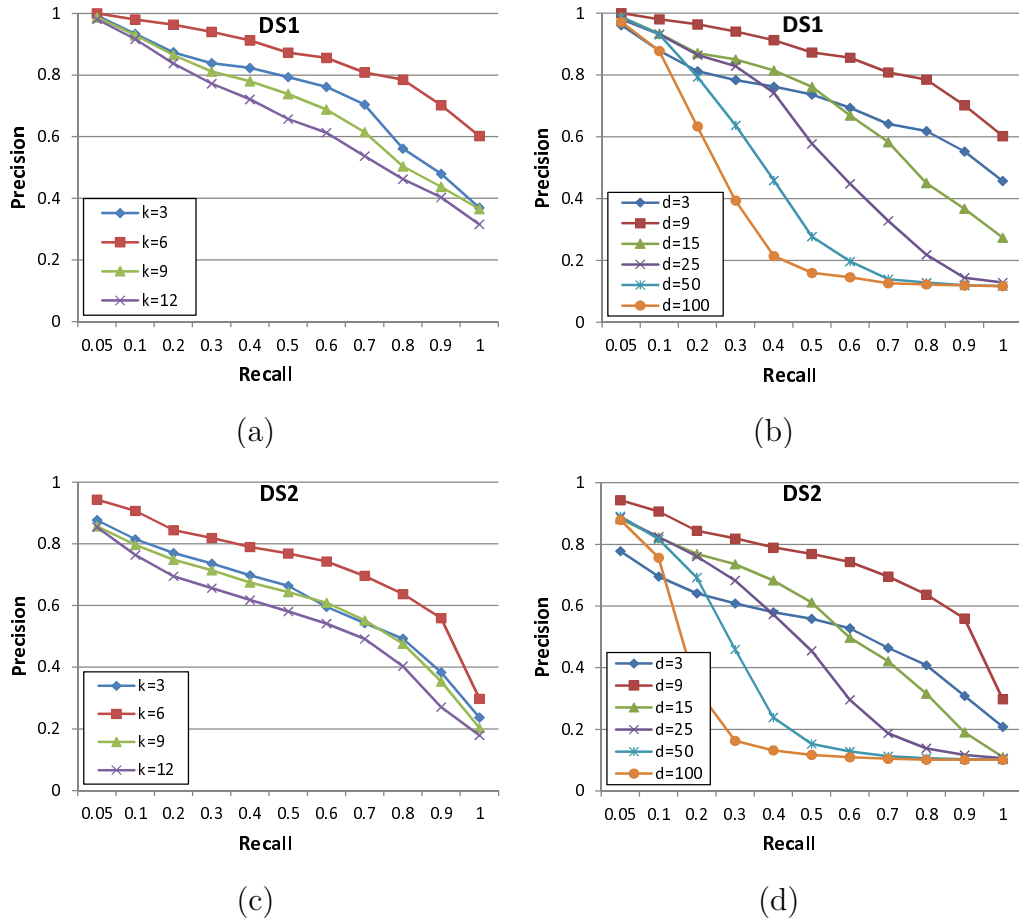


Figure 5: Precision-recall for DS1, DS2, by varying the  $k$  and  $d$  parameters in LE.

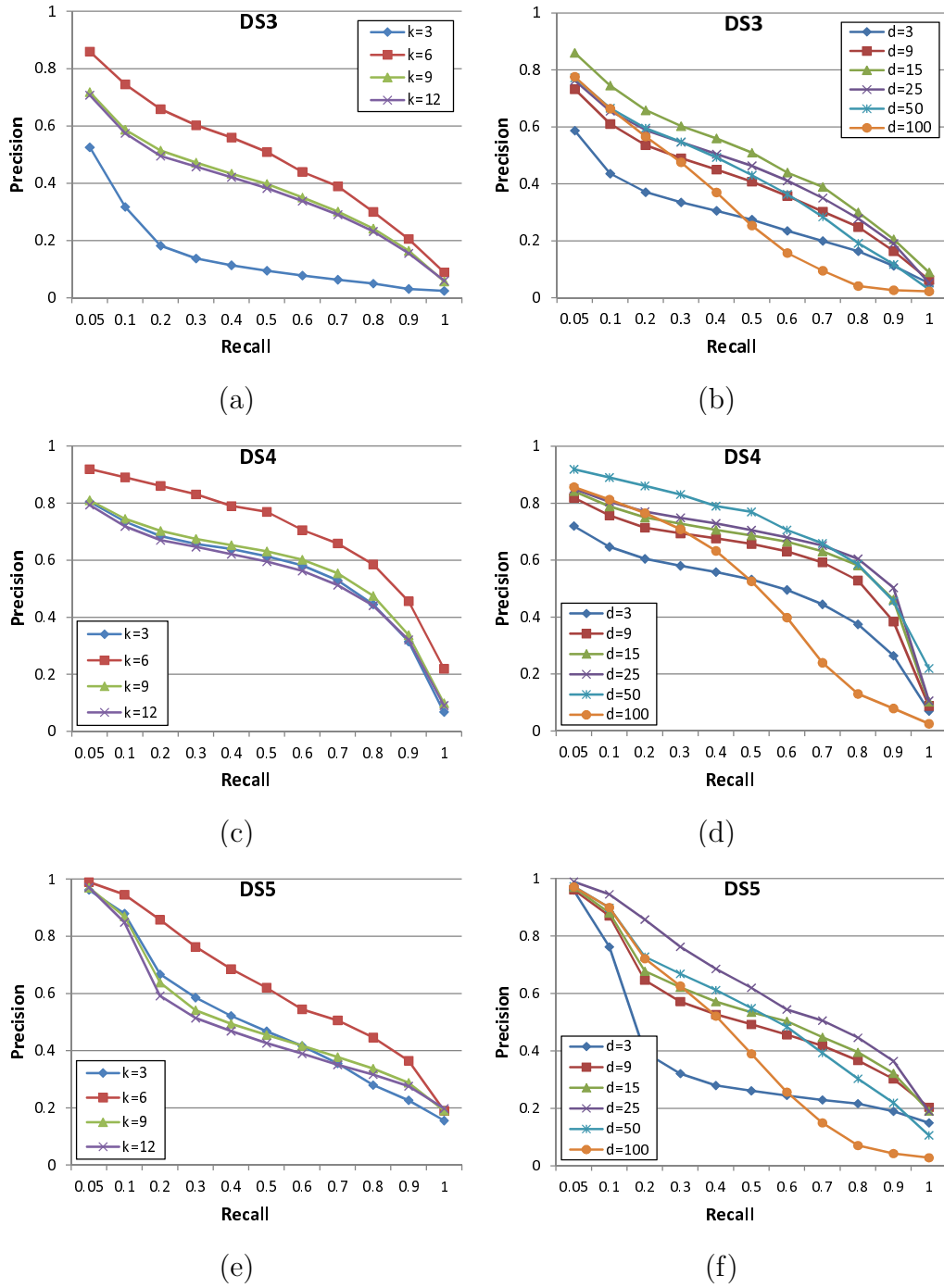


Figure 6: Precision-recall for DS3, DS4, DS5, by varying the  $k$  and  $d$  parameters in LE.



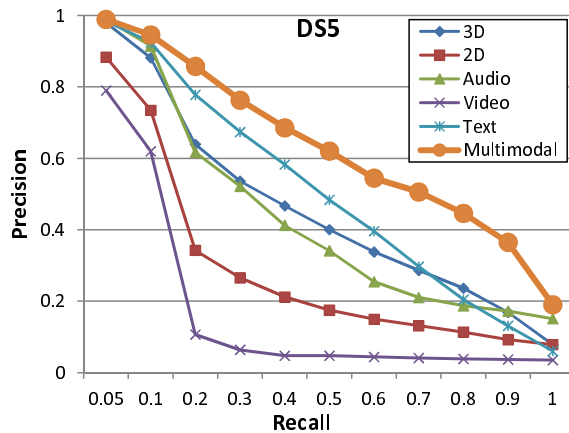


Figure 7: Precision-recall for DS5, comparing the performance of each modality content-based retrieval method with the proposed multimodal approach.

the retrieval accuracy of each modality’s content-based method to the proposed multimodal approach, in terms of average precision-recall. The evaluation is performed in DS5, which is a challenging dataset, since it consists of media of all 5 available modalities (Table 1). As expected, the retrieval accuracy of the proposed multimodal approach outperforms the retrieval accuracy of all monomodal content-based methods, because now the multimodal information is taken into account which is richer than each monomodal information separately. In Figure 8, we present an example of a MMD-query of dataset DS5, which describes a physical entity of a truck, by comparing the retrieval performance of the proposed multimodal method against the respective monomodal ones <sup>7</sup>.

---

<sup>7</sup>For presentation purposes the example consists of the 2D, 3D, video and text modalities.

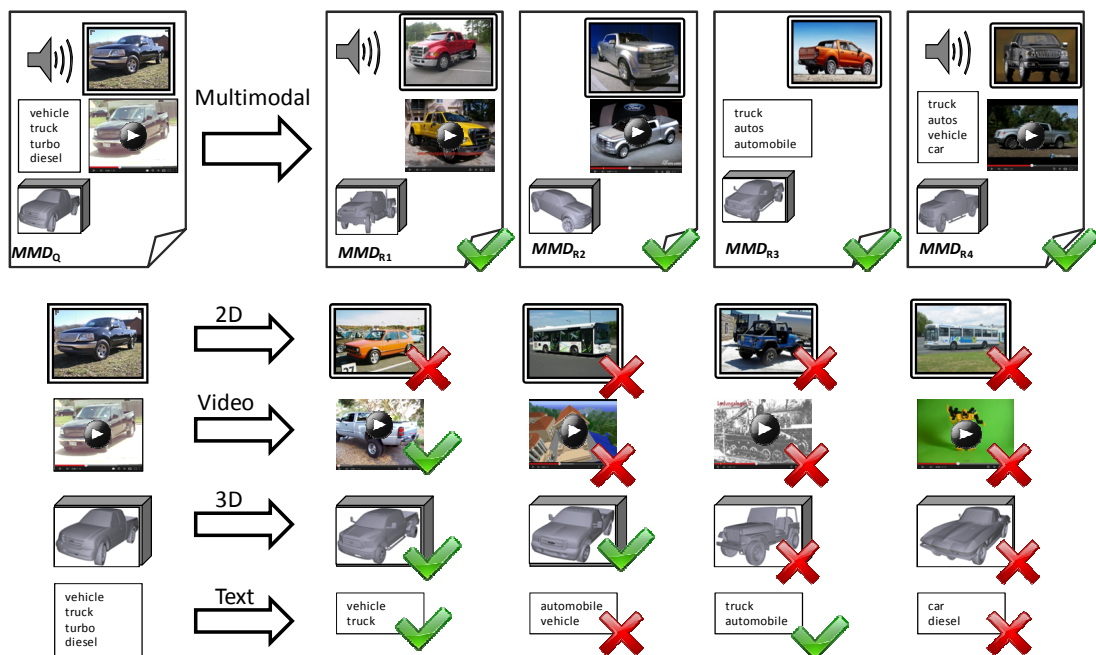


Figure 8: An example of a MMD-query of dataset DS5, which describes a physical entity of a truck, by comparing the retrieval performance of the proposed multimodal method against the respective monomodal ones.

#### 4.4. Comparison against state-of-the-art Nonlinear Dimensionality Reduction Methods

For the sake of completeness, LE was evaluated against two state-of-the-art nonlinear dimensionality reduction methods, the global Landmark Isomap (L-Isomap) [26] and the local approach of Locally Linear Embedding (LLE) [5]. L-Isomap is a global nonlinear technique, which tries to preserve the geodesic distances between all points and the landmarks, so as to maintain the data global structure. LLE and LE belong to the local nonlinear methods, where the data local structure is preserved. LLE embeds data points in a low dimensional space, by finding the optimal linear reconstruction in a small neighborhood, while LE restates the nonlinear mapping problem as an embedding problem for the vertices in a graph and uses the Laplacian graph to derive a smooth mapping. The input of the LE, L-Isomap and LLE methods is the multimodal similarity matrix,  $\mathbf{S}_{mult}$ , which is calculated according to Equation (4). As we can observe from Figure 9, LE outperforms L-Isomap and LLE, in all datasets. This happens because the LE method is able to preserve more efficiently the local neighborhood of each MMD in the embedded space, using the proposed global similarity matrix,  $\mathbf{S}_{mult}$ . This can be further explained by examining how the L-Isomap and the LLE methods work, where in contrast to LE the valuable information of  $\mathbf{S}_{mult}$  is not necessary preserved.

The global L-Isomap method finds the nearest neighbors according to  $\mathbf{S}_{mult}$  and then constructs a neighborhood graph, where each MMD is connected to each of its neighbors with an edge weighted by  $\mathbf{S}_{mult}$ . Then, it computes the shortest paths (geodesic distances) between all points and the

landmarks, so as to compute a new distance matrix  $\Delta$ . Next, the Multi-dimensional Scaling (MDS) method is performed to  $\Delta$ , so as to embed the data into the low-dimensional MMD space. However, the  $\Delta$  distance matrix contains the geodesic distances and thus, the information stored in  $\mathbf{S}_{mult}$  is not necessary preserved. Therefore, L-Isomap performs erroneous approximations of the geodesic distances and thus MMDs that lie close in the original metric space of  $\mathbf{S}_{mult}$  lie far in the MMD space, resulting in the low retrieval accuracy of L-Isomap.

The local method of LLE embeds MMDs in the low-dimensional space as a linear combination of their neighbors. LLE finds, for each MMD, the nearest neighbors based on  $\mathbf{S}_{mult}$ . Then, it computes a weight vector  $\vec{w}_x$  that best reconstructs MMD  $x$  by a linear combination of its nearest neighbors. Next, MMD  $x$  is embedded to a point  $y$  by minimizing the reconstruction error of  $y$  using  $\vec{w}_x$  and its corresponding nearest neighbors in the MMD space. However, the information of  $\mathbf{S}_{mult}$  is not necessary preserved in  $\vec{w}_x$  and thus, LLE fails to construct the MMD space accurately, especially for the large scale datasets of DS3 and DS4.

The LE method, described in Section 3.3, achieves high retrieval accuracy in all datasets, since it preserves the crucial information of  $\mathbf{S}_{mult}$  in the vertices of the Laplacian graph according to Equations (5), (6) and (7). Therefore, LE is able to map MMDs to the low-dimensional space more accurately than the L-Isomap and LLE methods. Consequently, LE preserves more efficiently the local neighborhood of each MMD in the embedded space, which furthermore results in LE’s high retrieval accuracy.

#### 4.5. Performance of the Proposed Method for Internal MMD-Queries

The proposed approach was also evaluated against the LRGA [37] and the SMMD [9] methods. The reason for selecting these methods is that both of them support multimodal MMD-queries, when those belong to the database. The respective results are presented in Figure 10, where it is shown that the proposed method outperforms the LRGA and SMMD methods in all cases. Following the proposed weighting strategy, our method is capable of preserving the local neighborhood of each MMD in the multimodal semantic space more accurately than LRGA and SMMD. This is confirmed by the high increase of the retrieval accuracy of the proposed method in DS5, where all 5 available modalities are involved.

#### 4.6. Performance of the Proposed Method for External MMD-Queries

In the final set of experiments, the proposed method was evaluated in the case of posing external MMD-queries. In particular, 12, 10, 100, 150, 43 external queries were posed for DS1, DS2, DS3, DS4 and DS5, respectively. The proposed approach was compared to SMMD, which also supports the case of posing external MMD-queries. For training the RBF network in the SMMD method, a number of  $CL$  predefined clusters are required. Therefore, through experimental configuration we concluded to 15, 15, 47, 50, 20 optimal values of  $CL$  clusters for DS1, DS2, DS3, DS4 and DS5, respectively. In Figure 11, the experimental results are depicted, where it is demonstrated that the proposed method achieves higher retrieval accuracy than SMMD. This happens because the SMMD method predicts the missing modalities of the external MMD-query based on the trained RBF network, whereas the proposed method actually projects the external MMD-query into the

already constructed multimodal semantic space. Note that high increase of the retrieval accuracy of the proposed method is achieved, especially for the case of (a) the large scale dataset DS4 compared to the achieved retrieval accuracy in DS3, where in the latter the text modality misses and of (b) DS5, where all 5 modalities are included. Therefore, we can conclude that the proposed method, in case of posing external MMD-queries, is capable of achieving high retrieval accuracy, along with the increase of the number of available modalities and the size of the database. This is of high importance, if we consider that the case of external MMD-queries reflects on the real-life scenario, where MMD-queries usually do not belong to the database.

#### 4.7. Computational Issues

In terms of computational efficiency, offline and online processing times are reported for SMMD and the proposed method, since both are able to perform multimodal search and retrieval either posing internal or external MMD-queries. According to the experimental results depicted in Table 2, we observe that the proposed method requires slightly more time than SMMD for constructing the multimodal semantic space, because the time of computing the multimodal similarity matrix  $\mathbf{S}_{mult}$  has to be added <sup>8</sup>, before the LE method is applied. Nevertheless, the proposed method requires less offline preprocessing time than SMMD in total, since the former does not include the clustering and the RBF training steps, as it happens in the case of the

---

<sup>8</sup>The computational times for calculating the monomodal similarities are omitted, since they are common for both methods. However, both methods can avoid calculating all-to-all similarities, by following indexing strategies for efficient similarity search [13].

latter in order to support the case of external queries. Additionally, it should be noted that both methods’ offline times increase with respect to the number of the included modalities and the dataset size. For example, although DS5 and DS2 are at the same dataset scale, higher time is required for DS5, since it consists of more modalities than DS2 (5 instead of 3 modalities). Moreover, for the SMMD method, clustering needs a proportional time to the size of MMDs, and the time required for training the RBF network becomes prohibitive along with the increase of either the database size (DS3, DS4) or the number of modalities (DS5).

	<b>DS1</b>	<b>DS2</b>	<b>DS3</b>	<b>DS4</b>	<b>DS5</b>
<b>SMMD</b>					
Const. mult. space	814	1,946	49 906	104 559	43 288
Clustering	15.6	94.7	432.9	1 638	123
RBF train.	1 010	14 910	14 977 740	44 694 820	529 984
<b>Prop. method</b>					
Const. mult. space	1 812	7 019	132 041	217 165	48 785

Table 2: Time of offline processing (msec)

In Table 3, we report the online computational times which correspond to the case of multimodal search and retrieval. As shown, both methods have equal times in case of posing internal MMD-queries, by following the same retrieval strategy. In particular, when the MMD-query  $Q$  belongs to the database, its multimodal descriptor vector  $\mathbf{y}_Q$  already exists, since it has already been mapped to the  $d$ -dimensional multimodal space according to (7). Thus,  $\mathbf{y}_Q$  is compared to multimodal descriptors  $\mathbf{y}_i$  ( $1 \leq i \leq N$ ) of the MMDs in the database, using the Euclidean distance, in order to retrieve

the most similar MMDs to  $Q$ . Since through experimental evaluation we concluded to the same number of  $d$  dimensions in the constructed semantic space for SMMD and in the proposed method, the retrieval time for both methods is proportional to the dataset size and therefore, equal in the case of internal queries. In the case of external MMD-queries, SMMD performance highly depends on the complexity of the RBF network, which is used to classify the monomodal descriptors of the external query to predict the center of the cluster that is closer to the MMD-query. Consequently, the whole performance of SMMD depends on the number of the  $CL$  predefined clusters and the number of modalities, which are included in the dataset. On the other hand, the proposed method requires  $O(M \cdot N)$  time to calculate  $M$  monomodal similarities and  $O((k + 1)^3)$  to map the external MMD-query to the already constructed semantic space, according to the analysis of [15]. Therefore, since it holds that  $N \gg k$ , the online retrieval complexity of the proposed method is transformed to  $O(M \cdot N)$ .

	<b>DS1</b>	<b>DS2</b>	<b>DS3</b>	<b>DS4</b>	<b>DS5</b>
<b>Int. MMD-Query</b>					
SMMD	0.28	0.55	2.89	3.71	0.85
Prop. method	0.28	0.55	2.89	3.71	0.85
<b>Ext. MMD-Query</b>					
SMMD	14.51	30.06	91.7	99.8	75.1
Prop. method	29.3	45.5	79.6	104.9	75.9

Table 3: Time of online processing (msec)



## 5. Conclusions

In this paper, a unified framework for multimodal content-based search and retrieval is presented, which supports internal and external MMD-queries. Following an innovative weighting strategy, all monomodal heterogeneous similarities are combined to a global MMD similarity, by considering (a) the different nature of the monomodal similarities and (b) the availability of modalities per MMD in the database. As a result, the local neighborhood of each media modality is preserved into the multimodal semantic space of MMDs, which is built using the Laplacian Eigenmaps method. Thus, within the constructed semantic space, internal MMD-queries are posed, in order to perform multimodal content retrieval. Additionally, the proposed method is able to handle external MMD-queries, by embedding them into the already constructed MMD space. In our experiments with five real multimodal datasets of different scale, we showed the superiority of the proposed approach against other state-of-the-art multimodal methods, in terms of retrieval accuracy and computational efficiency. Additionally, we experimentally demonstrated that the proposed method, especially in the case of posing external MMD-queries, is capable of achieving high retrieval accuracy, along with the increase of the number of available modalities and the size of the database.

## Acknowledgment

This work was supported by the EC project 3DVIVANT, GA-248420.

## References

- [1] E. Attalla, P. Siy, “Robust shape similarity retrieval based on contour segmentation polygonal multiresolution and elastic matching”, *Pattern Recognition*, 38(12), pp. 2229-2241, 2005.
- [2] J.J. Aucouturier, F. Pachet, “A scale free distribution of false positives for a large class of audio similarity measures”, *Pattern Recognition*, 41(1), pp. 272-284, 2008.
- [3] A. Axenopoulos, P. Daras, S. Malassiotis, V. Croce, M. Lazzaro, J. Etzold, P. Grimm, A. Massari, A. Camurri, T. Steiner, D. Tzovaras, “I-SEARCH: A Unified Framework for Multimodal Search and Retrieval”, *Future Internet Assembly 2012: From Technological Promises to Reality (FIA Book 2012)*, *Lecture Notes in Computer Science*, Volume 7281, pp 130-141, 2012.
- [4] M. Balasubramanian, E.L. Schwartz, J.B. Tenenbaum, V. de Silva, J.C. Langford, “The ISOMAP algorithm and topological stability”, *Science* 295, 7, 2002.
- [5] M. Belkin, P. Niyogi, “Laplacian eigenmaps for dimensionality reduction and data representation”, *Neural Comput.* 15(6), pp. 1373-1396, 2003.
- [6] R. Bellman, *Adaptive Control Processes: A Guided Tour*, Princeton University Press, 1961.
- [7] S. A. Chatzichristofis, Y. S. Boutalis, “CEDD: Color and Edge Directivity Descriptor - A Compact Descriptor for Image Indexing and Retrieval”,

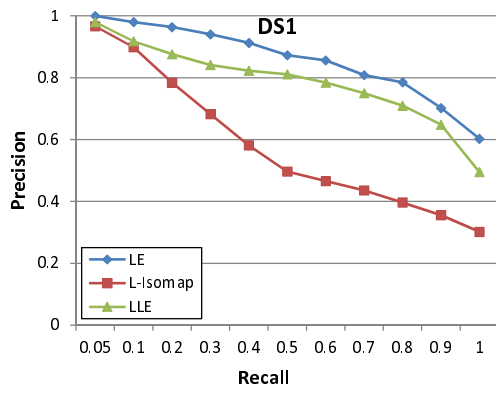
- Proc. of Int. Conf. on advanced research on Computer Vision Systems, pp. 312-322, 2008.
- [8] P. Daras, A. Axenopoulos, “A 3D Shape Retrieval Framework Supporting Multimodal Queries”, *Int. Journal of Computer Vision*, DOI 10.1007/s11263-009-0277-2, 2009.
- [9] P. Daras, S. Manolopoulou, A. Axenopoulos, “Search and Retrieval of Rich Media Objects Supporting Multiple Multimodal Queries”, *IEEE Trans. on Multimedia*, 4(3), pp. 734-746, 2012.
- [10] P. Ehlen and M. Johnston, ”Location grounding in multimodal local search”, *Proc. of ICMI-MLMI*, pp. 32:1-32:4, 2010.
- [11] T. Gao, Q. Dai, N.Y. Zhang, “3D model comparison using spatial structure circular descriptor”, *Pattern Recognition*, 43(3), pp. 1142–1151, 2010.
- [12] P. Geetha, V. Narayanan, “A Survey of Content-Based Video Retrieval”, *Journal of Computer Science* 4(6), pp. 474-486, 2008.
- [13] C. Gennaro, G. Amato, P. Bolettieri and P. Savino, “An approach to content-based image retrieval based on the Lucene search engine library”, *Proc. of European Conf. on Research and advanced technology for digital libraries*, 2010.
- [14] T. Hofmann, B. Schölkopf, and A. J. Smola, “Kernel methods in machine learning”, *Annals of Statistics* 36(3), pp. 1171-1220, 2008.

- [15] Peng Jia , Junsong Yin , Xinsheng Huang , Dewen Hu, “Incremental Laplacian eigenmaps by preserving adjacent information between data points”, *Pattern Recognition Letters*, 30(16), pp. 1457-1463, 2009.
- [16] L. Kennedy, S.-F. Chang, and A. Natsev, “Query-adaptive fusion for multimodal search”, *Proc. IEEE*, 96(4), pp. 567-588, 2008.
- [17] P. L. Lai, C. Fyfe, “Canonical correlation analysis using artificial neural networks”, *Proc. of European Symposium on Artificial Neural Networks*, 1998.
- [18] D. Li, N. Dimitrova, M. Li, I. K. Sethi, “Multimedia Content Processing through Cross-Modal Association”, *Proc. of ACM Int. Conf. on Multimedia*, 2003.
- [19] A. Mademlis, P. Daras, D. Tzovaras, M. G. Strintzis, “3D Object Retrieval using the 3D Shape Impact Descriptor”, *Pattern Recognition*, 42(11), pp. 2447-2459, 2009.
- [20] R. Ohbuchi, J. Kobayashi, “Unsupervised Learning from a Corpus for Shape-Based 3D Model Retrieval”, *ACM MIR*, 2006.
- [21] D. Rafailidis, A. Nanopoulos, and Y. Manopoloulos, “Nonlinear Dimensionality Reduction for Efficient and Effective Audio Similarity Searching”, *Multimedia Tools and Applications*, 51(3), pp. 881-895, 2011.
- [22] S.T. Roweis, L.K. Saul, “Nonlinear dimensionality reduction by locally linear embedding”, *Science* 290, pp. 2323-2326, 2000.

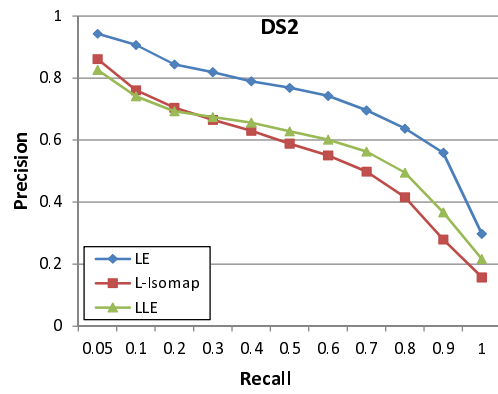
- [23] A. Ruta, Y. Li, X. Liu, “Real-time traffic sign recognition from video by class-specific discriminative features”, *Pattern Recognition*, 43(1), pp. 416-430, 2010.
- [24] K.E.A. van de Sande, T. Gevers, and C.G.M. Snoek, “Evaluating Color Descriptors for Object and Scene Recognition”, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32(9), pp. 1582-1596, 2010.
- [25] L.K. Saul, S. T. Roweis, “Think Globally, Fit Locally: Unsupervised Learning of Low Dimensional Manifolds”, *Journal of Machine Learning Research*, 2003.
- [26] V. de Silva, J. B. Tenenbaum, “Global versus local methods in nonlinear dimensionality reduction”, *Neural Information Processing Systems*, 15, pp. 705-712, 2003.
- [27] E. Tsamoura, V. Mezaris, I. Kompatsiaris, “Gradual transition detection using color coherence and other criteria in a video shot meta-segmentation framework”, *Proc. of ICIP 2008*, pp. 45-48, 2008.
- [28] D. Vranic, “3d model retrieval” Ph.D. Dissertation, University of Leipzig, 2004.
- [29] B. Wang, F. Pan, K.M. Hu, J.C. Paul, “Manifold-ranking based retrieval using k-regular nearest neighbor graph”, *Pattern Recognition*, 45(4), pp. 1569-1577, 2012.
- [30] X. Y. Wang, P.P. Niu, H.Y. Yang, “A robust digital audio watermarking based on statistics characteristics”, *Pattern Recognition*, 42(11), pp. 3057-3064, 2009.

- [31] C.H. Wei, Y. Li, W.Y. Chau, C.T Li, “Trademark image retrieval using syntetic features for describing global shape and interior structure”, *Pattern Recognition*, 42(3), pp. 386-394, 2009.
- [32] Wichern, Xue, Thornburg, Mechteley, Spanias: “Segmentation, Indexing, and Retrieval for Environmental and Natural Sounds”, *IEEE Trans. on Audio, Speech and Language Processing*, 2010.
- [33] F. Wu, H. Zhang, Y. Zhuang, “Learning Semantic Correlations for Cross-Media Retrieval”, *Proc. of IEEE Int. Conf. on Image Processing*, 2006.
- [34] X. Xie, L. Lu, M. L. Jia, H. Li, F. Seide, W.Y. Ma. “Mobile Search with Multimodal Queries”, *Proc. of IEEE*, 96(4), pp. 589-601, 2008.
- [35] X. Yang, S. Pang, and K. Cheng, “Mobile image search with multimodal context-aware queries”, *Proc. of Int. Workshop Mobile Vision*, 2010.
- [36] Y. Yang, F. Wu, D. Xu, Y. Zhuang, L.T. Chia, “Cross-media retrieval using query dependent search methods”, *Pattern Recognition* 43(8), pp. 2927-2936, 2010.
- [37] Y. Yang, D. Xu, F. Nie, J. Luo and Y. Zhuang, “Ranking with Local Regression and Global Alignment for Cross Media Retrieval”, *Proc. of ACM Int. Conf. on Multimedia*, pp. 175-184, 2009.
- [38] H. Zhang, F. Meng, “Multi-modal Correlation Modeling and Ranking for Retrieval”, *Advances in Multimedia Information Processing - PCM*, pp. 637-646, 2009.

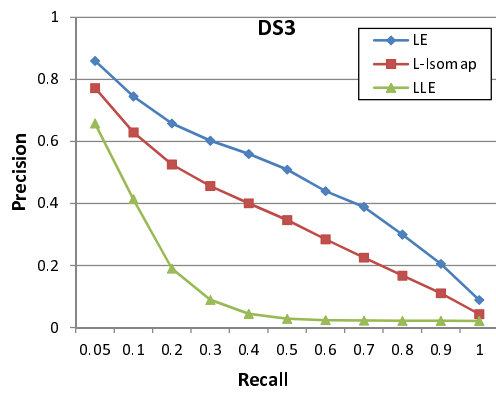
- [39] H. Zhang, J. Weng, “Measuring Multi-modality Similarities Via Subspace Learning for Cross-Media Retrieval”, *Advances in Multimedia Information Processing - PCM*, pp. 979-998, 2006.



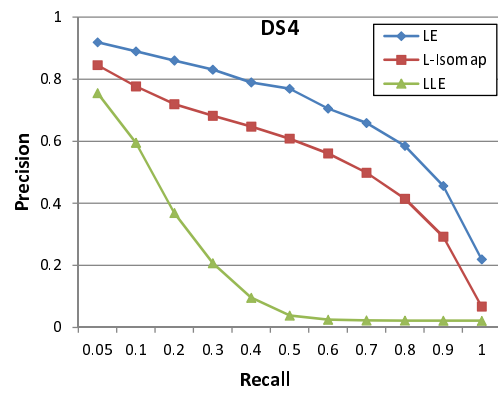
(a)



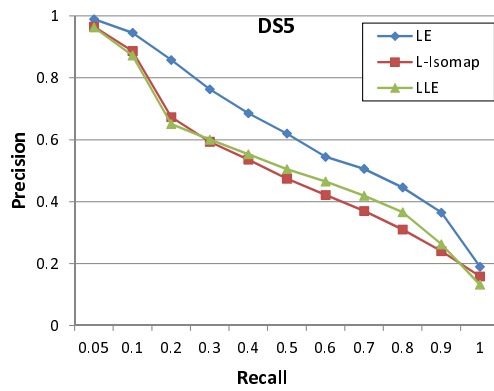
(b)



(c)



(d)



(e)

Figure 9: Precision-recall, comparing the nonlinear dimensionality reduction methods of LE, L-Isomap and LLE.



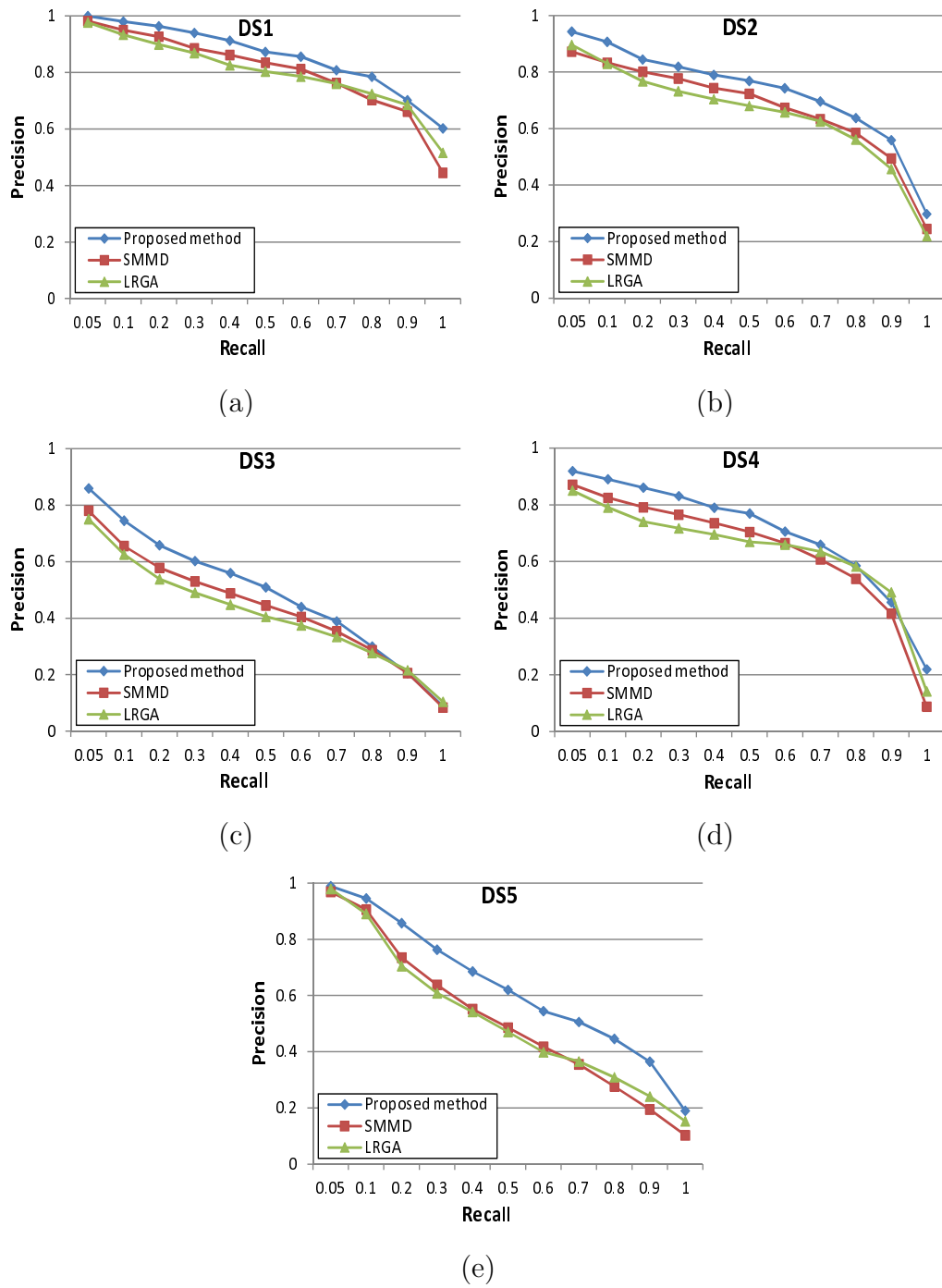
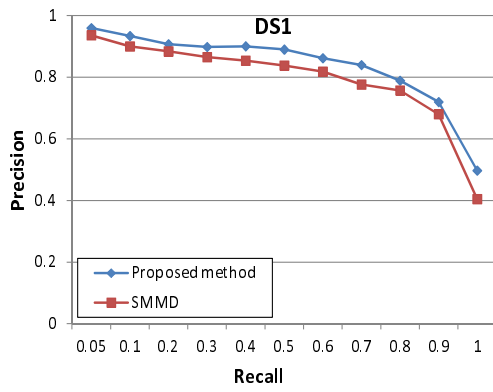
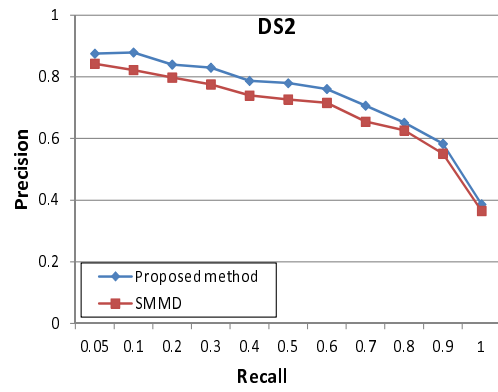


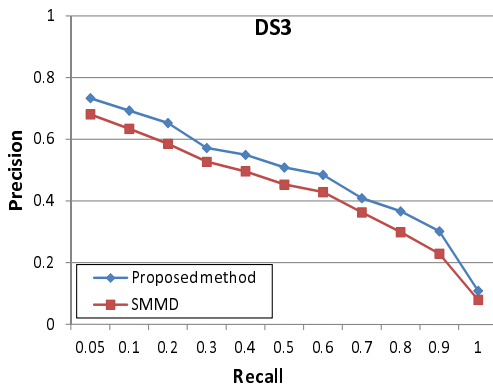
Figure 10: Precision-recall, comparing the proposed method with LRGA [37] and SMMD [9], for the case of internal MMD-queries.



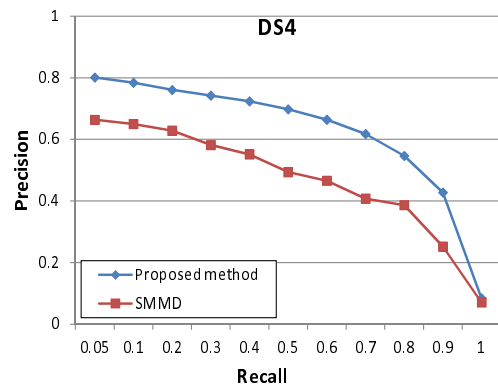
(a)



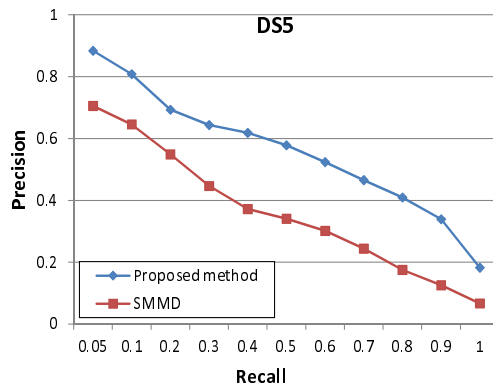
(b)



(c)



(d)



(e)

Figure 11: Precision-recall, comparing the proposed method with SMMD [9] for the case of external MMD-queries.