Multimodal Search and Retrieval using Manifold Learning and Query Formulation

Apostolos Axenopoulos\textsuperscript{*} \hspace{1cm} Stavroula Manolopoulou\textsuperscript{†} \hspace{1cm} Petros Daras\textsuperscript{‡}
Informatics and Telematics Institute \hspace{1cm} Informatics and Telematics Institute \hspace{1cm} Informatics and Telematics Institute
Centre for Research and Technology Hellas \hspace{1cm} Centre for Research and Technology Hellas

Abstract

In this paper a novel approach for multimodal search and retrieval is introduced. The searchable items are media representations consisting of multiple modalities, such as 2D images and 3D objects, which share a common semantic concept. The proposed method combines the low-level feature distances of each separate modality to construct a new low-dimensional feature space, where all media objects are 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.


Keywords: Multimedia description, 2D/3D content descriptor extraction, multimodal search and retrieval.

1 Introduction

Due to the widespread availability of multimedia content over the Internet, the need for methods for effective search and retrieval of this content emerged. Towards this direction, several approaches for content-based multimedia retrieval have been introduced. These methods exploit the low-level features, which are automatically extracted from media objects, in order to retrieve semantically similar objects. Most of the methods are specialised to deal with only one single modality, namely 3D objects [VANAMALI 2010], images [HE 2004], video [GEEITHA 2008] or audio [WAN 2005]. However, the emerging demands of Internet users asking for multimedia types simultaneously brought up a new generation of content-based methods, known as Cross-media Retrieval ([YANG 2009] [TONG 2005] [ZHANG 2006] [WU 2006]) or Multimodal Retrieval. These methods allow users to enter multimodal queries and retrieve multiple types of media simultaneously. Such an approach is presented in this paper, introducing several novel characteristics that will further be analyzed.

\textsuperscript{*}e-mail: axenop@iti.gr
\textsuperscript{†}e-mail: manolop@iti.gr
\textsuperscript{‡}e-mail: daras@iti.gr

1.1 Related Work

Dealing with multimedia data involves confronting the obstacle of high dimensionality. Thus, it is often assumed that the data lie on an embedded non-linear manifold within the higher-dimensional space. By properly unfolding this manifold structure, a more representative feature space of lower dimension is achieved. \textit{Manifold learning}, which has been successfully adopted by many single-modality content-based retrieval methods ([OHBUCHI 2006] [VANAMALI 2010] [HE 2004]), can be easily extended to address cross-modal and multimodal retrieval problems. The most representative attempts in this field are given in the sequel.

Yang et al. [YANG 2009] proposed a cross-media retrieval method, based on a structure called Multimedia Document (MMD). The paper introduced the concept of MMD distance, which intuitively merges the dissimilarities between different modalities as a weighted sum of their mono-modal distances. The method applies Multidimensional Scaling 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 Local Regression and Global Alignment (LRGA). Similarly in [ZHANG 2006], Zhang et al. investigated the intra- and inter-media correlations to build a map from heterogeneous multi-modality (text, image, audio) feature spaces, called "multimedia bags", into a semantic subspace created using Laplacian Eigenmaps, called Multi-modality Laplacian Eigenmaps Semantic Subspace (MLESS). Furthermore, Wu et al. [WU 2006] proposed a method which constructs an isomorphic multimodal subspace based on Canonical Correlation Analysis (CCA), where a general distance function using polar coordinates is defined.

Manifold ranking has been also used to improve the retrieval performance of methods that deal with mono-modal data [ZHOU 2003]. In [OHBUCHI 2006], Ohbuchi et al. created a lower dimensional manifold using Laplacian Eigenmaps (LE) to enhance the discriminative power of shape features extracted from 3D models. Then, a Radial Basis Function (RBF) network was used to approximate the manifold in order to map there a query not belonging to the training set. Another dimensionality reduction algorithm, similar to LE, is Locally Linear Embedding (LLE) [SAUL 2003]. He at al. in [HE 2004] introduced a Manifold-Ranking Based Image Retrieval (MRBIR) scheme using a weighted graph. Through iterations, each data point gains a ranking score according to the neighboring nodes. A modified version of the previous approach, the “Modified Manifold Ranking (MMR)”
algorithm [VANAMALI 2010] was proposed for the improvement of the retrieval performance in a 3D object retrieval method. Inspired by the previous manifold ranking application on mono-modal data, the authors of [TONG 2005] tried to extend this graph-based semi-supervised learning to multi-modal data, by creating an independent graph for each modality.

1.2 Method Overview

In multimodal search and retrieval, the searchable items are rich media representations, which enclose multiple media types of the same semantics into a media container. The supported modalities can be 2D images, 3D objects, audio files or video. In this paper, we borrowed the term Multimedia Document (MMD) [YANG 2009] to refer to these media containers (we could either choose the term Multimedia Bag [ZHANG 2006]). An MMD can span from very simple media items (e.g. a single image or an audio file) to highly complex multimedia collections (e.g. a 3D object together with multiple 2D images and audio files) along with accompanying information (Figure 1). When a user refers to an MMD, s/he directly refers to all of its constituting parts. For simplicity, the modalities used for our experiments are 3D objects and 2D images, although the proposed framework can be easily extended to include other media types as well.

The proposed method consists of two consecutive processes. During the first process, which is depicted in Figure 2, the construction of the Multimodal Feature Space takes place. More specifically, given as input a database of MMDs the mono-modal descriptors are extracted from each modality (2D image or 3D object). These descriptors are used to construct an adjacency matrix. By applying Laplacian Eigenmaps to the adjacency matrix a new multimodal feature space of lower dimensionality is created, were every MMD is mapped to a data point.

The second process is presented in Figure 3 and involves the multimodal search and retrieval. During the training phase, a Radial Basis Function (RBF) network is constructed using a predefined dataset of MMDs. The constructed 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 (testing), a query MMD, which does not belong to the database is transformed to a multimodal descriptor vector, using the RBF function. This multimodal descriptor vector is directly matched with the multimodal descriptors of the database MMDs and the most similar MMDs are retrieved.

Comparing with similar cross-media retrieval methods, the paper introduces two clear innovations: the first is that the non-zero elements of the adjacency matrix, which correspond to pairs of neighbouring MMDs, are not MMD distances 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 neighbours of each modality are assigned the same non-zero value to construct the adjacency matrix. The second innovative feature of this work is that it supports multimodal queries. In already existing cross-media retrieval methods [YANG 2009][ZHANG 2006], the query of one single modality is used to retrieve items of different modalities, but using two or more modalities simultaneously as input is not supported. In this paper, multimodal querying is fully supported, allowing users to enter as query an entire MMD. 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.

Multimodal search could be very useful in many application areas. Regarding 3D object retrieval, for instance, a query 3D object may not always be available. Through multimodal search, the user is able to retrieve 3D objects by using as queries different input modalities, such as images/photos, hand-drawn sketches or even sounds. This makes multimedia search more personalised and adapted to the user preferences.

The rest of the paper is organized as follows. In section 2, the construction of the Multimodal Feature Space is analyzed. In section 3, the multimodal search and retrieval procedure is described and in the following section the experimental results are presented. Finally, conclusions are drawn in section 5.

2 Construction of Multimodal Feature Space

During the construction of the multimodal feature space, all MMDs of a database, irrespective of their constituting modalities, are represented as d-dimensional points in a new feature space. In this new feature space, semantically similar MMDs will lie close to each other with respect to a common distance metric (such as the L-2 distance). The methodology, which is usually followed, is known as manifold learning, where it is assumed that the multimodal data lie on a non-linear low-dimensional manifold. The majority of manifold learning approaches are based on the computation of the k-nearest neighbours among all items of the dataset in order to construct an adjacency matrix. In our case, the items of the dataset are MMDs. The k-nearest neighbour computation for an MMD is not trivial, though, since it requires merging descriptors of heterogeneous modalities into one unified distance metric. Although such unified distances have been already presented to address cross-media retrieval [YANG 2009][ZHANG 2006], they suffer from several limitations. The main limitation is that since these distances are weighted sums of mono-modal distances, the weights are highly dependent on the discriminative power of each separate mono-modal descriptor, which makes the distance measure unreliable, especially when more than two modalities are merged.

To address the above limitation, an alternative approach based on Laplacian Eigenmaps (LE) is presented in this paper. The reason for choosing LE is that they support input adjacency matrices with only zeros/ones instead of accurate distances, i.e. when items i, j are neighbours, the item W_{ij} of the adjacency matrix is assigned the value 1 instead of the distance between i and j. Since the items of the adjacency matrix are MMDs, the neighbourhood criterion is determined as follows: two MMDs, i and j are neighbours if and only if at least one pair of their constituting items of the same modality are neighbours. If the two MMDs do not have items of common modality they are not considered as neighbours. Neighbourhood 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.

2.1 Computing Nearest Neighbours within MMDs

In this step, a more detailed description of the creation of Multimodal Adjacency Matrix is presented. Given a multimedia database of N MMDs and p different modalities, the goal is to compute the k-nearest neighbours for every MMD, 1≤i≤N. For simplicity, we assume that each MMD, consists of exactly one item per modality, although it is possible to have only few modalities in MMD, as well as more than one items of the same modality.

In order to compute the k-nearest neighbours of MMD, the nearest neighbours of each separate modality need to be determined. Let a media item within MMD of m-th modality (1≤m≤P) that is represented by the descriptor vector \( x_i^m \). For the m-th modality, a distance measure is defined as \( d^m(x_i^m,x_j^m) \) to calculate the mono-modal similarities. The \( k_m \)-nearest neighbours of \( x_i^m \) are retrieved by ranking all the media items of m-th modality (\( x_j^m \)) within the database, with respect to their mono-modal distances \( d^m \). The ranked list of \( k_m \)-nearest neighbours of \( x_i^m \) is defined as:

\[
\text{NeighList}_{i}^m = \text{index}_{i1}^m, \text{index}_{i2}^m, \ldots, \text{index}_{ik_m}^m
\]

(1)

where \( \text{index}_{ik_m}^m \) is the index of the MMD which corresponds to the media item of m-th modality, ranked as the first nearest neighbour of \( x_i^m \). \( \text{index}_{i1}^m, \text{index}_{i2}^m, \ldots, \text{index}_{ik_m}^m \) are the indices of the MMDs corresponding to the 2\textsuperscript{nd}, ..., \( k_m \textsuperscript{th} \) ranked items, respectively. Similarly, \( p \) lists of nearest neighbours are extracted, one for each modality.

The final k-nearest neighbours of MMD, are computed by taking equal number of first neighbours from each list \( \text{NeighList}_{i}^m \), \( 1≤m≤P \), i.e. \( k/p \) neighbours, with \( (k/p)<k_m \). In case an MMD, appears in the \( k/p \) neighbours of more than one lists \( \text{NeighList}_{i}^m \), this MMD is counted only once. The remaining positions in the k-nearest neighbours list are then filled with the next closest MMDs.

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

Finally, a Nxk matrix, \( \text{NN}_{N,MMD} \), is created, where each row \( i \) represents the k-nearest neighbours of MMD,.
2.2 Laplacian Eigenmaps

Using the $NN_{nn}$ matrix, the Laplacian Eigenmaps create a multimodal feature space of low dimension, where every MMD is represented as a $d$-dimensional vector, that is the $N \times d$ matrix $Y$. The LE algorithm consists of the following steps:

1. Construct the graph $G$, by connecting nodes (i.e. MMDs) $i$ and $j$ with an edge, if $j$ is among the $k$-nearest neighbors of $i$.
2. Produce the $N \times N$ adjacency matrix, $W$, of $G$:
   \[ W_{i}^{j} = \begin{cases} 
   1, & \text{if MMD } j \text{ belongs to } k \text{-nearest neighbors of MMD } i \\
   0, & \text{otherwise} \end{cases} \]  
   (6)
3. Create a $N \times N$ diagonal matrix $H$: $H_{ii} = \sum_{j} W_{i}^{j}$.
4. Create a $N \times N$ Laplacian matrix: $L = H - W$.
5. Solve the generalized eigenproblem $LY = \lambda HY$ to find the eigenvalues $\lambda$ and the eigenvectors $Y$ of $L$.
6. Sort eigenvalues in an ascending order and keep the least $d$ corresponding eigenvectors (excluding the first one).

The $d$ selected eigenvectors correspond to $d$-dimensions of the new multimodal feature space, where all database MMDs are mapped. 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 motivation for choosing binary values (1, 0) to construct $W$ (instead of using actual distances) was to overcome the heterogeneity of descriptors from multiple modalities. Different descriptors require different distance metrics that cannot be put together in the same equations.

3 Multimodal Search and Retrieval

In the existing cross-media retrieval approaches, a query of a single modality is used to retrieve objects of different modalities. However, to the best of our knowledge, none of the existing methods supports queries of multiple modalities simultaneously. The framework proposed in this paper allows multimodal queries (the query can be an MMD with multiple 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, computed with the method described in the previous section, is matched against the multimodal descriptor vectors of the rest MMDs of the database and the most relevant results are retrieved.

In case the query MMD does not belong to the database (Figure 4), its multimodal descriptor vector is not available. Therefore, it cannot be directly matched with the descriptor vectors of the database MMDs. A solution to this problem would be to compute the query descriptor vector using machine learning algorithms. In this work, an appropriately trained Radial Basis Function (RBF) network [Bors 2001] was used to map the initial mono-modal descriptors of a query MMD to the low-dimensional multimodal feature space.

The selected RBF network is a multiple-input – multiple-output network, where the multiple inputs are the mono-modal descriptors of the MMD’s constituting modalities and the $d$ outputs correspond to the $d$-dimensions of the multimodal descriptor vector. If $D_{m}$ is the dimension of the $m$-th modality’s descriptor vector ($m = 1,...,p$), the total number of inputs to the RBF is:

\[ D = \sum_{m=1}^{p} D_{m}. \]  

During training, the $D$ descriptors of all database MMDs are given as input to the RBF taking as output their corresponding $d$-dimensional descriptors. In the retrieval phase, the query MMD enters the RBF using its $D$ descriptors as input. Then, the RBF function predicts its $d$-dimensional multimodal descriptor vector. The latter is used for similarity matching with the MMDs within the database.

An interesting property of RBF networks is that they support missing inputs. That is the performance of an RBF is not affected when multiple inputs are empty. Thus, even if one or more modalities are missing from several database MMDs 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.

![Figure 4: Multimodal retrieval using an MMD that does not belong to the database.](image)

4 Experimental Results

The proposed method was experimentally evaluated on a database of 264 MMDs. This database was compiled by the authors and it consists of 3D objects and 2D images. The images are snapshots of the 3D objects, either black/white silhouettes or depth images. MMDs are divided in 12
categories: animals (27), spheroid objects (17), conventional airplanes (64), delta airplanes (20), helicopters (20), cars (20), motorcycles (12), tubes (10), couches (14), chairs (20), fish (20) and humans (20). The total number of 3D objects is 159 and of 2D images 312. The 3D object descriptors were extracted using the Spherical Trace Transform (STT). A detailed description of the algorithm is available at [ZARPALAS 2006]. The dimension of STT is 1080. The image descriptors consist of 2D Polar-Fourier coefficients, Zernike moments and Krawtchouk moments and form a descriptor vector of size 212. A detailed description of the algorithm is available at [DARAS 2009].

To compute mono-modal dissimilarities among 2D image descriptors and 3D objects the L-2 distance was used. The values of several key parameters were also determined experimentally. Regarding the $k$-nearest neighbours of MMDs for the construction of the adjacency matrix, it was found that for $k=6$, the method achieved the maximum performance in the given dataset. Similarly, for the dimension of the multimodal feature space it was found that $d=6$.

In Figure 5, the precision-recall diagrams of the proposed method against other similar cross-media retrieval approaches are presented. A definition of precision and recall values is available at [ZARPALAS 2006]. The proposed method was compared in our experimental dataset of 264 MMDs with the Local Regression Global Alignment (LRGA) method [YANG 2009], the Modified Manifold Ranking (MMR) method [VANAMALI 2010] and the Locally Linear Embedding (LLE) [SAUL 2003]. It must be noted that all methods were implemented by us in C++, using a PC with 3GB RAM running Windows XP.

These results, though, are not indicative of real-life multimodal search tasks, where the queries usually do not belong to the MMD database. In this case, the multimodal descriptors of the query MMD are not available, thus the distance between the query and the database MMDs cannot be directly calculated. LRGA, MMR and LLE methods do not support multimodal queries, but they adopt a rather simplified solution to address queries that do not belong to the database. More specifically, the query, which should be of one single modality, is matched only with the media items of the same modality, with respect to their mono-modal distances and the most relevant corresponding MMDs are retrieved. Then, the retrieval process is repeated using as queries the first retrieved MMDs. Moving one step further, the proposed method not only supports multimodal queries, but it can estimate the multimodal descriptor of an MMD not belonging to the database, using an RBF network.

In Figure 6, the precision-recall diagrams for queries that do not belong to the database are shown. In this scenario, 24 query MMDs, two from each semantic category, are selected. Since LRGA, MMR and LLE do not support multimodal queries, the 24 query MMDs consist of one single modality, either 3D objects or 2D images. The proposed method clearly outperforms the other approaches, which demonstrates its ability to deal with real queries.

Multimodal retrieval can be very useful in real-life multimedia retrieval problems. Especially in 3D object retrieval, the user may not have a 3D model available in his/her local storage in order to use as query. On the other hand, a query image is more likely to be available. An example of 2D-to-3D multimodal retrieval is given in Figure 7. This demo has been developed in the context of the EU-funded project I-SEARCH. The url of the website is: http://www.isearch-project.eu/isearch/.

5 Conclusions

This paper proposed a novel framework for multimodal search and retrieval. Multimedia data are organised in rich media containers, namely MMDs, which enclose semantically
similar media items. Using a manifold learning method based on Laplacian Eigenmaps, a new low-dimensional multimodal feature space is created, where all MMDs are represented as data points. During multimodal search and retrieval, multimodal queries are matched with the database MMDs with respect to their multimodal descriptor vectors. In case the query does not belong to the database, a RBF network predicts its multimodal descriptor with high accuracy. Experiments conducted in a database of MMDs, consisting of 3D objects and 2D images, showed that the proposed method demonstrates superior performance over existing cross-media retrieval approaches, especially for queries not belonging to the database. Moreover, the method accepts queries of multiple modalities simultaneously, which is not supported by any of the cross-modal retrieval methods reported so far.

![Figure 7](image-url)  
**Figure 7:** A 2D-to-3D multimodal retrieval scenario: the query is an image of a chair and the retrieved results are similar 3D objects. The demo is available at the website of the I-SEARCH project.

**Acknowledgements**

This work was supported by the EU-funded project I-SEARCH.

**References**


