Multimedia Search and Retrieval using Multimodal Annotation Propagation and Indexing Techniques

Michalis Lazaridis, Apostolos Axenopoulos, Dimitrios Rafailidis and Petros Daras

Informatics and Telematics Institute, Centre for Research and Technology Hellas, Thessaloniki, Greece email: {*lazar,axenop,drafail,daras*}*@iti.gr*

Abstract

In this paper, a novel framework for multimodal search and retrieval of rich media objects is presented. The searchable items are media representations consisting of multiple modalities, such as 2D images, 3D objects and audio files, which share a common semantic concept. A manifold learning technique based on Laplacian Eigenmaps was appropriately modified in order to merge the low-level descriptors of each separate modality and create a new low-dimensional multimodal feature space, where all media objects can be mapped irrespective of their constituting modalities. To accelerate search and retrieval and make the framework suitable even for web-scale applications, a multimedia indexing scheme is adopted, which represents each object of the dataset by the ordering of a number of reference objects. Moreover, the hubness property is introduced in this paper as a criterion to select the most representative reference objects, thus, achieving the maximum possible performance of indexing. The content-based similarity of the multimodal descriptors is also used to automatically annotate the objects of the dataset using a predefined set of attributes. Annotation propagation is utilized to approximate the multimodal descriptors for multimodal queries that do not belong to the dataset.

Keywords: Multimodal search and retrieval, Multimedia indexing, Annotation Propagation, Hubness

1. Introduction

Internet was designed and primarily used by scientists for networking research and exchanging information between each other. However, the explosion of the World Wide Web (which has started as a document repository) and its successful descendants (Semantic Web and Web 2.0), along with the widespread availability of digital recording devices, improved modeling tools, advanced scanning mechanisms as well as display and rendering devices, are rapidly transforming the Internet to a fully fledged 3D collaborative environment that facilitates services, interaction and communication. It is therefore now possible for users to rapidly move from a mainly textual-based to a media-based "embodied" Internet, where rich audiovisual content (images, graphics, sound, videos, etc.), 3D representations, virtual and augmented reality worlds, serious games, life-logging applications, multimodal yet affective utterances, become a reality.

This new environment, in which a dramatic increase of net-based audiovisual and 3D object databases that have been produced by professional and amateur users is observed, drives demands towards more sophisticated information representation, filtering, aggregation and search tools for

achieving more efficient information retrieval. Since information is currently perceived, stored and processed in various forms, leading to vast amounts of heterogeneous multimodal data, an optimal search and retrieval engine should allow users to express their query in any form most suitable for them and retrieve content in various forms providing the users with a complete view of the retrieved information. Based on the above facts, a clear need for a new generation of multimodal search engines is emerging. These will be able to handle specific types of multimedia (text, 2D image, sketch, video, 3D objects, audio and combination of the above), which can be used as queries and retrieve any available relevant content of any of the aforementioned types.

The need to move beyond traditional text-based retrieval approaches has led the scientific community to invent novel efficient concepts for multimedia retrieval. Most of the methods presented in the last years were focused on the extraction of low-level features (e.g. color, texture, shape, etc.) automatically from content (content-based retrieval). While there are numerous content-based techniques that achieve retrieval of one single modality, such as 3D objects, images, video or audio, only few of them are able to retrieve multiple modalities simultaneously. The latter, which are referred in the literature as multimodal retrieval methods, are expected to bring a new vision to multimedia search engines, since they will enable users 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, search and retrieval methods that are based only on low-level features are not able to achieve the discriminative efficiency of a human. This has driven the research on other, context-based techniques suitable for classification, which at the same time improve significantly the retrieval performance of content-based search engines. The use of semantic annotation seems to be the most powerful technique in this area. Semantic annotation relies on the "attachment" of an amount of information on each multimedia object. In traditional multimedia databases, manual annotation of all database items is a work that requires significant labor. More specifically, the database multimedia objects are presented to the human annotator, one after another, and s/he decides whether the object possesses or not a specific attribute. Manual annotation becomes non-functional in modern database systems, where continuous renewal and management of a constantly increasing volume of information is crucial. The process of annotation propagation focuses on this problem. The key question to better understand the nature of annotation propagation is: "how can we automatically expand annotations of certain already manually annotated multimedia objects to other objects that have the same or similar attributes without presenting them to the user for manual annotation"?

In an attempt to provide a complete solution for search and retrieval of rich multimedia content over modern databases, the framework proposed in this paper combines the advantages of multimodal search with those of annotation propagation into a unified system. Moreover, an effective technique, which is appropriate for web-scale indexing, is adopted, extended and integrated to the proposed framework so as to achieve optimized search and retrieval of rich media content even from large, web-scale databases.

1.1. 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 [1, 2, 3], 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.

Regarding cross-modal retrieval, several methods adopt a well-known technique called Canonical Correlation Analysis (CCA) [4], which constructs an isomorphic subspace (CCA subspace) in order to learn multi-modal correlations of media objects. In [5], a CCA-based method is presented, which defines a general distance function in the CCA subspace using polar coordinates. It also utilizes relevence feedback to improve the results. Similarly, in [6], a method for cross-modal association called Cross-modal Factor Analysis (CFA) is introduced. 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 [4] and Latent Semantic Indexing [7]. In [8], 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 [9] they extended their work by proposing a structure for cross-media indexing over large multi-modal media databases.

Moving beyond cross-media retrieval, several attempts have been made to support retrieval of more than two modalities simultaneously. In [10], 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 [11], a structure called Multimedia Document (MMD) is introduced to define a set of multimedia objects (images, audio and text) that carry the same semantics. After creating a Multimedia Correlation Space (MMCS), where every MMD is represented as a data point, a ranking algorithm was applied, which uses a local linear regression model for each data point and then it globally aligns all of them through a unified objective function. However, the above methods demonstrate significant retrieval accuracy only when the query objects belong to the dataset. For query objects that do not belong to the dataset, the methods provide weak support and only by applying relevance feedback they are able to produce acceptable results.

Closely related to content-based multimedia retrieval is the task of assigning semantic annotations, either manually or automatically to the media items of a database, which can improve the performance of content-based search systems. Research on annotation propagation focuses on the investigation of methods that achieve correct automatic annotation to the items of a dataset with the minimum manual effort. Annotation propagation in multimedia databases is still an open research topic, although several approaches have been proposed so far which offer acceptable and realistic solutions to the problem. Almost all of these methods deal with one single modality at a time, e.g. image annotation, video annotation, 3D content annotation and so on. Multimodal annotation propagation involves the task of automatically assigning semantic attributes to the items of a dataset, which may consist of multiple media types. To achieve this, the multimodal correlations among the items' constituting modalities need to be identified, using techniques similar to the ones of cross-modal and multimodal search and retrieval. A presentation of the most important methods for mono-modal annotation propagation, dealing with 3D and 2D content, is given in the sequel. It is worth to mention that, to the best of our knowledge, there is no method

for multimodal annotation propagation presented so far.

In [12], a classification-based, keyword propagation method is presented. The proposed framework consists of an image database that links images to semantic annotations, a similarity measure that integrates both semantic features and image features, and a machine learning algorithm to iteratively update the semantic network and to improve the systems performance over time. In [13], the use of the maximum entropy approach for the task of automatic image annotation is proposed. Maximum entropy allows one to predict the probability of a label in the test data, when labeled training data are available. The technique allows for the effective capturing of relationships between features. In [14], an attempt is made to propagate semantics of the annotations, by using WordNet [15], a lexicographic arrangement of words, and low-level features extracted from the images. The hierarchical organization of WordNet leads to the concept of implication / likelihood among words. In [16], an automatic image annotation system is proposed, which integrates two sets of Support Vector Machines (SVMs), namely the Multiple Instance Learning (MIL)-based (applied to image blocks) and global feature-based SVMs (using global color and texture features), for annotation. Finally, an approach for automatic annotation propagation in 3D object databases is proposed in [17], based on propagation of probabilities through neuro-fuzzy controllers, using a combination of low-level geometric and high-level semantic information.

In the approaches described above for content-based search & retrieval and annotation propagation, similarity search is performed by matching the descriptor of a query object with the descriptors of all database objects. However, this raises scalability issues for web-scale datasets, therefore, more efficient approaches are needed. Towards this direction, approximate similarity search aims to address the scalability issues by minimizing the amount of required computations during content-based search. The field of approximate similarity search can be divided into two major areas: techniques that transform the space of the object's descriptors and techniques that optimize the procedure for accessing and analyzing the data to be indexed and searched. Space transformation mainly addresses dimensionality reduction techniques, which aim to obtain more compact representations of the original data that capture the information necessary for higher-level decision making. Two representative dimensionality reduction techniques are Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). PCA is an unsupervised method that maximally preserves the variance of the data, and LDA is a supervised method that achieves maximal class separation by maximizing the ratio of between-class variance to the within-class variance. In approximate search techniques based on Vector Approximation (VA)-Files [18], dimensionality reduction is obtained by quantizing the original data objects. Other techniques that fall in the category of space transformation are FastMap [19], mainly used in vector spaces, and MetricMap [20] suitable for metric spaces.

Techniques that optimize the procedure for accessing and analyzing the data are basically aiming at reducing the space to be examined through the identification of the most "informative" parts of this space. M-Trees [21] use a hierarchical decomposition of the space. A technique that uses a proximity measure to decide which tree nodes can be pruned, even if their bounding regions overlap the query region, is proposed in [22]. One of the early works that suggested the use of inverted index for Content-Based Information Retrieval (CBIR) is the Viper system [23], in which images are indexed by a huge number of visual features that can either be present or absent in each image, as words in a text document. The approach used in this paper [24] is a hybrid approach since it uses a space transformation, while it also reduces the number of object comparisons needed during the indexing and search. The idea at the basis of these techniques is that when two objects are very close one to each other they 'see' the world around them

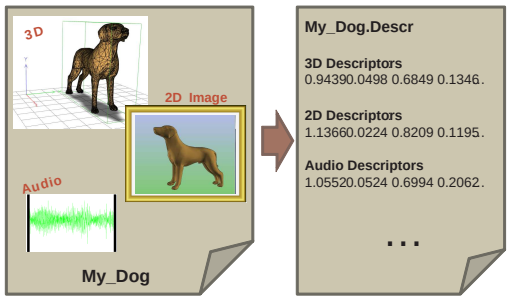


Figure 1: An example of Content Object, which describes the physical entity "My_Dog".

in the same way. Accordingly, we can use a measure of dissimilarity between the views of the world at different objects, in place of the distance function of the underlying metric space. The concept was tested and evaluated using a dataset of 106 million images, taken from Flickr (www.flickr.com) and described by MPEG-7 visual descriptors, with very promising results [25].

The rest of the paper is organized as follows: in Section 2, an overview of the overall framework is given. In Section 3, the multimodal descriptor extraction and indexing are analyzed, which include the construction of the multimodal feature space and the proposed large-scale indexing technique. In Section 4, the multimodal annotation propagation method is described followed by the proposed framework for multimodal search and retrieval. Section 5 analyzes the experimental results. Finally, conclusions are drawn in Section 6.

2. Method Overview and Contributions

2.1. The Concept of Content Object

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 labelling each media instance separately. This approach has been already followed in both [11] and [10], where authors introduced new structures to organize data based on their semantic correlations, namely Multimedia Documents (MMDs) and Multimedia Bags, respectively.

Following and extending this concept, the whole proposed framework is based on a multimedia structure called "*Content Object (CO)*". However, unlike MMDs and Multimedia Bags, COs are not just collections of different media items. A CO 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 accompanied with multiple 2D images and audio files). Moreover, a CO may include additional metadata related to the media, such as textual information, classification information, real-world data (location or time-based), etc. When a user refers to a CO, s/he directly refers to all of its constituting parts.

An example of a Content Object is given in Figure 1. This CO describes the physical entity "My_Dog" and consists of a 3D representation (VRML model), 2D images (snapshots) of the dog, as well as its sound (wav file of the barking sound). In the current work, 3D objects, 2D images and sounds are considered as the constituting modalities of COs. Further extensions of the proposed framework, in order to include other modalities, are planned for future work.

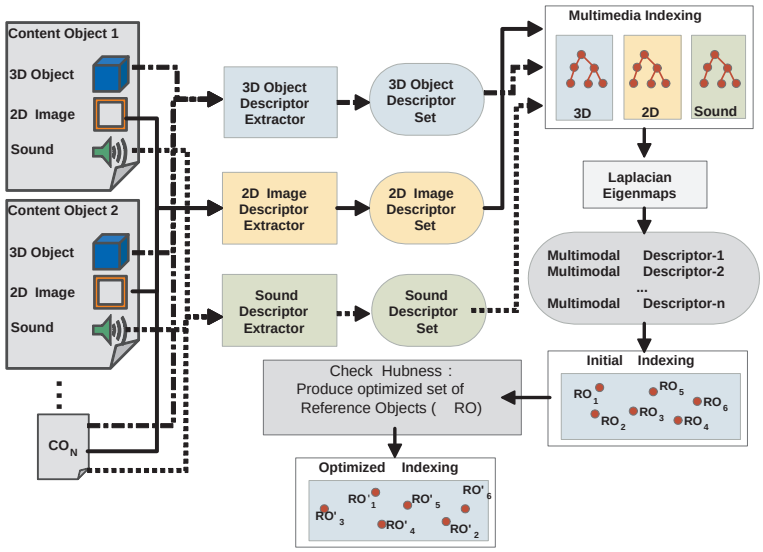


Figure 2: Multimodal descriptor extraction and indexing

2.2. *The Proposed Framework*

The proposed framework consists of the following stages: multimodal descriptor extraction and indexing (Figure 2), multimodal annotation propagation and search & retrieval (Figure 3).

Given a dataset of Content Objects, during the first stage (Figure 2), low-level descriptors are extracted for each of their constituting modalities. Then, using an appropriate manifold learning method based on Laplacian Eigenmaps (LE), the low-level descriptors are mapped to a new low-dimensional feature space. In this feature space, semantically similar COs, irrespective of their constituting modalities, are described by multimodal descriptor vectors close to each other in the Euclidean space. The mono-modal descriptors are indexed, using one indexing structure for each modality, to accelerate the computation of neighbors in the LE method. With the creation of the new multimodal feature space, content-based search and retrieval of COs from the dataset can be performed by directly matching their new multimodal descriptors, resulting in a fully multimodal approach. In order to facilitate faster retrieval, especially when it comes to large-scale (or even web-scale) datasets, an appropriately selected indexing scheme is adopted to index the multimodal descriptors. This indexing technique, which was initially proposed in [24], is further enhanced by intuitively selecting the optimal reference objects. The idea is to identify which COs are the most representative among the whole dataset of COs, by computing the hubness property of each CO.

The concept of annotation propagation is the utilization of the user-provided information for training the annotation system, so that it can be later used for the automatic annotation. During the manual annotation phase (Figure 3), a set of reference COs are presented to the user (annotator) through an annotation interface. The user assigns one or more attributes to the COs, derived from a list of predefined attributes. Then, the system propagates the attributes to the non-annotated COs without any external help. Again, the hubness property of COs is exploited here to select a set of the most representative COs that will be provided for manual annotation. This step improves the automatic annotation and achieves the best percentage of correct automatic

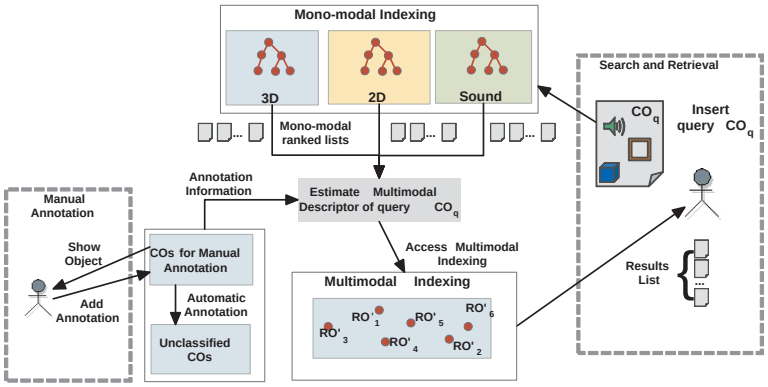


Figure 3: The annotation propagation and multimodal search and retrieval procedures.

annotation with the minimum percentage of manually annotated objects.

During the multimodal search and retrieval phase (Figure 3), a CO_q, which does not belong to the database, is given as query to the system. Since the CO_q's multimodal descriptor is not available, an approximation of the multimodal descriptor is computed as follows: firstly, the CO_q's constituting modalities are used as separate queries; the mono-modal descriptors of the query access the corresponding mono-modal indexing structures, which were described above (Figure 2), and one mono-modal ranked list for each of the query CO's modalities is retrieved. For the k-Nearest Neighbors of each ranked list, the COs, which correspond to the retrieved media items, are identified. Then, the attribute that is most frequent among these COs is assigned to the new query CO_q and the centroid of the multimodal descriptors of only the COs with this attribute is used as an approximate multimodal descriptor of CO_q. The approximate descriptor vector accesses the multimodal indexing to retrieve similar results. The contribution of annotation propagation here is crucial, since the propagated attributes of each CO are used as a criterion to decide which COs will participate in the centroid of multimodal descriptors (i.e. the approximate descriptor of CO_q). If this annotation information was not available, the approximation of the multimodal descriptor of CO_q would not be feasible. In other words, if an approximation was made including all k-Nearest Neighbors of each ranked list, irrespective of their attributes, the approximate descriptor would contain outliers which would affect the retrieval accuracy of the method.

The proposed method introduces the following innovative features:

It combines a new multimodal descriptor with an efficient indexing scheme. Indexing is applied to both the mono-modal descriptors, to avoid computation of large distance matrices during the creation of the multimodal feature space, and the multimodal descriptors, to facilitate faster multimodal retrieval in large scale.

It improves the retrieval accuracy of indexing, by selecting the optimal reference objects. Instead of randomly selecting reference objects from the dataset, the proposed method exploits the hubness property of the multimodal descriptors and picks the most representative ones.

It provides multimodal annotation propagation using minimum number of manually annotated COs. Again, this is achieved by selecting the most representative COs (hubs) for manual annotation, instead of random selection.

It exploits annotation propagation for multimodal search and retrieval when query COs do

not belong to the indexed dataset. Since the multimodal descriptor of a query CO that does not belong to the dataset cannot be directly calculated, multiple ranked lists are retrieved from the mono-modal search tasks; then, the first ranked objects of each ranked list are used to approximate a multimodal descriptor of the query CO. Annotation propagation here contributes so as to have a more accurate approximation.

3. Multimodal descriptor extraction and indexing

3.1. Construction of a multimodal feature space

During the construction of the multimodal feature space, all COs of a dataset, irrespective of their constituting modalities, are represented as d-dimensional points in a new feature space. In this feature space, semantically similar COs 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 [1, 2], 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 neighbors among all items of the dataset in order to construct an adjacency matrix. In our case, the items of the dataset are COs. The k-nearest neighbor computation for a CO is not trivial though, since it requires merging descriptors of heterogeneous modalities into one unified representation. Although such unified distances have been already presented to address cross-media retrieval [10, 11], they suffer from the following limitation: 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 neighbors, 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 COs, the neighborhood criterion is determined as follows: two COs, 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 COs 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.

Given now a multimedia dataset of N COs and p different modalities, the goal is to compute the k-nearest neighbors for every CO_i, $1 \leq i \leq N$. For simplicity, we assume that each CO_i consists of exactly one item per modality, although it is possible to have only few modalities in CO_i as well as more than one items of the same modality. Let a media item within CO_i of m-th modality ($1 \leq m \leq p$) be represented by the descriptor vector $\mathbf{x}_i^m$. For the m-th modality, a distance measure is defined as $d^m(\mathbf{x}_i^m, \mathbf{x}_j^m)$ to calculate the mono-modal dissimilarity. The k_m-nearest neighbors of $\mathbf{x}_i^m$ are retrieved by ranking all the media items of m-th 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:

$$\mathbf{NeighList}_{CO_i}^m = \{index_{CO_i}^m(1), index_{CO_i}^m(2), \cdots, index_{CO_i}^m(k_m)\} \quad (1)$$

where $index_{CO_i}^m(1)$ is the index of the CO which corresponds to the media item of m-th modality, ranked as the first nearest neighbor of $\mathbf{x}_i^m$. $index_{CO_i}^m(2), \cdots, index_{CO_i}^m(k_m)$ are the indices of the

COs corresponding to the $2^{nd}, \cdots, k_m^{th}$ ranked items, respectively. Similarly, p lists of nearest neighbors are extracted, one for each modality. The final k-nearest neighbors of CO_i are computed by taking equal number of first neighbors from each list $\mathbf{NeighList}_{CO_i}^m$, $1 \leq m \leq p$, i.e. k/p neighbors, with $(k/p) < k_m$. In case a CO_j appears in the k/p neighbors of more than one lists $\mathbf{NeighList}_{CO_i}^m$, this CO_j is counted only once. The remaining positions in the k-nearest neighbors list are then filled with the next closest COs.

In the general case that a CO consists of less than p modalities, more nearest neighbors are taken from each modality, in order to keep the number k of the neighboring COs the same. As an example, let $k = 6$ be the number of k-nearest neighbors of CO_i. If CO_i consists of $p = 2$ modalities, we need $(k/p) = 3$ nearest neighbors from each modality. If CO_i consists of $p = 1$ modality, we need $(k/p) = 6$ nearest neighbors, all from the same modality. Finally, a $N \times k$ matrix, $\mathbf{NN}_{CO}$, is created, where each row i represents the k-nearest neighbors of CO_i.

The $\mathbf{NN}_{CO}$ matrix is taken as input to create a $N \times N$ adjacency matrix $\mathbf{W}$, where:

$$W_{ij} = \begin{cases} 1, & \text{if } CO_j \text{ belongs to } k\text{-neighbors of } CO_i. \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

The Laplacian Eigenmaps (LE) use the matrix $\mathbf{W}$ as input to create a multimodal feature space of low dimension, where every CO is represented as a d-dimensional vector, that is the $N \times d$ matrix $\mathbf{Y}$. A more detailed description of the LE algorithm is available in [33]. The motivation for choosing binary values (1, 0) to construct $\mathbf{W}$ (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.

3.2. An indexing scheme for web-scale retrieval

It is obvious that for the creation of the $N \times N$ adjacency matrix $\mathbf{W}$, a $N \times N$ distance matrix is required, which stores the pair-wise distances among all database's COs. 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 [ref] 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 $\mathbf{S} = \{o_1, o_2, \ldots, o_M\}$ be a set of M media objects and d a distance function between objects of S. Let $\mathbf{RO} \subset \mathbf{S}$ be a set of reference objects chosen from $\mathbf{S}$. An object $o_i \in \mathbf{S}$ can be represented as the ordering $\bar{o}_i$ of the reference objects RO according to their distance d from o_i, as follows: $\bar{o}_i \in O_{d,o_i}^{RO}$, where O_{d,o_i}^{RO} is the ordered list containing all objects of $\mathbf{RO}$, ordered according to their distance d from o_i. The position in O_{d,o_i}^{RO} of a reference object $ro_j \in \mathbf{RO}$ is

denoted as $O^{RO}_{d,o_i}(ro_j)$. The distance between two objects in the transformed domain is given by $\bar{d}(\bar{o}_1,\bar{o}_2) = SFD(O^{RO}_{d,o_1}, O^{RO}_{d,o_2})$, where SFD is the Spearman Footrule Distance, which is used as a measure to compare ordered lists:

$$SFD(O^{RO}_{d,o_1}, O^{RO}_{d,o_2}) = \sum_{ro \in RO} \mid O^{RO}_{d,o_1}(ro) - O^{RO}_{d,o_2}(ro) \mid \tag{3}$$

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 $\mathbf{S}$, $i = 1,2,\ldots,M$. 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 $\mathbf{S}$. The approximate ordering of $\mathbf{S}$ with respect to q can be obtained by computing the distance $\bar{d}(\bar{q},\bar{o}_i)$, $\forall o \in \mathbf{S}$. 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 $ro_j \in \mathbf{RO}$ is a list of pairs $(o_i, O^{RO}_{o_i}(ro_j)), o_i \in S$, that is a list where each object o_i of the dataset $\mathbf{S}$ is associated with the position of the reference object ro_i 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:

$$\begin{aligned} & ro_1 \rightarrow ((o_1, O^{RO}_{o_1}(ro_1)), \ldots, (o_M, O^{RO}_{o_M}(ro_1))) \\ & \ldots \\ & ro_n \rightarrow ((o_1, O^{RO}_{o_1}(ro_n)), \ldots, (o_M, O^{RO}_{o_M}(ro_n))) \end{aligned} \tag{4}$$

where M is the size of the dataset $\mathbf{S}$ and n is the size of the set of reference objects **RO**. A more detailed description of the algorithm is available in [24]. By using the above indexing structure, search within the dataset $\mathbf{S}$ 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 of reference objects *RO*. According to [24], the following inequality must hold so that the retrieval performance is not affected:

$$Size(RO) \geq 2 \cdot \sqrt{Size(S)} \tag{5}$$

The multimedia indexing scheme described above is applied to the mono-modal descriptors, to avoid computation of large distance matrices during the creation of the multimodal feature space, as well as to the multimodal descriptors, to facilitate faster multimodal retrieval in large scale. In the case of the mono-modal descriptors, the indexing algorithm is applied for each modality separately, thus, the dataset $\mathbf{S}$ is the set of media items o of the same m-th modality and d is the distance metric $d^m(\mathbf{x}^m_i, \mathbf{x}^m_j)$ that computes the dissimilarity between the mono-modal descriptors $\mathbf{x}^m$ of the m-th modality. Similarly, in the case of the multimodal descriptors, the dataset $\mathbf{S}$ is the set of COs and d is the distance metric that computes the dissimilarity between their corresponding d-dimensional descriptors, which were extracted by using the LE method (Section 3.1).

3.3. Computing Hubness to improve Indexing

The multimodal indexing scheme described in the previous subsection selects randomly COs from the dataset in order to create the set of reference objects. Similarly, the methods for automatic classification that will be described in the following section select a random sample of the

dataset for manual annotation (training set). However, random selection may not always produce the optimal sample, especially in cases when only a small percentage of the dataset is selected. In this paper, an intuitive way to select sample COs is introduced. The aim is to pick the most representative COs not only for the creation of the **RO** set in indexing but also for the creation of the set of COs for manual annotation in annotation propagation.

In order to estimate which COs are the most representative among the objects of the dataset, their distance concentration will be taken into account. The latter denotes the tendency of distances between all pairs of points in high-dimensional data to become almost equal. Distance concentration and the meaningfulness of nearest neighbors in high dimensions has been thoroughly explored in [26, 27, 28]. Distance concentration is strongly associated with the "*hubness*" phenomenon, which has been recently observed in several application areas involving sound and image data [29, 30, 31].

Our approach regarding the selection of representative COs is based on the following hypothesis: "*the most representative COs for manual annotation and for the creation of the RO set in indexing should be among the hubs of the CO dataset*". Therefore, *hubness* has to be computed. Since we are interested in identifying hubs in our dataset without taking into account the classification information, the empirical analysis performed in [32] is followed:

In particular, two key factors reveal hubness in a dataset: a) *skewness* and b) *position* of COs in the d-dimensional space (the multimodal feature space described in subsection 3.1). Concerning the first key factor for revealing hubness, a formal definition follows:

Hubness H_k is the distribution of k-occurrences: the number of times each CO appears among the k nearest neighbors of other COs in a data set. The computation of nearest neighbors is based on the L2-distance in the d-dimensional multimodal feature space, which is noted as *LEdist*. We consider $\mathbf{M} = \{\mathbf{m}_1, \mathbf{m}_2, \cdots, \mathbf{m}_N\}$ the multimodal descriptor vectors of all COs of the dataset. Let functions p_{ik}, where $i,k \in \{1,2,\cdots,N\}$ with N is the total number of COs in the dataset, be defined as:

$$p_{i,k}(\mathbf{m}) = \begin{cases} 1, & \text{if } \mathbf{m} \text{ belongs to } k\text{-neighbors of } \mathbf{m}_i \text{ according to } \textit{LEdist}. \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

In this setting, the hubness $H_k = \sum_{k=1}^{n} p_{i,k}(\mathbf{m})$ is defined as the number of COs that have $\mathbf{m}$ included in their list of k nearest neighbors. In [32], it is proved that the emergence of hubness is demonstrated by increasing the skewness (S_{H_k}) of the distribution H_k. To characterize the asymmetry of H_k, in particular the skewness, the standardized third moment of the distribution of k-occurences is used, which is defined as:

$$S_{H_k} = \frac{E(H_k - \mu_{H_k})^3}{\sigma_{H_k}^3} \tag{7}$$

where μ_{H_k} and σ_{H_k} are the mean and standard deviation, respectively. If the skewness value is close to zero, there is no hubness. Otherwise for large skewness values, hubness is revealed, especially for high dimensions d (of the multimodal feature space described in subsection 3.1). In Table 1 the skewness value for different dimensions is presented. We validate the existence of hubness for high dimensions, up to 20 (large skewness). Since there is insignificant difference of skewness value for dimensions 20, 25 and 30, while at the same time higher dimensionality increases the computational cost for both eigenvalue computation in LE method [33] and the L2-distance in dissimilarity matching, 20 dimensions of the multimodal feature space are finally kept.

It should be stressed here that hubness does not always imply good classification accuracy. In several cases, an increase in dimensionality may produce questionable results. This is due to the fact that extremely high dimensionality makes distance concentration even more intense. This results in a remarkably high number of hubs, which do not constitute any more a subset of representative objects for the given multimodal feature space [32]. In other words, multiple outliers are identified among the hubs set, which result in "bad" occurrences, i.e. these hubs appear in many k-NN lists but they are not of the same semantic class to the queries. This implies that there should be a tradeoff between high dimensionality, which improves hubness, and not an extremely high dimensionality, which would produce hubs with "bad" occurrences.

Dimensions (d)	Skewness (S_{H_k})
5	0.2313
10	1.2661
15	1.9484
20	3.1263
25	3.3541
30	3.3951

Table 1: Dimensions of the multimodal feature space and skewness values for the examined data set.

Concerning the second key-factor for revealing hubness, to ensure our assumption of selecting 20 dimensions, we demonstrate that the position of a CO in the multimodal d-dimensional space has significant effect on its k-occurrences value, in particular the existence of hubness. Figure 4 plots for each $\mathbf{m}$, its $H_k(\mathbf{m})$, against its L2-distance from the data set mean, for dimensions 5, 10, 20 and 30 in (a), (b) (c) and (d) respectively. The value of k is set to 50, but analogous observations can be made with other values of k. As dimensionality increases, stronger correlation between the position of COs and hubness emerges, implying that more COs are getting closer to the mean. However, for $d = 30$ (Figure 4(d)), a very large number of COs is close to the mean, and optimal selection of representative COs is not feasible (there are several COs selected as hubs, which produce "bad" occurrences).

Consequently, hubs are points with very high k-occurrences which effectively represent "popular" nearest neighbors and can be found on the top left corners in Figures 4 (b), (c), (d). On the contrary, antihubs are points which appear in very few, if any, k-NN lists of other points and can be found on the bottom right corner in the same Figures. Since antihubs are far away from all other points, they can be regarded as distance-based outliers [34]. In [27, 29], hubness is considered in a similar way; more specifically, hub is defined as a point which appears in the nearest neighbors of many points in the dataset. In [27, 29], a considerable amount of the resulting hubs produced "bad" occurrences, which were also treated as outliers. In our case, however, the selection of hubs was beneficial. This can be proven from the results in annotation propagation and indexing (Section 5), in which the hubs-based selection of COs for manual annotation and ROs, respectively, improved the performance.

Consequently, the first l^I COs with the highest value of hubness H_k are selected as reference objects (RO) for the multimodal indexing. In the experiments section, the choice of the optimal number l^I of hubs will be justified.

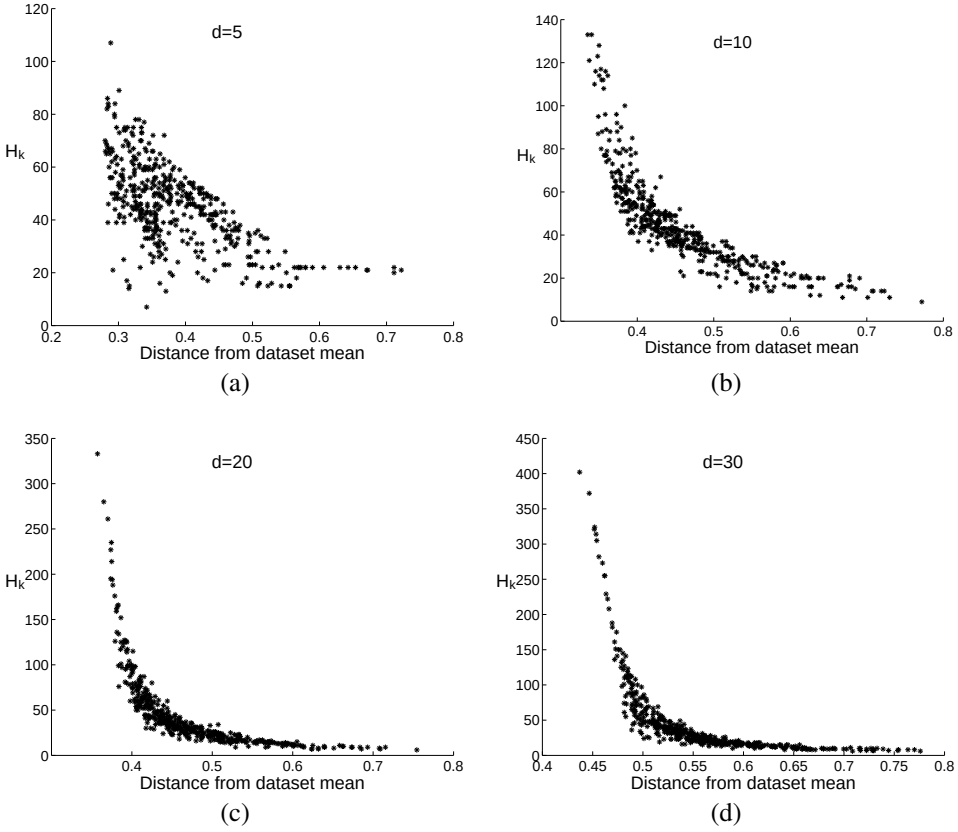


Figure 4: Plots for each descriptor vector $\mathbf{m}$, its $H_k(\mathbf{m})$, against its L2-distance from the data set mean, for dimensions (a) 5 (b) 10 (c) 20 (d) 30. As dimensionality increases, stronger correlation between the position of COs and hubness emerges (higher H_k values).

4. Multimodal Annotation Propagation, Search and Retrieval

4.1. Automatic Annotation for Content Objects of the Dataset

The central idea behind annotation propagation is the utilization of the user-provided information (manual annotation) for training the annotation system, so that it can be later used for the automatic annotation. The annotation propagation procedure can be formulated as: given a set of multimodal descriptor vectors $\mathbf{M} = \{\mathbf{m}_1, \ldots, \mathbf{m}_l, \mathbf{m}_{l+1}, \ldots, \mathbf{m}_N\}$ and an attribute set $\mathbf{A} = \{a_c, c = 1, \ldots, C\}$, the first l^A vectors $\mathbf{m}_i$ $(i \leq l^A)$ are labeled as $y_i \in \mathbf{A}$; and the remaining vectors $\mathbf{m}_u$ $(l^A + 1 \leq u \leq N)$ are to be labeled. In general, an object may have more than one attributes $a_c \in \mathbf{A}$, however, in the current implementation, it is assumed that a Content Object of the dataset can have maximum one attribute. This resembles a classification problem, i.e. the system predicts the category that a CO belongs to. It is also assumed that the attributes are predefined and the manual annotation is correct.

In annotation propagation, two operation modes can be distinguished: the *training mode* (manual annotation) and the *on-line mode* (automatic annotation). During the training mode, a set of COs along with the attributes list are presented to the user (annotator) for manual annotation, through an annotation interface. The user assigns to each CO an attribute from the attribute list. These annotated COs are used to train an appropriately selected classifier. During the on-line mode, the trained classifier assigns automatically attributes (classifies) to the non-annotated COs without any external help.

Choosing the "right" training examples for manual annotation is crucial for the learning ability of the proposed annotation system. In common classification problems, the training samples are randomly selected, a process which does not produce the optimal classification accuracy [35]. In order to achieve higher performance, an increase of the percentage of training samples is required. However, the latter is not trivial when dealing with very large datasets, in which the size poses an obstacle to manual annotation. To overcome this limitation, several methods have been introduced, which focus on selecting the optimal train set, so as to achieve acceptable classification accuracy with the minimum number of train samples.

A well-known approach for optimal selection of training samples is called *Active Learning*. Active learning has been extensively used in relevance feedback applications [36], while little effort has been devoted to apply active learning to annotation propagation. In [17], an Active Learning approach is used, which selects the most informative training examples for manual annotation. The criterion for selecting a specific object for manual annotation is the maximization of the knowledge that will be gained for the system through the manual annotation of this object. This is an iterative procedure, that is, after selecting an object for manual annotation, all probabilities of the dataset need to be recalculated in order to present to the user the next sample.

In order to avoid the extensive computations of active learning, a new approach for optimal selection of training samples is introduced in this paper. The approach has been inspired by the hubness property of the multimodal descriptors, which was presented in Section 3.3. The criterion for selecting the most representative training samples of the CO dataset is similar to the criterion for selecting reference objects for the indexing scheme. Again, the hubness H_k of each CO is computed and the first l^A COs with the highest value of hubness are selected for manual annotation. In the experiments section, the choice of the optimal number l^A of hubs will be justified. It is worth to mention that the number l^A of the first hubs for manual annotation is not necessarily the same as the number l^I of the first hubs for RO selection in indexing.

After manually annotating a representative sample of the dataset, an appropriately selected classifier is assigned the task to automatically annotate the remaining COs. In this paper, the

k-Nearest-Neighbor Classifier (kNN) was adopted among several well-known classifiers due to its high classification accuracy (Section 5). Let $\mathbf{M}_A = \{\mathbf{m}_1, \ldots, \mathbf{m}_{N_A}\}$ the descriptor vector set of the manually annotated COs and $\mathbf{Y}_A = \{y_1, \ldots, y_{N_A}\}$ their corresponding attributes, with $y_i \in \mathbf{A}$ and N_A the number of manually annotated COs. Let also $\mathbf{m}_u$ be the multimodal descriptor vector of an unclassified CO. The dissimilarity of $\mathbf{m}_u$ with the descriptor vector of an annotated CO $\mathbf{m}_i$ is given by their L-2 distance as follows:

$$dis(u,i) = \sqrt{\sum_{j=1}^{d} (m_u(j) - m_i(j))^2} \tag{8}$$

where d is the dimension of the multimodal descriptor vector $\mathbf{m}_i = \{m_i(1), \ldots, m_i(d)\}$. The kNN classifier classifies the new CO as follows: let $Score_u^c$ the score of the new CO to have attribute $a_c \in \mathbf{A}$, $c \in \{1, \ldots, C\}$. C different scores are computed, one for each attribute. The score is given by the following equation:

$$Score_u^c = \sum_{\mathbf{m}_i \in K_{NN}, y_i = a_c} \frac{1}{dis(u,i)} \tag{9}$$

where K_{NN} is the set of $\mathbf{m}_i$, which are the k-nearest neighbors of $\mathbf{m}_u$. The attribute a_c with the maximum score $Score_u^c$ is assigned to the new CO. This procedure is repeated for all unclassified COs of the dataset.

4.2. Multimodal Search using as Queries COs that do not belong to the Dataset

In state-of-the-art cross-modal 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 supports the option to enter multiple query modalities simultaneously. As an example, a CO can be used as query to retrieve semantically similar COs. The constituting modalities of the retrieved COs may be different from the query's modalities, which is a clear step forward in the field of multimodal retrieval. By using as query a CO from the dataset, the retrieval procedure is straightforward: the multimodal descriptor vector of the query CO, which was computed using the proposed LE-based method, is matched against the multimodal descriptor vectors of the rest COs of the dataset and the most relevant results are retrieved. The situation, though, is different when dealing with queries which do not belong to the dataset. These query COs were not included in the LE-based learning process and thus, their multimodal descriptor vectors are not available. Therefore, they cannot be directly matched with the COs of the dataset.

In the complex case where the query does not belong to the dataset, 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 3.1, by adding the query to the initial dataset, a faster and more approximate solution should be 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 [1], where a Radial Basis Function (RBF) network was applied to map the initial low-level descriptors of 3D objects to a new feature space of lower dimension. However, in [1], 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, a novel approach, which exploits the results of automatic annotation described in the previous section, is presented. Let CO_q be a query Content Object that does not belong to the dataset. Without loss of generality, we assume that CO_q consists of two modalities and it is represented by the mono-modal descriptors of each modality: $\mathbf{x}_q^1$ and $\mathbf{x}_q^2$. At a first stage, two mono-modal searches (one per modality) are performed in parallel, using the mono-modal indexing schemes presented in Section 3.2. The ranked lists of k_1 and k_2-nearest neighbors of $\mathbf{x}_q^1$ and $\mathbf{x}_q^2$, respectively, are defined as in equation (1):

$$\begin{aligned} \mathbf{NeighList}_{CO_q}^1 &= index_{CO_q}^1(1), index_{CO_q}^1(2), \cdots, index_{CO_q}^1(k_1) \\ \mathbf{NeighList}_{CO_q}^2 &= index_{CO_q}^2(1), index_{CO_q}^2(2), \cdots, index_{CO_q}^2(k_2) \end{aligned} \tag{10}$$

where $index_{CO_q}^1(1)$ is the index of the CO of the dataset, which corresponds to the media item of the 1^{st} modality, ranked as the first nearest neighbor of $\mathbf{x}_q^1$; $index_{CO_q}^1(2), \cdots, index_{CO_q}^1(k_1)$ are the indices of the COs corresponding to the $2^{nd}, \cdots, k_1^{th}$ ranked items, respectively, and so on. We also keep equal number of k-first neighbors, i.e. $k_1 = k_2 = k$. In order to compute the score $Score_q^c$ of CO_q to have a specific attribute a_c, we modify equation (9) as follows:

$$Score_q^c = \sum_{m=1}^{2} \Big(\sum_{i \in \mathbf{NeighList}_{CO_q}^m, y_i = a_c} 1 \Big) \tag{11}$$

in other words, the score $Score_q^c$ is accumulated by 1, when CO_i, which is among the k-first neighbors of CO_q's m^{th} modality (i.e. belongs to ranked list $\mathbf{NeighList}_{CO_q}^m$), has attribute equal to a_c. Again, the attribute a_c with the maximum score $Score_q^c$ is assigned to CO_q.

The procedure of automatically assigning an attribute to a query CO that does not belong to the dataset differs from the procedure described in Subsection 4.1 in the following aspects: first of all, when the query CO does not belong to the dataset, the k-nearest neighbors are calculated per modality using the mono-modal indexing and are retrieved from the entire dataset, while for COs of the dataset (Subsection 4.1) the k-nearest neighbors are calculated using the multimodal descriptor vectors and are retrieved from the list of manually annotated COs only. Moreover, in (11), the weighting factor $1/dis(u,i)$ of (9) is not taken into account. This is due to the heterogeneity of distance metrics between descriptors of different modalities, as explained in Subsection 3.1. Finally, it is worth to mention that, in the case of a query CO not belonging to the dataset, the attributes of the nearest neighbors are not 100% correct, since the automatic annotation does not achieve 100% accuracy. However, the retrieval performance still remains high, as it will be presented in the Experiments section.

Let, now, $\mathbf{M}_{NN}$ be the set of multimodal descriptors $\mathbf{m}_i$, where $i \in NeighList_q^1 \bigcup NeighList_q^2$ and $y_i = y_q$, i.e. the set of nearest neighbors of both ranked lists of the query CO_q, which are assigned the same attribute with the query. Since the multimodal descriptor of CO_q is not available, an approximate multimodal descriptor vector $\tilde{\mathbf{m}}_q$ can be estimated by:

$$\tilde{\mathbf{m}}_q = \frac{1}{|M_{NN}|} \sum_{\mathbf{m}_i \in M_{NN}} \mathbf{m}_i \tag{12}$$

A similar approach to compute approximate multimodal descriptors was presented in [10], where, instead of automatically annotating the "items" of the dataset, the user marks explicitly which "items" from the ranked list are relevant to the query (relevance feedback). In a similar sense, the method proposed in this paper resembles an approach of implicit relevance feedback,

where the user feedback is the manual annotation, which is automatically propagated to the COs of the dataset. Thus, annotation propagation is exploited here to improve the accuracy in multimodal retrieval.

The approximate multimodal descriptor vector $\tilde{\mathbf{m}}_q$ is used as query to retrieve similar COs with respect to multimodal descriptor similarity. The multimodal retrieval framework described above can be applied even when the query CO consists of more than two modalities, without any further extensions. The only difference, here, is that more mono-modal ranked lists will be taken into account to create the approximate multimodal descriptor.

In case a new CO is introduced to the dataset, it can be easily inserted to the multimodal index by exploiting the proposed framework. More specifically, the approximate multimodal descriptor of the CO is computed and the indexing algorithm updates the index structure of the dataset, so that the new CO can be retrieved. Thus, the procedure of applying the LE-based algorithm for constructing the multimodal feature space, every time a new CO is inserted, can be avoided, which saves computational time. The only limitation here is that the new CO should belong to one of the predefined semantic categories of the dataset. Otherwise, the option to introduce new attributes to the dataset should be supported, probably through an appropriate automated tool. The latter is currently not available, but it is an interesting topic for further research.

5. Experimental Results

For the experimental evaluation of the proposed method, a multimodal dataset was compiled by us, since, to the best of our knowledge, no benchmark dataset for multimodal retrieval is available. For the creation of the dataset, three different modalities were used, namely 3D objects, 2D images and sounds. A total number of 495 COs is created, classified into 10 categories: tetrapods (50), birds (49), airplanes (50), helicopters (50), cars (50), motorcycles (50), guns (49), ships (49), string instruments (49) and missile (49). To create these COs, 266 3D objects, 370 2D images and 283 sounds were used. The 3D objects were collected from well-known 3D shape benchmarks, such as the Princeton Shape Benchmark (PSB) [37], the SHREC 2009 and SHREC 2010 Generic Shape Benchmarks [38, 39]. The 2D images were produced as renderings of the 3D objects, while the sounds were collected from the Internet and manually linked to the corresponding visual objects. The dataset can be downloaded from the following url:

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

The 3D object descriptors were extracted using the combined Depth-Silhouette-Radialized Extent (DSR) descriptor [40]. The 2D image descriptors consist of 2D Polar-Fourier coefficients, Zernike moments and Krawtchouk moments [41]. The images are snapshots of the 3D objects with no background information, however, the framework can be extended to include real images (this implies also that the above 2D shape-based descriptors should be replaced by the appropriate descriptors for real images). Finally, the audio descriptors are extracted using the algorithm presented in [42].

The set of attributes for the multimodal annotation propagation procedure reflects the classification scheme of the dataset, i.e.:

$$A = \{tetrapod, bird, airplane, helicopter, car, motorcycle, gun, ship, string, missile\} \quad (13)$$

In the first series of experiments, each CO from the dataset is used as query to retrieve similar COs. In this phase, the retrieval accuracy of the indexing scheme presented in [24] is compared to the enhanced indexing proposed in this paper, which exploits the hubness property to select the optimal set of reference objects. Regarding multimodal annotation propagation, the performance of automatic annotation, when the set of manually annotated COs is randomly selected, is compared with the performance, when the hubness property is used to select the optimal COs for manual annotation. Several classifiers are also tested in this stage. The contribution of automatic annotation to multimodal search for queries that do not belong to the dataset is demonstrated in a dataset of 50 new COs (5 COs from each of the above categories). Since the creation of a web-scale multimodal dataset is not a trivial task, we do not have actual experimental results for very large datasets. However, we are able to prove, through some complexity computations, that the proposed indexing scheme scales well when the size of the dataset increases.

5.1. Performance evaluation using the proposed indexing scheme

In order to evaluate the performance of the proposed indexing method, the following metrics were used: *Nearest Neighbor*, *Tier-1 Precision*, *Tier-2 Precision* as well as the *Recall* measure. The first three evaluation measures share the similar idea, that is, to check the ratio of COs in the query's class that also appear within the top K matches, where K can be 1, the size of the query's class, or the double size of the query's class. More specifically, for a class with N_C members, $K = 1$ for Nearest Neighbor, $K = N_C - 1$ for the first tier, and $K = 2 * (N_C - 1)$ for the second tier. The final score is an average over all the objects in database. The Recall measure is defined as follows:

$$R = \frac{\#(S \bigcap S^A)}{\#S} \tag{14}$$

where S and S^A are the ordering of the k-closest objects to the query CO found by an all-to-all similarity search algorithm and by the indexing method, respectively [24]. In Figure 5, the Recall measure for different values of k-retrieved objects is presented. The proposed indexing approach, which uses hubs as ROs, is compared to the indexing approach presented in [24], where ROs are randomly selected. When the number of ROs is derived from (5), i.e. 45 ROs are considered for the dataset of 495 COs, both indexing methods demonstrate similar performance. However, if we decrease significantly the number of ROs, the performance of our method based on hubs is not affected, while the performance of the method based on random selection of ROs is seriously degraded. More specifically, when 10 ROs are used, the proposed indexing method achieves higher performance, especially for $k \leq 25$. As an example, when $k = 10$, i.e. the first 10 retrieved objects, the proposed indexing has a recall value equal to 0.9, which means that its retrieval accuracy is very close to the accuracy achieved by all-to-all similarity matching. The corresponding recall value for the indexing proposed in [24] is less than 0.8. The above results demonstrate the contribution of hubness to the choice of reference objects. It must be noted here that in the cases of random selection of ROs 20 different random seeds were used and the results presented in Figure 5 are the average values.

While for a dataset of 495 objects, the number of ROs is not a critical issue, for web-scale datasets, it is expected that the number of ROs in indexing would significantly affect the response times of the system. Therefore, the contribution of hubness to the selection of ROs in indexing is a significant step forward. Towards this direction, experiments of indexing using large-scale datasets is an interesting challenge for further research.

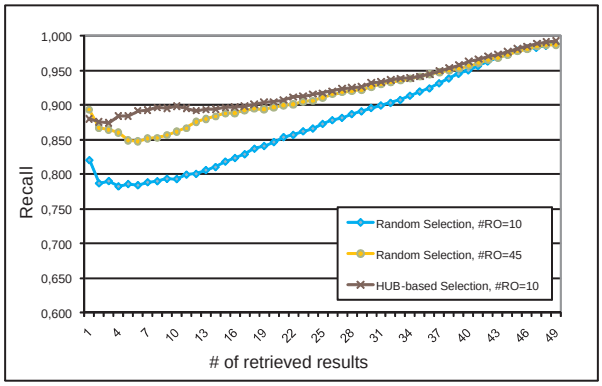


Figure 5: The Recall measure for different numbers of retrieved COs, using the following variations of indexing: a) 10 ROs randomly selected, b) 45 ROs randomly selected and c) 10 ROs selected using hubness.

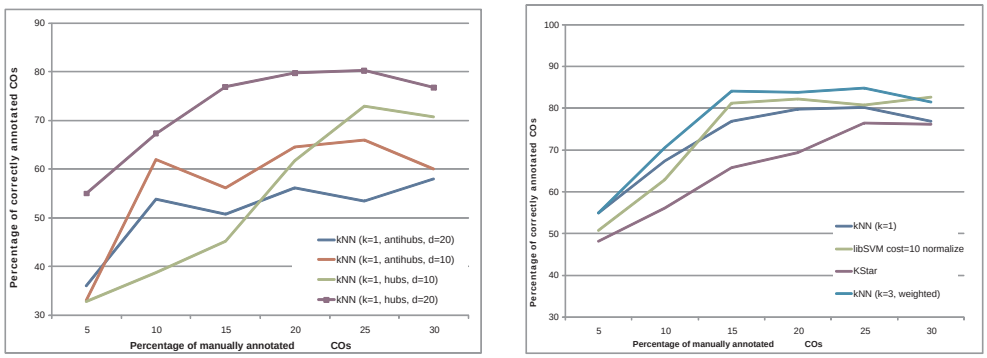
In Table 2, the NN, Tier1 and Tier2 measures are presented for all-to-all similarity search and the two indexing approaches (with random ROs and with hubs as ROs). In these experiments, the dimension of the multimodal descriptors was $d = 20$ and the number of first hubs, which is also equal to the number of selected ROs, was $l^I = \#ROs = 10$.

	All-to-all Similarity	Indexing (Random ROs)	Indexing (Hubs as ROs)
NN	0.846	0.794	0.829
Tier1	0.636	0.636	0.636
Tier2	0.384	0.382	0.384

Table 2: The results of Nearest Neighbor (NN), First-tier (Tier1) and Second-tier (Tier2) for all-to-all similarity search and the two types of indexing (with random ROs and with hubs as ROs).

From Table 2, it is obvious that our enhanced indexing method (using hubs) achieves higher retrieval performance, with respect to the NN metric, than the indexing presented in [24]. Moreover, the value of NN for our enhanced indexing is close to the one of all-to-all similarity search. For the Tier1 and Tier2 metrics, the values are similar for the 3 methods. According to the diagram in Figure 5, for a number of retrieved objects $k = 50$, which is equal to the number of objects in Tier1, the recall for both indexing methods is almost equal to 1, which means that all three methods have retrieved the same percentage of relevant objects.

The efficiency of the proposed indexing method with respect to computational cost and storage requirements cannot be demonstrated using the multimodal dataset that was used in the current work. In order to illustrate the capabilities of indexing, a theoretical example follows. Let a database of $N = 50000$ COs. Each CO is represented by a d-dimensional multimodal descriptor vector ($d \approxeq 20$), according to the method based on LE (Section 3.1). For a query CO_q, one-to-all similarity search within this dataset, without using indexing, involves $N \times d = 10^6$ calculations, if a simple distance metric (such as L-2 distance) is used. If the proposed indexing method is adopted, a total of $d \times \#RO + \#RO \times N'$ calculations is required, where $\#RO$ is the number of reference objects and N' is the number of objects in each inverted file (equation (4)), which are actually accessed. According to [24], not all of the $N = 50000$ COs of the inverted file need to be accessed for a given query CO_q. Only a number of $N' = 50$ objects is enough to obtain the accu-



(a) (b)

Figure 6: Performance of automatic annotation for different percentage of manual annotation: a) comparison of hubs with antihubs selected as objects for manual annotation, with multimodal descriptor dimensionality d=10 and d=20; b) comparison of different classifiers, using hubs and d=20.

rate results. If (5) is used to calculate the number of ROs, then $\#RO = 447$ and the total number of calculations is reduced to 31290 (~ 30 times faster). If, now, hubness is used for the selection of ROs, the number $\#RO$ and the number of calculations can be further reduced (the exact number of ROs in this case needs to be experimentally determined). 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, without indexing, is proportional to $N \times N$. On the other hand, the storage requirements of the proposed indexing method is much less than $\#RO \times N = 2 \cdot N^{3/2}$, which is more compact than the distance matrix.

The task of creating a large-scale multimodal dataset is not trivial, thus, the current work lacks sufficient experimental results to prove the efficiency of large-scale indexing. This remains a challenge for future work, since multimodal search is an open research topic and leaves space for further achievements.

5.2. Performance evaluation of multimodal annotation propagation

In order to assess the accuracy of the proposed automatic annotation method, the following procedure was followed: a subset of the CO dataset was selected for manual annotation and the remaining COs were automatically classified. Three different options for selecting COs for manual annotation were taken: a) random selection, b) selection of the first l^A hubs and c) selection of the first l^A antihubs. With respect to random selection, various random seeds were used. For each random seed, the classification accuracy for a specific amount of manually annotated COs demonstrated significant variations. To verify this, for all different runs we applied statistical pair-wise t-test; the calculated differences of means were significant at level 0.05. This means that, for $5\% - 30\%$ manual annotation, it is not possible to randomly select a representative train set and the classification accuracy is highly dependent on the random seed. Therefore, no results for random selection were presented.

In Figure 6 (a), a comparison of hubs with antihubs selected as objects for manual annotation is depicted. The multimodal descriptor dimensionality was set to d=10 and d=20. It is obvious that antihubs demonstrate a very unstable behavior comparing with that of hubs. Moreover, for

dimension d=10, the percentage of correct automatic annotation for hubs is very low comparing with that of hubs with dimension d=20. This makes sense since for dimension d=10, where the hubness is not very obvious (Figure 4), while for d=20, the hubness is more distinct. It is worth to mention that for percentage of manual annotation higher than 20%, the correct automatic annotation is not improved, it rather decreases for values higher than 25%. This can be explained as follows: since the amount of manually annotated COs is taken from the set of hubs, an increase in the percentage of manually annotated COs results in an increase in the number of l^A first hubs. But moving from the top left corner of the diagram in Figure 4 (c) to the bottom right corner results in taking also antihubs along with hubs, which decreases the accuracy. Therefore, a value of about 15% of manual annotation seems to be optimal for this dataset.

In Figure 6 (b), the performance of automatic annotation using hubs of dimension $d = 20$ is presented for some of the most representative classifiers, namely *Nearest Neighbor*, *libSVM*, *kStar* and *k-Nearest-Neighbors*. The implementation of these methods was obtained from the Weka [43] library. The latter appears to outperform the other methods. The parameters of the *k-Nearest-Neighbors* that achieved the best performance are: $k = 3$ and distance-weighted scoring. From the diagram it is also obvious that a 15% of manually annotated objects is sufficient to achieve more than 83% correct automatic annotation, which is not improved if we increase the amount of manually annotated objects. This results in an optimal number of manually annotated objects $l^A = 74$. It must be noted that the number of l^A best hubs for the manual annotation is not the same as the number of the l^I best hubs for the RO selection in indexing.

5.3. Performance of the proposed approach for queries that do not belong to the dataset

The task of retrieving COs when the query does not belong to the dataset is more complex since the multimodal descriptor of the query CO cannot be directly extracted. The method proposed in this paper for queries that do not belong to the dataset exploits the automatic annotation in order to compute approximate multimodal descriptors (Section 4.2). Therefore, high accuracy in automatic annotation is crucial for high retrieval accuracy. The performance (NN, Tier1, Tier2) of the proposed framework for multimodal retrieval of COs that do not belong to the dataset is depicted in Table 3. The following three methods were compared:

RBF-based retrieval: a Radial Basis Function (RBF) network was used to train an RBF function using as training set the multimodal descriptors of the initial dataset of 495 COs. After training, the RBF network is able to predict the multimodal descriptor of a new CO, by taking as input its mono-modal descriptors. RBF was initially adopted in [1] to map 3D object descriptors into a new low-dimensional feature space.

Proposed method (automatic annotation): at first, a number of $l^A = 74$ COs (first hubs) were used for manual annotation and the remaining 421 objects of the dataset were automatically annotated using the method presented in Section 4.1. Then, for a CO_q that does not belong to the dataset, the multimodal retrieval framework presented in Section 4.2 is applied. The numbers of k-Nearest Neighbors in (10) were set to $k_1 = k_2 = k = 3$ and the attribute values y_i in (11) are the estimated attributes (that occurred through automatic annotation). Finally, the approximate descriptor of each query CO_q was computed using equation (12).

Proposed method (actual attributes): in this case, the multimodal retrieval framework presented in Section 4.2 is applied just as in the previous case. The difference here is that, instead of automatically annotating the remaining 421 objects of the dataset, the actual attributes of all 495 objects were used. Thus, in equation (11) the attribute values y_i are not the estimated but the original ones. Obviously, this case does not correspond to a real scenario, it is only used as a ground truth for comparison with the proposed method.

After the estimation of the approximate multimodal descriptor of each CO_q, using the above methods, the multimodal indexing scheme presented in Section 3.2 was used. The dimension of the multimodal descriptors was $d = 20$ and the number of ROs, which is also equal to the number of l^I-first hubs, was $l^I = \#ROs = 10$.

	RBF-based retrieval	Proposed Method (actual attributes)	Proposed Method (automatic annotation)
NN	0.822	0.835	0.827
Tier1	0.638	0.643	0.640
Tier2	0.382	0.387	0.384

Table 3: Comparison of the proposed multimodal retrieval framework with the a method based on RBF, in terms of Nearest Neighbor (NN), First-tier (Tier1) and Second-tier (Tier2), using the 50 COs objects that do not belong to the dataset.

It is obvious that the retrieval performance using the actual attributes is higher than using automatic annotation. This was expected since automatic annotation is not 100% accurate. However, in order to use the actual attributes, manual annotation of the whole dataset is required, which is not realistic, especially for large datasets. On the other hand, the proposed multimodal retrieval framework using automatic annotation shows promising results and it outperforms the approach based on RBF. Although the performance of the latter two methods is comparable, the proposed one was eventually chosen. One of the main reasons for this choice is that RBF takes as input the concatenated vector of all mono-modal descriptors of the CO's constituting modalities. Considering large descriptor vectors per modality as well as more than one media items per modality in the same CO, results in increased complexity of training the RBF network. The proposed method does not suffer from similar complexity problems since each modality within the CO is treated as a separate ranked list, while the use of mono-modal indexing reduces significantly the complexity of the algorithm. Moreover, every time new COs are introduced, RBF would require further training of the dataset, while the proposed method requires only to propagate the annotations to the new COS, without any further training.

In Figure 7, the ranked lists produced by the proposed approach, for query COs that do not belong to the dataset, are presented (the 6-first retrieved results are shown). The proposed approach supports both mono-modal and multimodal queries. It is worth to mention that given a query of one modality (e.g. sound), our method returns COs that do not necessarily contain the sound modality, which is a clear step forward in multimodal retrieval.

6. Conclusions

In this paper, a novel approach for multimodal search and retrieval was presented. The proposed framework achieves retrieval of rich media objects, namely the Content Objects (COs), which consist of multiple modalities. The framework creates a new multimodal feature space, where all COs, irrespective of their constituting modalities can be mapped. Thus, each CO can be represented by a multimodal descriptor. An appropriate indexing scheme is utilized to index these multimodal descriptors so as to accelerate search and retrieval and make the proposed search framework suitable even for web-scale applications. Additionally, a new approach for multimodal annotation propagation is proposed to assign attributes to the COs of the dataset with the minimum possible manual effort. These attributes are utilized to approximate the multimodal

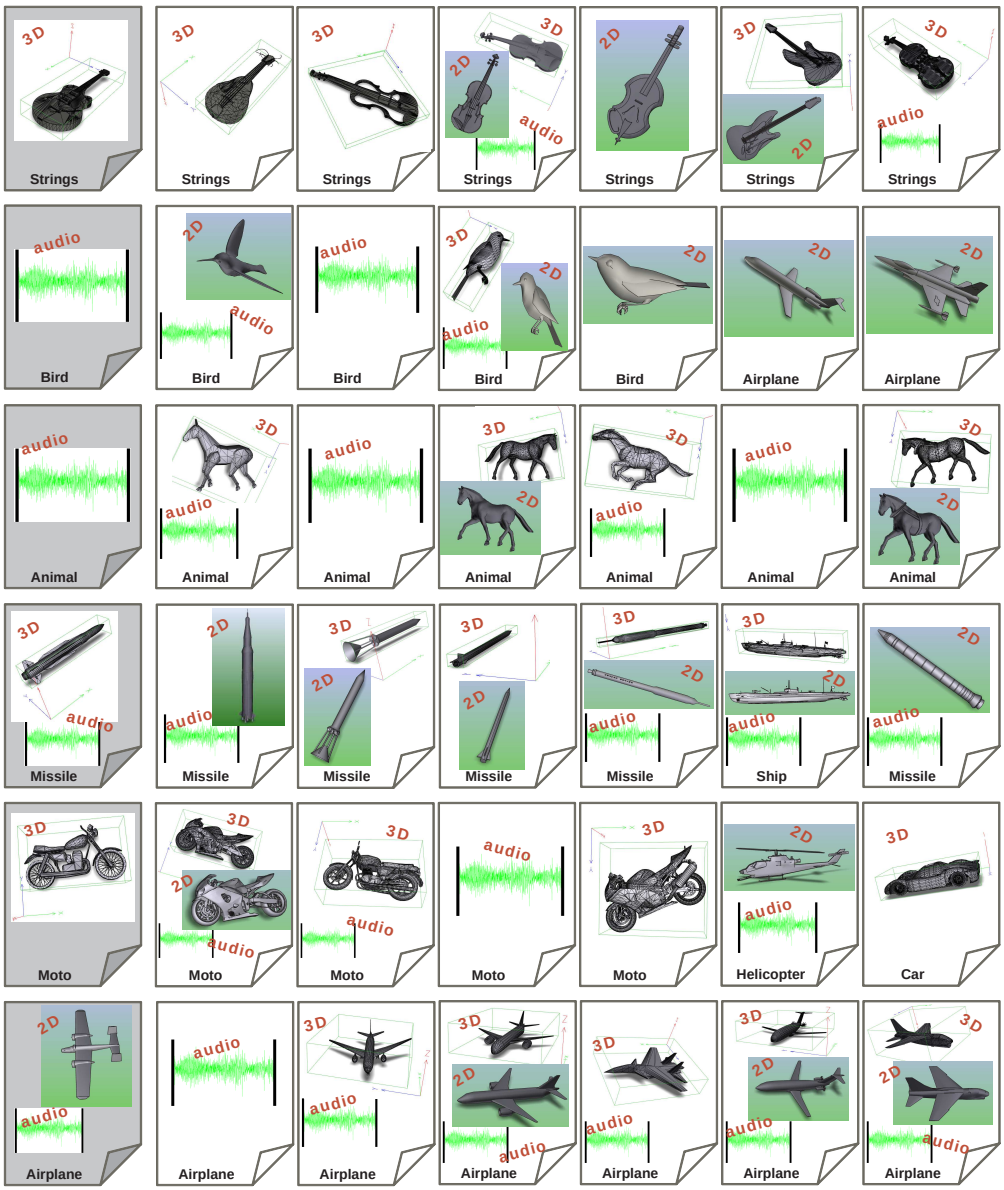


Figure 7: The ranked lists produced by the proposed approach, for query COs that do not belong to the dataset. The query CO is at the first column, while at the rest columns the 6-first retrieved results are shown.

descriptor of a query CO that does not belong to the dataset, which produces improved retrieval results.

The performance of the proposed method was evaluated by using an appropriately constructed multimodal dataset. The method achieves quite promising results both in terms of automatic annotation propagation and multimodal search and retrieval. However, the performance of the proposed framework was not tested for web-scale search tasks, which is essential to demonstrate the full potential of multimodal indexing. The reason is that creating a large-scale multimodal dataset is not a trivial task but it remains a challenge for future work.

Acknowledgment

This work was supported by the EC-funded projects ***I-SEARCH*** and ***ASSETS:***

http://www.isearch-project.eu/isearch/

http://www.assets4europeana.eu/

References

[1] R. Ohbuchi, J. Kobayashi, Unsupervised Learning from a Corpus for Shape-Based 3D Model Retrieval, ACM MIR, Santa Barbara California USA, (2006).

[2] L.K. Saul, S. T. Roweis, Think Globally, Fit Locally: Unsupervised Learning of Low Dimensional Manifolds, Journal of Machine Learning Research, (2003).

[3] J.R. He, M.J. Li, H.J. Zhang, H.H. Tong, C.S. Zhang, Manifold-Ranking Based Image Retrieval, ACM MM, New York USA, (2004).

[4] Pei L. Lai and Colin Fyfe, Canonical correlation analysis using artificial neural networks, Proc. European Symposium on Artificial Neural Networks (ESANN), (1998).

[5] F. Wu, H. Zhang, Y. Zhuang, Learning Semantic Correlations for Cross-Media Retrieval, International Conference in Image Processing, IEEE, (2006).

[6] Dongge Li, Nevenka Dimitrova, Mingkun Li, Ishwar K. Sethi, Multimedia Content Processing through Cross-Modal Association, Proceedings of the eleventh ACM international conference on Multimedia (MM'03), (2003), USA.

[7] Mingkun Li, Dongge Li, Nevenka Dimitrova, and Ishwar K. Sethi, Audio-visual talking face detection, Proc. International Conference on Multimedia and Expo (ICME), pp. 473-476, Baltimore, MD, July (2003).

[8] Zhuang, Y.-T., Yang, Y., Wu, F., Mining Semantic Correlation of Heterogeneous Multimedia Data for Cross-media Retrieval. IEEE TMM 10(2), 221-229 (2008).

[9] Yi Zhuang, Qing Li, and Lei Chen, A Unified Indexing Structure for Efficient Cross-Media Retrieval, DASFAA 2009, LNCS 5463, pp. 677-692, (2009).

[10] H. Zhang, J. Weng, Measuring Multi-modality Similarities Via Subspace Learning for Cross-Media Retrieval, Advances in Multimedia Information Processing - PCM, (2006).

[11] Y. Yang, D. Xu, F. Nie, J. Luo and Y. Zhuang, Ranking with Local Regression and Global Alignment for Cross Media Retrieval, ACM MM, Beijing China, (2009).

[12] H. Zhang, Z. Su, Improving cbir by semantic propagation and cross modality query expansion, In Proceedings of the international workshop on MultiMedia Content-Based Indexing and Retrieval, Brescia, Italy, (2001).

[13] J. Jeon, R. Manmatha, Using maximum entropy for automatic image annotation, Lecture notes in computer science (2004) 2432.

[14] B. Shevade, H. Sundaram, Vidya: An experiental annotation system, Arts Media and Engineering Program, Arizona State University (2003).

[15] Wordnet, A lexical database for English: http://wordnet.princeton.edu/

[16] X. Qi, Y. Han, Incorporating multiple svms for automatic image annotation, Pattern Recognition Volume 40 (No. 2) (2007) 728741.

[17] M. Lazaridis, P. Daras, A neurofuzzy approach to active learning based annotation propagation for 3d object databases, Eurographics Workshop on 3D Object Retrieval (EGW3DOR), Crete, Greece, (2008).

[18] R. Weber and K. Boehm. Trading quality for time with nearest neighbor search. In C. Zaniolo, P. C. Lockemann, M. H. Scholl, and T. Grust, editors, Proceedings of the 7th International Conference on Extending Database Technology (EDBT 2000), Konstanz, Germany, March 27-31, (2000), volume 1777 of Lecture Notes in Computer Science. Springer, 2000.

[19] C. Faloutsos and K.-I. Lin. FastMap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. In M. J. Carey and D. A. Schneider, editors, Proceedings of the 18th ACM International Conference on Management of Data (SIGMOD 1995), San Jose, California, USA, May 22-25, (1995), pages 163-174. ACM Press, 1995.

[20] X. Wang, J. T.-L. Wang, K.-I. Lin, D. Shasha, B. A. Shapiro, and K. Zhang. An index structure for data mining and clustering. In Knowledge and Information Systems, volume 2, pages 161-184. Springer, (2000).

[21] P. Zezula, P. Savino, G. Amato, and F. Rabitti. Approximate similarity retrieval with m-trees. VLDB J., 7(4):275-293, (1998).

[22] G. Amato, F. Rabitti, P. Savino, and P. Zezula. Region proximity in metric spaces and its use for approximate similarity search. ACM Trans. Inf. Syst., 21(2):192-227, (2003).

[23] Squire, D.M., Mueller, W., Mueller, H., Pun, T.: Content-based query of image databases: inspirations from text retrieval. Pattern Recognition Letters 21(13-14), 1193-1198 (2000); Selected Papers from The 11th Scandinavian Conference on Image

[24] G. Amato, P. Savino. Approximate similarity search in metric spaces using inverted files. In: Proceedings of the 3rd International Conference on Scalable Information Systems (InfoScale 2008), pp. 1-10. ICST (2008)

[25] C. Gennaro, G. Amato, P. Bolettieri, and P. Savino. An approach to content-based image retrieval based on the Lucene search engine library. Proceeding ECDL'10 Proceedings of the 14th European conference on Research and advanced technology for digital libraries Springer-Verlag Berlin, Heidelberg (2010)

[26] Kevin S. Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is nearest neighbor meaningful? In Proceedings of the 7th International Conference on Database Theory (ICDT), volume 1540 of Lecture Notes in Computer Science, pages 217235. Springer, (1999).

[27] Alexander Hinneburg, Charu C. Aggarwal, and Daniel A. Keim. What is the nearest neighbor in high dimensional spaces? In Proceedings of the 26th International Conference on Very Large Data Bases (VLDB), pages 506515, (2000).

[28] Charu C. Aggarwal and Philip S. Yu. Outlier detection for high dimensional data. In Proceedings of the 27th ACM SIGMOD International Conference on Management of Data, pages 3746, 2001.; Damien Francois, Vincent Wertz, and Michel Verleysen. The concentration of fractional distances. IEEE Transactions on Knowledge and Data Engineering, 19(7):873886, (2007).

[29] Jean-Julien Aucouturier and Francois Pachet. A scale-free distribution of false positives for a large class of audio similarity measures. Pattern Recognition, 41(1):272284, (2007).

[30] George Doddington, Walter Liggett, Alvin Martin, Mark Przybocki, and Douglas Reynolds. SHEEP, GOATS, LAMBS and WOLVES: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. In Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP), (1998). Paper 0608.

[31] Austin Hicklin, Craig Watson, and Brad Ulery. The myth of goats: How many people have fingerprints that are hard to match? Internal Report 7271, National Institute of Standards and Technology (NIST), USA, (2005).

[32] M. Radovanovic, A. Nanopoulos, M. Ivanovic Hubs in Space: Popular Nearest Neighbors in High-Dimensional Data. Journal of Machine Learning Research (JMLR), 11:24872531, (2010).

[33] Mikhail Belkin and Partha Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Comput. 15 (2003), no. 6, 13731396.

[34] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. Introduction to Data Mining. Addison Wesley, (2005).

[35] C. Zhang, T. Chen, An active learning framework for content based information retrieval, Tech. rep., CMU-AMP-01-04 (2002).

[36] Hong Zhang and Fanlian Meng, Multi-modal Correlation Modeling and Ranking for Retrieval, PCM 2009, LNCS 5879, pp. 637646, (2009).

[37] P. Shilane, P. Min, M. Kazhdan, and T. Funkhouser (2004). The Princeton shape benchmark. In Proceedings of the shape modeling international (SMI 04) (pp. 167178). Genova, Italy, June 2004.

[38] C. Akgul, A. Axenopoulos, B. Bustos, M. Chaouch, P. Daras, H. Dutagaci, T. Furuya, A. Godil, S. Kreft, Z. Lian, T. Napoleon, A. Mademlis, R. Ohbuchi, P. L. Rosin, B. Sankur, T. Schreck, X. Sun, M. Tezuka, Y. Yemez, A. Verroust-Blondet, M.Walter, SHREC 2009 - Generic Shape Retrieval Contest, 30th International conference on EUROGRAPHICS 2009, workshop on 3D object retrieval, Munich, Germany, Mar (2009).

[39] T.P. Vanamali, A. Godil, H. Dutagaci,T. Furuya, Z. Lian, R. Ohbuchi, In: M. Daoudi, T. Schreck, M. Spagnuolo, I. Pratikakis, R. Veltkamp (eds.), SHREC'10 Track: Generic 3D Warehouse, Proceedings of the Eurographics/ACM SIGGRAPH Symposium on 3D Object Retrieval, (2010).

[40] Vranic, D. (2004). 3d model retrieval. Ph.D. Dissertation, University of Leipzig.

[41] P. Daras, A. Axenopoulos, A 3D Shape Retrieval Framework Supporting Multimodal Queries, SPRINGER, International Journal of Computer Vision, DOI 10.1007/s11263-009-0277-2, Jul (2009).

[42] Wichern, Xue, Thornburg, Mechteley, Spanias: Segmentation, Indexing, and Retrieval for Environmental and Natural Sounds, IEEE Transactions on Audio, Speech and Language Processing, March (2010).

[43] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, The WEKA Data Mining Software: An Update, SIGKDD Explorations, Volume 11, Issue 1, (2009).