Unsupervised Visual and Textual Information Fusion in CBMIR Using Graph-Based Methods

Julien Ah-Pine, Gabriela Csurka, Stéphane Clinchant

Unsupervised Visual and Textual Information Fusion in CBMIR using Graph based Methods

Julien Ah-Pine, University of Lyon
Gabriela Csurka, Xerox Research Centre Europe
Stéphane Clinchant, Xerox Research Centre Europe

Multimedia collections are more than ever growing in size and diversity. Effective multimedia retrieval systems are thus critical to access these datasets from the end-user perspective and in a scalable way. We are interested in repositories of image/text multimedia objects and we study multimodal information fusion techniques in the context of content based multimedia information retrieval. We focus on graph based methods which have proven to provide state-of-the-art performances. We particularly examine two of such methods: cross-media similarities and random walk based scores. From a theoretical viewpoint, we propose a unifying graph based framework which encompasses the two aforementioned approaches. Our proposal allows us to highlight the core features one should consider when using a graph based technique for the combination of visual and textual information. We compare cross-media and random walk based results using three different real-world datasets. From a practical standpoint, our extended empirical analyses allow us to provide insights and guidelines about the use of graph based methods for multimodal information fusion in content based multimedia information retrieval.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval model,Search Process*

General Terms: Algorithms,Theory

Additional Key Words and Phrases: Content based multimedia information retrieval, Information fusion, Graph based methods, Cross-media similarity, Random Walk, Visual reranking

1. INTRODUCTION

With the continuous growth of communication technologies, the information that we consult, produce and communicate whatever the communication device we use, has been richer and richer in terms of the media it is composed of. The web has particularly contributed to the production of such multimedia or multimodal data. For instance, web pages from news agencies websites are texts illustrated with pictures or videos; photo sharing websites, such as FlickR, store pictures annotated with tags; video hosting websites, such as Youtube, are again examples of multimedia data repositories. Apart from the web, we have also witnessed the development of new services that rely on digital libraries made of

Author's address: J. Ah-Pine, University of Lyon, ERIC Lab, 5, avenue Pierre Mendes France 69676 Bron Cedex, France. Email : julien.ah-pine@eric.univ-lyon2.fr
G. Csurka and S. Clinchant, Xerox Research Centre Europe, 6 chemin de Maupertuis, 38240 Meylan, France. Emails : gabriela.csurka@xrce.xerox.com and stephane.clinchant@xrce.xerox.com

data composed of several media. In museums for example, there are more and more multimedia applications using text, image, video and speech in order to better plunge the visitor into the historical context of the piece of art she is consulting. New generations of television devices now propose on-line interactive media, on-demand streaming media and so on. The ever-growing production of multimodal data has brought the multimedia research community to address the problem of effectively accessing multimedia repositories from the end-user perspective and this in a scalable way. Accordingly, multimedia data search has been a very active research domain for the last decades.

There are different ways to search a multimedia repository. As for video or image datasets such as Youtube or FlickR, we typically index those media by means of the title, metadata, tags or text associated to or surrounding them. Then, we can search those multimodal objects by using text queries and text based search engines. There are different reasons we use text to retrieve videos or images. Firstly, it is not always possible for the user to query a collection by examples, since the search engine cannot always provide her with examples of videos or images that represent the type of items she would like to retrieve. Secondly, videos or images are stored in machines in a computational representation consisting of low-level features which do not carry by their own high-level semantics. In other words, it is a strong challenge to effectively associate low-level features extracted from videos or images with high-level features such as keywords or tags. This problem is known as the semantic gap. As a consequence of these two difficulties, we generally use the text media for content based multimedia information retrieval (CBMIR) in order to have more relevant search results.

If the text is the core media to use in order to access a multimedia repository effectively, it is however beneficial to use other media during the search process. Indeed, most of research works about multimedia information fusion have shown that combining different modalities to address CBMIR tasks, even with simple strategies, is beneficial. In this paper, we are interested in this topic and we particularly address the combination of visual and textual information. We thus deal with repositories whose items are multimedia objects made of an image associated with a text. Besides, we place ourselves in an unsupervised setting which means that the system we propose does not assume any training set that would help us to learn how to fill the semantic gap between images and texts. In such a context, there are different multimedia information fusion methods. But in our case, we focus on graph based techniques which became very popular in the information retrieval community since the development of techniques like PageRank or Hits [Brin and Page 1998a; Kleinberg 1999; Langville and Meyer 2005].

In a nutshell, the goals of this paper are the following ones:

- We provide an extended state-of-the-art of the main families of approaches for unsupervised visual and textual information fusion. We survey early, late and transmedia methods and this allows us to position our work. The model we propose actually seeks to leverage features of several techniques of the literature with the aim of better filling the semantic gap between visual and textual information.
- We study and compare two popular graph based methods that were originally proposed in two different research communities. On the one hand, we analyze the cross-media similarity approach proposed in [Clinchant et al. 2007; 2008] for image retrieval in the context of ImageCLEF multimedia retrieval tasks. On the other hand, we investigate the random walk approach proposed in [Hsu et al. 2007b; 2007a] for video retrieval

and used in several TRECVID tasks. We show that the two techniques are actually related and can be embedded into a single graph based framework. This generalization allows us to better compare the two techniques and enable us to examine the main points and settings when using graph based methods in CBMIR.

- We analyze two different multimodal search scenarios. In the first scenario, we suppose that the user can only use a text query in order to retrieve images. Multimedia objects of the repository are indexed using their text part and a text based search engine is used in a first place. In a second time, we use the visual information of the (text based) retrieved objects in order to improve the search results. This multimedia search scenario is referred throughout this paper as the *asymmetric* case since the user can only use a text query. In contrast, in the second scenario, the user can use a multimedia query which means that she can enter a text query accompanied with one or several images as examples of her information need. We refer to this second case as the *symmetric search scenario*.
- We tested our multimedia retrieval model with 3 different image/text datasets which have distinct features. We conduct many tests in order to have a better analysis of the core points in the use of the graph based methods under study and in the context of image/text multimedia retrieval. Our experimental results allow us to provide insights and guidelines about how to set the parameters of the unified graph based technique we propose.

The rest of this paper is organized as follows. In section 2, we review the main families of unsupervised multimodal fusion approaches and their features. We take into consideration both the asymmetric and the symmetric search scenarios. In section 3, we discuss the use of graph based methods to fuse visual and textual information and we detail the cross-media similarities and the random walk based techniques. In section 4, in light of the material exposed in sections 2 and 3, we introduce our multimedia relevance model which relies on a generalisation of the visual reranking method and on a unifying graph based framework. Then, we describe in section 5 the experimental settings we conducted on three real-world multimedia collections in order to validate our work. In section 6, we present the experimental results we obtained with different fusion strategies drawn from our multimedia retrieval model. We finally discuss some other advantages of our proposal concerning complexity issues when addressing large collections and we provide some guidelines on how to use the generalized graph based approach we propose. In section 7, we summarize our main findings.

2. UNSUPERVISED MULTIMEDIA FUSION TECHNIQUES IN CBMIR

A general reference about multimedia information access, its challenges and its basic techniques can be found in [Rueger 2010]. This book covers the common topics in multimedia information retrieval such as feature extraction, distance measures, supervised classification (also known as automatic tagging) and fusion of different experts. In this paper, we are particularly interested in multimedia fusion techniques and the literature on this topic is very vast. In this section, we attempt to depict the main families of fusion methods for visual and textual information. It is important to mention that we place ourselves in an unsupervised context which means that we do not use any learning technique in our framework. We can mention at least two research communities that have been addressing this research topic actively. On the one hand, there are the research teams that have participated

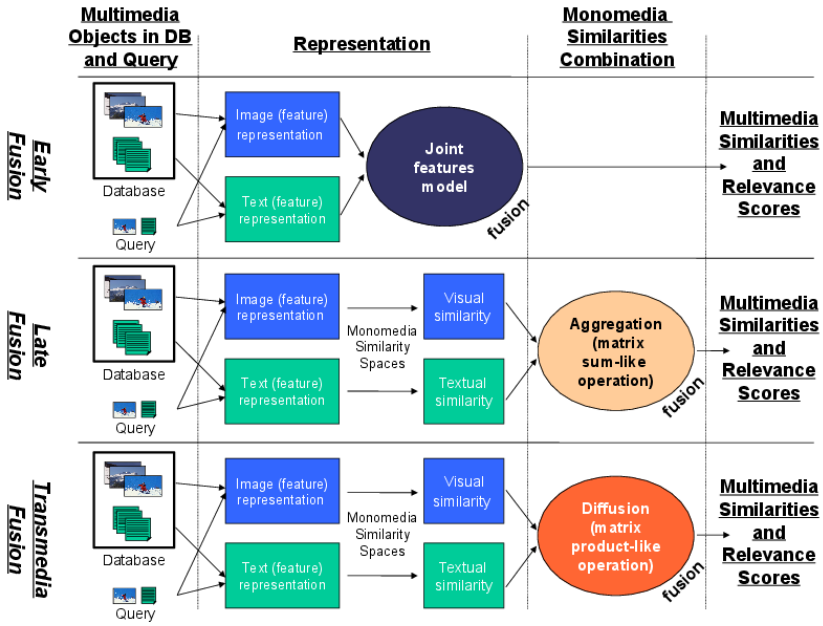

Fig. 1. Early, late and transmedia fusion.

in the TRECVID workshop series and have focused their research efforts on video retrieval [Smeaton et al. 2006]. On the other hand, we can quote the research groups involved in the ImageCLEF meetings and which have been interested in the tracks related to image and multimedia retrieval [Müller et al. 2010]. In the former research community it is usually assumed that the user does not have any example query and the common way to search a multimedia collection relies solely on textual queries. In contrast, in the latter research community, it is generally assumed that the user information need is expressed by a multimedia query composed of a short text query and one or several examples of images. We present in the following, broad families of unsupervised multimedia fusion techniques that have been studied for the two distinct search scenarios. However, in section 2.4 we also point to some research papers that address multimedia fusion techniques from a supervised or a semi-supervised perspective and which show some connections with our work.

2.1 The symmetric case with an image query and a text query

Most of the techniques developed in this context fall into three different categories: early, late and transmedia fusion. We depict these three families of approaches by distinguishing the inherent steps they are composed of. This is summarized in Figure 1. In the following, we assume that the multimedia query can be considered similarly as any item of the multimedia collection that is to say an object made of an image part and a text part. Given a multimedia query, the search process amounts to measuring a multimedia similarity between the query and the multimedia items in the repository.

The early fusion approaches represent the multimedia objects in a multimodal feature space designed *via* a joint model that attempts to map image based features to text based features and *vice versa*. The simplest early fusion method is to concatenate both image

and text feature representations (see *e.g*. [Snoek et al. 2005; Clinchant et al. 2007; Müller et al. 2010]). However, more elaborated joint models such as Canonical Correlation Analysis have been investigated [Mori et al. 1999; Lavrenko et al. 2003; Vinokourov et al. 2003; Rasiwasia et al. 2010]. In the same vein, [Magalhães and Rüger 2010] presents an information theoretic framework that could also fit into this family of fusion approaches.

By contrast, late fusion and transmedia fusion strategies do not act at the level of the monomedia feature representations but rather at the level of the monomedia similarities [Clinchant et al. 2007; Bruno et al. 2008]. In these contexts, we assume that we have effective monomedia similarities and that it is better to combine their respective decisions rather than attempting to bridge the semantic gap at the level of the features.

Concerning late fusion techniques, they mainly seek to merge the monomedia relevance scores by means of aggregation functions. In other words, they can be seen as functions that take as inputs the vectors of monomedia similarities between the query and the documents of the database. In that case, the simplest aggregation technique used is the mean average [Escalante et al. 2008] but more elaborated approaches have been studied (*e.g*. [Caicedo et al. 2010; Müller et al. 2010; Csurka and Clinchant 2012; Wilkins et al. 2010]).

As far as transmedia fusion methods are concerned, they act like similarity diffusion processes. Unlike late fusion methods which implicitly consider documents of the database independent from each other, transmedia techniques leverage the information conveyed by the similarity relationships between each pair of documents. Such methods go beyond late fusion approaches by taking into account not only the similarity vectors of the query with the objects but also the monomedia similarity matrices of elements of the dataset. Transmedia fusion techniques typically involve mixing monomedia similarity matrices by means of matrix multiplication operations [Wang et al. 2004; Pan et al. 2004; Hsu et al. 2007b; Clinchant et al. 2007]. In these relevance models, we usually carry out the similarity diffusion process by using pseudo-relevant items only, which are given by the k nearest neighbors. Furthermore, the transmedia principle seeks to spread one type of monomedia similarities to the other one. Thereby, these methods can also be understood as a generalization of the monomodal pseudo-relevance feedback mechanism in information retrieval (see *e.g*. [Ruthven and Lalmas 2003]).

It is important to mention that there are other ways to categorize the different multimedia fusion techniques. In the recent survey paper [Zha et al. 2012] for example, other terms are used, nevertheless, they basically correspond to the definitions given above with the following mappings: early, late and transmedia fusion are named latent space based, linear fusion and graph based fusion.

2.2 The asymmetric case with a text query only

In addition to the three previously discussed types of fusion methods, [Zha et al. 2012] cites another category named visual reranking. This fourth family of techniques assumes that multimedia collections are accessed using textual queries solely. Therefore, in this context, there is an explicit asymmetry between image and text in the multimedia search scenario.

Visual reranking techniques proceed in two steps (see Figure 2): firstly, based on the similarities with the text query they find the most relevant objects from a semantic viewpoint; then, they employ the visual similarities between objects of the database in order to refine the textual similarity based ranking.

The common assumption that all visual reranking techniques make is that visually sim-

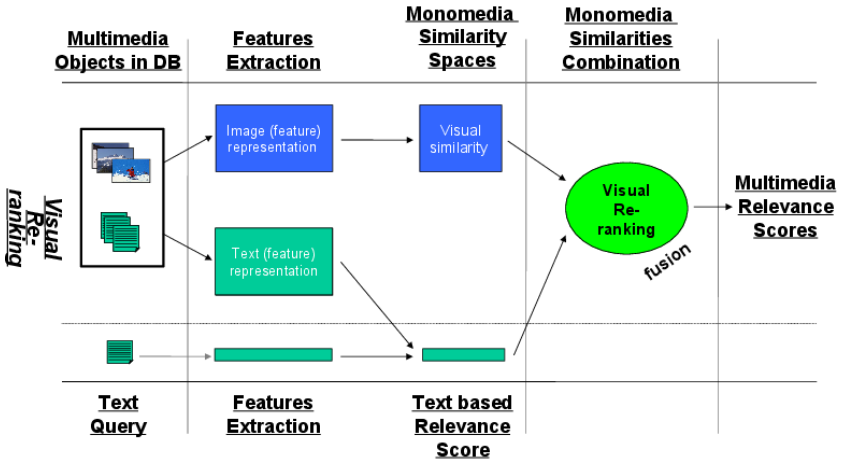

Fig. 2. Visual reranking.

ilar images should have similar relevance scores [Morioka and Wang 2011]. However, different approaches are used to re-arrange the top retrieved items by the text similarities in order to take this principle into account. According to [Zha et al. 2012], we can categorize visual reranking techniques into three subcategories: classification based, clustering based and graph based.

In the first case, pseudo-positive and pseudo-negative documents are sampled from the text based ranked list then a learning to rank algorithm is trained on the visual features (see *e.g.* [Liu 2009] for a general reference on learning to rank methods). Afterwards, objects are re-ordered according to the scores provided by the trained classifier. The critical point is the sampling method used to select pseudo-training examples. The simplest strategy considers items at the top of the list as pseudo-positive and items at the bottom as pseudo-negative but more sophisticated approaches have been proposed [Tian et al. 2008; Yang and Hanjalic 2010; Morioka and Wang 2011].

As for clustering based visual reranking, the main idea is to cluster the list of text based retrieved items and to re-arrange them such that objects that are visually highly similar and have high initial text retrieval scores are favored [Jardine and van Rijsbergen 1971; Hsu et al. 2006; 2007a].

Graph based methods consider multimedia documents as nodes of a graph and the different types of relationships they share as edges. Examples of weighted edges between objects are visual similarities or textual similarities but depending on the application other types of relations can be considered. Graph analysis techniques are then employed in order to infer new features in the goal of re-arranging the text based ranked list of items. One such method, inspired by the well-known PageRank [Brin and Page 1998b; Langville and Meyer 2005; Franceschet 2011], was proposed in [Hsu et al. 2007b; 2007a]. It is based on random walks over a stochastic matrix which is deduced from the fusion of visual and textual similarities, and the stationary probability distribution over the nodes is then additionally used to rerank the initial retrieved list. In the same vein, [Craswell and Szummer 2007] proposed a Markov random walk model with backward and forward steps. They

found out that the best performances were obtained with a long backward walk with high self-transition probability.

2.3 Graph based techniques in both search scenarios

Transmedia fusion techniques we introduced in paragraph 2.1 are technically similar to graph based methods presented in the previous paragraph. Indeed, both approaches use similarity matrices to respectively rank or rerank multimedia items. Graph based methods have proven to be state-of-the-art techniques for many information retrieval tasks (see *e.g.* [Brin and Page 1998b; Langville and Meyer 2005; Franceschet 2011]). In CBMIR too, they have demonstrated their advantages over early or late fusion approaches in many research works (see *e.g.* [Müller et al. 2010; Zha et al. 2012]). We thus focus on such methods in this paper. Besides, there has been few research works that address CBMIR in a symmetric search scenario and using graph based methods. Consequently, in this paper we study both the asymmetric and the symmetric search scenarios with such techniques in order to have a better comparison between them.

Before presenting in more details the two graph based fusion techniques, we present in the next paragraph some additional references that also address multimedia information fusion and/or multimedia retrieval but in a supervised or semi-supervised context.

2.4 Multimedia fusion in a supervised or a semi-supervised context

In [Gao et al. 2013], the authors use hypergraph learning to design a joint visual-textual representation of multimedia objects. This method amounts to an early fusion scheme. Another early fusion approach was presented in [Natsev et al. 2007] and which addresses multimedia query expansion for both the text and the image parts. This work relies on an intermediate representation of multimedia information in a predefined visual-concept lexicon. Classification models are used to map the queries to the lexicon. Then based on pseudo-relevance feedback, different query expansion and score reranking methods are proposed. Similarly, [Rodriguez-Vaamonde et al. 2013] uses an intermediate representation which is based on visual classifiers in their case. To build these classifiers they download images from Google or Bing using query words and they represent these images by classemes (attribute-based image descriptors). In a second step, images in the web pages are classified using these classifiers and the scores are used to rerank the multimodal documents (in their cases the web pages). The reranking is also supervised as a set of training queries with relevance scores are used to learn the parameters of the latter algorithm.

Other related works that are worth mentioning are the following ones [Wang et al. 2009; Wang et al. 2009; Wang et al. 2012]. These papers address video semantic annotation and web image search in a semi-supervised fashion. The general framework used in these contributions is formulated as an optimization problem that simultaneously deal with the late fusion of monomedia similarity matrices and graph semi-supervised learning. The solutions of the optimization problems can be formulated using normalized graphs Laplacian and iterated algorithms are proposed to infer the relevance scores which are further used for annotating videos or ranking images.

Our framework differs from these research works with regard to the followings points:

- We do not use any learning models nor external resources (such as a domain ontology or downloaded image set) and we only rely on the surrounding text of images which is a more general setting.

Notations	Definitions
v	Subscript indicating the visual part of an item
t	Subscript indicating the textual part of an item
$q = (q_v, q_t)$	Multimedia query (which reduces to $q = (q_t)$ in the asymmetric search scenario)
$d = (d_v, d_t)$	Multimedia object in the database
n	The number of multimedia objects or documents in the database
$s_v(q, .)$	Visual similarities (row) vector of q with documents of the database
$s_t(q, .)$	Textual similarities (row) vector of q with documents of the database
l	The number of top elements retained from s_t for semantic filtering
S_v	Visual similarity (square) matrix between pairs of documents
S_t	Textual similarity (square) matrix between pairs of documents
$\mathbf{K}(., k)$	k nearest neighbor thresholding operator acting on a vector
$x_{(i)}$	Similarity diffusion process starting from the text modality
$y_{(i)}$	Similarity diffusion process starting from the visual modality
cm_{tv}, cm_{vt}	Cross-media similarities
rw_{tv}, rw_{vt}	Random walk based scores
gd_{tv}, gd_{vt}	Generalized diffusion model

Table I. Notations and definitions.

- We emphasize the transmedia principle in the diffusion process which mixes the mono-media similarity matrices and relevance scores differently from late fusion.
- Since no learning phase is required, in our case, we avoid the annotation burden and also the time complexity problem underlying such methods.

After having introduced a classification of the most used multimedia information fusion strategies, we introduce in the next section, the graph based fusion methods we are going to embed in our multimedia relevance model.

3. CROSS-MEDIA SIMILARITIES AND RANDOM WALK BASED SCORES

For convenience, we start by introducing in Table I the notations we will use in the rest of the paper. Note that we assume that the different similarities or scores are all non negative numbers. Then, we review two popular image/text graph based fusion techniques in CBMIR, namely the cross-media similarities and the random walks based scores, considering their use in the two different search scenarios we mentioned previously. Finally, we present a new graph based framework that allows generalizing these two popular techniques.

3.1 Methods based on cross-media similarities

Cross-media similarities studied in this paper refer to the research works developed in [Clinchant et al. 2007; Ah-Pine et al. 2009] and which has proven to give top-ranked retrieval results on several ImageCLEF multimedia search tasks[1] [Müller et al. 2010]. They draw inspiration from cross-media relevance models [Jeon et al. 2003] and intermedia feedback methods [Maillot et al. 2006] by generalizing the pseudo-feedback idea present in the cross-media relevance model.

We can explain the cross-media similarity mechanism using the following illustration (see also Figure 3). Given a text query q_t, we first find the most similar items in the collection with regard to the textual similarities. Then, we select pseudo-relevant objects d which are the k nearest neighbors. Next, we look at the pseudo-relevant objects' visual

[1]For more details, please visit `www.imageclef.org`

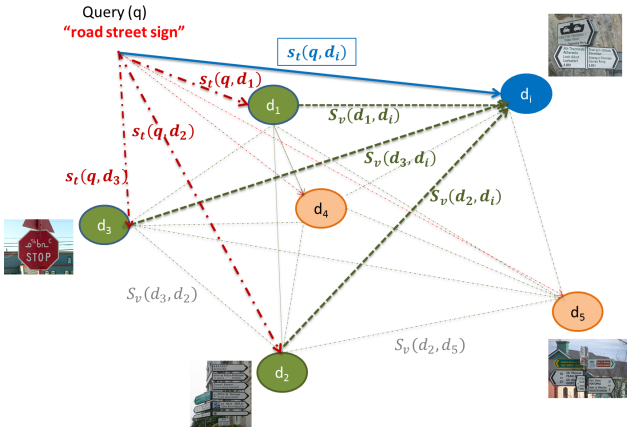

Fig. 3. Given a text query q, the cross-media relevance score can be computed as $cm_{tv}(q,d^i) = s_t(q,d^1)S_v(d^1,d^i)+s_t(q,d^2)S_v(d^2,d^i)+s_t(q,d^3)S_v(d^3,d^i)$, where d^1, d^2 and d^3 have the highest textual similarities with the query. Note that the sum is over the nearest neighbors of the query, and the complementary visual information of the documents that are close to the query are exploited.

similarities profiles $S_v(d,.)$. We then combine these visual similarity scores linearly and we obtain a cross-media similarity measure between the text query and the multimedia objects of the database. Formally such cross-media similarities are defined as follows:

$$cm_{tv}(q,.) = \mathbf{K}(s_t(q,.),k) \cdot S_v \tag{1}$$

where $\mathbf{K}(.,k)$ is an operator that takes as input a vector and gives a zero value to elements whose score is strictly lower then the k^{th} highest score and the $\cdot$ symbol represents the regular matrix multiplication operation.

The previously introduced cross-media similarity, denoted $cm_{tv}(q,.)$, propagates the text similarities of pseudo-relevant objects to their visual similarities which can be seen as a transmedia pseudo-relevance feedback mechanism. This operation is non commutative and we can design a cross-media similarity, $cm_{vt}(q,.)$, propagating visual similarities to textual similarities, providing that we are also given an image query q_v. We then obtain:

$$cm_{vt}(q,.) = \mathbf{K}(s_v(q,.),k) \cdot S_t \tag{2}$$

These cross-media similarities attempt to bridge the semantic gap between visual and textual information by enriching one modality by the other using monomedia nearest neighbors as proxies. Once the cross-media similarities are computed we can linearly combine them with monomedia similarities as follows:

$$rsv_{cm}(q,.) = \alpha_t s_t(q,.) + \alpha_v s_v(q,.) + \alpha_{tv} cm_{tv}(q,.) + \alpha_{vt} cm_{vt}(q,.) \tag{3}$$

where $\alpha_t, \alpha_v, \alpha_{tv}, \alpha_{vt}$ are real parameters that sum to one.

The formula given in Eq. 3 encompasses different particular sub-cases:

- $\alpha_{tv} = \alpha_{vt} = 0$, leads to the classical late fusion technique using a weighted mean as an aggregation function.
- $\alpha_v = \alpha_{vt} = 0$, gives a cross-media based approach to address CBMIR tasks in the context of the asymmetric search scenario.

- $\alpha_v = \alpha_{tv} = 0$, is one particular combination that gave top-ranked results on different ImageCLEF tasks [Clinchant et al. 2007; Ah-Pine et al. 2008; Ah-Pine et al. 2009].

3.2 Methods based on random walks

The PageRank algorithm proposed in [Brin and Page 1998b; Langville and Meyer 2005; Franceschet 2011] has been an important step forward in the development and success of search engines such as Google. It is therefore not surprising that multimedia information fusion based on graph modeling using random walks has been addressed by several researchers [Pan et al. 2004; Hsu et al. 2007b; Tian et al. 2008; Ma et al. 2010]. In this paper we particularly study the method proposed in [Hsu et al. 2007b; 2007a], where it is assumed that each image is a node of a graph and two images are linked with a weighted edge if there exists a multimodal contextual similarity between them. Depending on the application, the definition of such multimodal contextual similarities can vary. Typically, we assume that they are given by a linear combination of some visual and textual similarities.

The research work described in [Hsu et al. 2007b] deals with video retrieval. In the latter paper, the authors propose to use near-duplicate detection measures as for visual similarities between video stories. Text similarities are derived from automatic speech recognition and machine translation transcripts and measured by a mutual information approach.

In our perspective, we are concerned with image/text data and we assume generic image based and text based similarity matrices which are respectively denoted by S_v and S_t. Using the notations given in Table I, the multimodal contextual similarity matrix according to [Hsu et al. 2007b], that we denote by C, can be interpreted as follows:

$$C = \beta S_t + (1 - \beta) S_v \tag{4}$$

where $\beta \in [0, 1]$. We transform the latter matrix C by multiplying it with the diagonal matrix D of size $n \times n$ where:

$$D(i, i) = \frac{1}{\sum_{j=1}^{n} C(i, j)} \tag{5}$$

We obtain a stochastic matrix $P = D \cdot C$. The general term $P(i, j)$ is interpreted as the probability to go from "state" i to "state" j where these indices respectively refer to documents d^i and d^j. We then compute the random walk's stationary probability distribution over the documents and employ it to rerank the list retrieved by the text based scores. However, to further fuse visual and textual information, the random walk is biased towards documents with higher textual similarity values with the text query. In other words, we add a prior based on the text scores in the random walk process. Note that such a prior can also be interpreted as a restart process or a personalization vector in other information retrieval tasks (*e.g.* [Brin and Page 1998b; Langville and Meyer 2005]).

Formally, if we denote by $x_{(i)}$ the row vector of size $1 \times n$ of the state probabilities at iteration i and consider $\gamma \in [0, 1]$, we have:

$$x_{(i)} = (1 - \gamma) x_{(i-1)} \cdot P + \gamma s_t(q, .) \tag{6}$$

In order to obtain the state stationary distribution, we iterate the previous updating equation until convergence which yields to the following definition:

$$x_\infty = (1 - \gamma) x_\infty \cdot P + \gamma s_t(q, .) \tag{7}$$

In [Hsu et al. 2007b] only the asymmetric search scenario with a text query solely was treated. In this paper we consider the extension of this approach when we are also given an image query. Accordingly, we use a similar random walk process but with a prior depending on the initial image based scores $s_v(q,.)$ and define the related stationary distribution:

$$y_\infty = (1-\gamma) y_\infty \cdot P + \gamma s_v(q,.) \tag{8}$$

Let us denote $rw_{tv}(q,.) = x_\infty$ and $rw_{vt}(q,.) = y_\infty$. We can linearly combine these graph based scores with the initial monomedia similarities and design the following final relevance score:

$$rsv_{rw}(q,.) = \alpha_t s_t(q,.) + \alpha_v s_v(q,.) + \alpha_{tv} rw_{tv}(q,.) + \alpha_{vt} rw_{vt}(q,.) \tag{9}$$

where $\alpha_t, \alpha_v, \alpha_{tv}, \alpha_{vt}$ are real parameters that sum to one.

We can consider the following particular cases:

- $\alpha_{vt} = \alpha_{tv} = 0$, leads to the classic late fusion technique for Eq. 3.
- $\alpha_t = \alpha_v = \alpha_{vt} = 0$, is a combination that reduces to $rw_{tv}(q,.)$. It assumes the asymmetric search scenario and corresponds to the FRTP case in [Hsu et al. 2007b].
- $\alpha_v, \alpha_{vt} > 0$, is, to our knowledge, a new extension of the method which assumes the symmetric search scenario.

4. A MULTIMEDIA RETRIEVAL MODEL COMBINING TEXT QUERY BASED SEMANTIC FILTERING AND A UNIFYING GRAPH BASED FRAMEWORK

To introduce our CBMIR model, we begin with the description of the text based semantic filtering which represents the first level of image/text information fusion in our model. This step amounts to assuming that the text query is the main semantic source with regard to the user information need and it should be treated in a more specific way similarly as in the asymmetric search scenario based on the visual reranking paradigm. Secondly, we show how we can embed both the cross-media similarities and the random walk approaches in a unifying graph based model. This unifying model is then employed to rerank the semantically relevant items selected after the first step. In that perspective, our proposal emphasizes different types of strategies for fusing visual and textual similarities *via* graph based techniques.

4.1 Text query based semantic filtering of multimedia items

As argued in section 1, many multimedia retrieval systems rely on an asymmetric search scenario where only the text part of multimedia objects are indexed and the user accesses the collection by using a text query and a text search engine.

Accordingly, one first aspect of our multimedia model is to employ the text query provided by the user in order to filter the multimedia collection keeping only the most semantically relevant items. Our strategy in that regard is simple: any element of the multimedia repository that does not belong to the top l list given by the pure textual similarities is discarded. We will call that operation *semantic filtering*.

This text query based semantic filtering not only aims at selecting semantically relevant multimedia elements but it also allows decreasing the overall complexity of our graph based model[2]. Indeed, after the semantic filtering step, our system applies graph techniques

[2]It is important to mention that the price to pay using this approach is that relevant images with no or completely

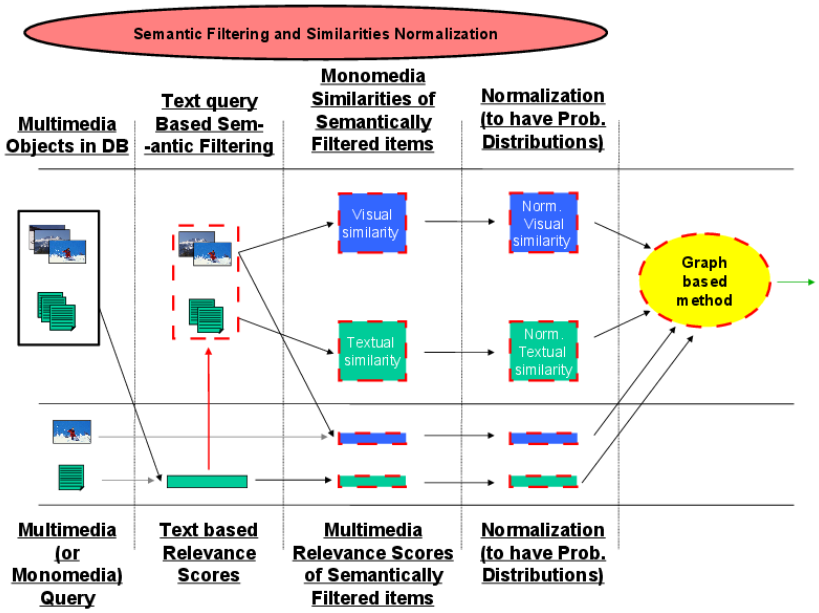

Fig. 4. Pre-processing, text query based semantic filtering and normalization.

to the l selected items only and not to the whole dataset. Since graph methods generally have a quadratic cost in terms of memory and at least a cubic computational complexity and since we assume $l \ll n$, then our model has a memory complexity reducing from $O(n^2)$ to $O(l^2)$ and a time complexity decreasing from $O(n^3)$ to $O(l^3)$.

In order to not burden the formulas, *we will not introduce new notations* to refer to the subset of top l text query based semantically filtered items. However, it is important that the reader keeps in mind that, in the sequel, all similarity vectors and matrices only involve the subset of l selected elements. In Figure 4, we depict the first feature of our multimedia retrieval model which applies the text query based semantic filtering to the query and to the multimedia items of the database.

4.2 A unifying graph based framework

The second feature of our multimedia retrieval model aims at defining a unifying framework for graph based methods that encompasses the similarity diffusion process strategies underlying both the cross-media similarities and the random walk based scores. Note that in order to embed these two approaches in the same model, we assume that all similarities have been normalized so that we manipulate probability distributions. Henceforth, we assume that $s_t(q,.)$, $s_v(q,.)$, and rows of S_t and S_v have non negative values and that they all sum to one.

This constraint is due to the random walk method but the cross-media approach does

unrelated text cannot be retrieved. Only a much more costly visual search engine could recover them. Nevertheless, our system could still be used as a first step as it would allow one to enrich the visual query and to yield an improved visual search.

not initially require such a normalization and other possibilities exist. We will come back to this point later on in section 6.3. The normalization step occurs just after the text query based semantic filtering and just before applying graph based methods as depicted in Figure 4.

To establish our unifying graph based model, let us start by studying the random walk approach a little bit deeper and let us consider the following formula:

$$x_\infty = (1-\gamma)x_\infty \cdot P + \gamma x_\infty \cdot e \cdot s_t(q,.) \tag{10}$$

where e is the $l \times 1$ vector full of 1. In the previous equation, the sub-part $x_\infty \cdot e$ reduces to 1 since x_∞ is a probability distribution. Therefore Eq. 10 and Eq. 7 are strictly equivalent. But, in Eq. 10, we can factorize the term x_∞ to obtain:

$$x_\infty = x_\infty \cdot [(1-\gamma)P + \gamma e \cdot s_t(q,.)] \tag{11}$$

Let us introduce the following matrix of size $l \times l$:

$$Q_{tv} = (1-\gamma)P + \gamma e \cdot s_t(q,.) \tag{12}$$

Using this matrix, Eq. 11 can be re-written as $x_\infty = x_\infty \cdot Q_{tv}$. The solution of this equation is the same as the solution of $(x_\infty)^\top = (Q_{tv})^\top \cdot (x_\infty)^\top$ where the right superscript $\top$ states for the transpose operation on vectors and matrices. From the latter relation we can see that the stationary probability distribution of the random walk is related to an eigen-decomposition problem [Langville and Meyer 2005]. Indeed, x_∞ is clearly the eigenvector of $(Q_{tv})^\top$ associated[3] to the eigenvalue 1. Since Q_{tv} is a stochastic matrix, 1 is the highest eigenvalue. As a result, x_∞ is the leading eigenvector of $(Q_{tv})^\top$. One efficient way to compute the leading eigenvector of a square matrix is the power method [Langville and Meyer 2005]. Thus, in practice, we iterate the following equation until convergence in order to determine $rw_{tv}(q,.)$:

$$(x_{(i)})^\top = (Q_{tv})^\top \cdot (x_{(i-1)})^\top \tag{13}$$

Since $x_{(0)}$ is a probability distribution then so are the vectors $x_{(i)}, i > 0$ and x_∞ represents the stationary distribution which is proportional to the leading eigenvector of Q_{tv}.

Let us now consider the following general formula:

$$x_{(i)} \propto \mathbf{K}(x_{(i-1)}, k) \cdot [(1-\gamma)D \cdot (\beta S_t + (1-\beta)S_v) + \gamma e \cdot s_t(q,.)] \tag{14}$$

From the previous developments, it follows that Eq. 14 embeds the random walk approach described by Eq. 11 since $rw_{tv}(q,.)$ is recovered if we use the following parameters: $x_{(0)}$ is the uniform distribution, $\beta > 0$, $\gamma > 0$, $k = l$ in the $\mathbf{K}$ nearest neighbors operator and $i = \infty$ (iterations until convergence). Indeed, by setting $\beta > 0$ and $\gamma > 0$, we recover the matrix Q_{tv} in Eq. 12. Moreover, if $k = l$ then $\mathbf{K}(x_{(i-1)}, l)$ is exactly $x_{(i-1)}$ and x_∞ is thus $rw_{tv}(q,.)$.

It turns out that the cross-media similarity approach $cm_{tv}(q,.)$ given in Eq. 1 is also a particular case of Eq. 14. Indeed, it corresponds to the following setting: $x_{(0)} = s_t(q,.)$, $\beta = 0$, $\gamma = 0$, $k < l$ in the $\mathbf{K}$ nearest neighbors operator and $i = 1$ (only one iteration). In that case, we do not use any preliminary fusion of monomedia similarities ($\beta = 0$) and we do not use a text based prior in the similarity diffusion process ($\gamma = 0$). However, the text

[3]Note that the eigenvector x_∞ is independent from x_0, for which the only constraint is to be a probability distribution, for example the uniform distribution.

relevance scores are used as the initial distribution. Furthermore, the similarity diffusion process is limited both in terms of "time" and "space" since the number of iterations is very limited ($i = 1$) and since the diffusion only relies on the nearest neighbors ($k < l$).

Eq. 14 actually combines the idea of considering only a few nearest neighbors in the similarity diffusion process as in the case of the cross-media, while doing several iterations (until stability) as in the case of the random walk.

Apart from these two particular interpretations, there is another case which is worth defining and which corresponds to the following set of parameters: $x_{(0)} = s_t(q,.)$, $\beta > 0$, $\gamma > 0$, $k < l$ in the $\mathbf{K}$ nearest neighbors operator, $i = \infty$ (iterations until convergence). This particular approach, that we call generalized diffusion and denote $gd_{tv}(q,.)$, represents an intermediary situation between $rw_{tv}(q,.)$ and $cm_{tv}(q,.)$.

Note that when $k = l$, the random walk is guaranteed to converge and the limit value does not depend on the initialisation. However, when $k < l$, we do not have a theoretical guarantee of the convergence of the generalized diffusion. Nevertheless, we have experimentally observed that after several iterations the scores $x_{(i)}$ become stable. It seems that the set of top k documents remains unchanged after a few iterations which could explain the convergence.

We have shown that Eq. 14 is a general graph based approach which generalizes both $rw_{tv}(q,.)$ and $cm_{tv}(q,.)$ methods. Similarly, we propose the following formula that allows us to generalize the symmetric relations $rw_{vt}(q,.)$ and $cm_{vt}(q,.)$:

$$y_{(i)} \propto \mathbf{K}(y_{(i-1)}, k) \cdot [(1-\gamma)D \cdot (\beta S_v + (1-\beta)S_t) + \gamma e \cdot s_v(q,.)] \qquad (15)$$

In this case, $y_{(i)}$ is related to a similarity diffusion process using an image based prior. In the right member of Eq. 15, what formally changes as compared to Eq. 14, is the substitution of t by v and *vice versa*. As stated in the introduction, this formula makes it possible to consider the symmetric search scenario that has been less investigated in the context of the random walk approach. In such a case, we also have an image query and we can thus consider using the random walk technique for multimedia fusion using an image based prior $s_v(q,.)$.

Eq. 15 generalizes $rw_{vt}(q,.)$ given by Eq. 8 by setting the parameters as followed: $y_{(0)}$ is the uniform distribution, $\beta > 0$, $\gamma > 0$, $k = l$ in the $\mathbf{K}$ nearest neighbors operator and $i = \infty$ (iterations until convergence). Similarly, Eq. 15 embeds the cross-media similarities $cm_{vt}(q,.)$ defined by Eq. 2 which is given by the following setting: $y_{(0)} = s_v(q,.)$, $\beta = 0$, $\gamma = 0$, $k < l$ in the $\mathbf{K}$ nearest neighbors operator and $i = 1$ (only one iteration).

In addition, we consider the generalized diffusion $gd_{vt}(q,.)$ defined using the following parameters values: $y_{(0)} = s_v(q,.)$, $\beta > 0$, $\gamma > 0$, $k < l$ in the $\mathbf{K}$ nearest neighbors operator, $i = \infty$ (iterations until convergence).

This unifying framework encompasses the cross-media similarities and the random walk based method for CBMIR. Eq. 14 and Eq. 15 enable us to have a better understanding of the main differences between these two techniques from a conceptual point of view. However, our proposal suggests more than a simple comparison of those two approaches, it invites to a deeper analysis of what are the key points when using graph based techniques in CBMIR. To this end, we have also defined the generalized diffusion scores $gd_{tv}(q,.)$ and $gd_{vt}(q,.)$.

We depict in Figure 5 the unified formulation of graph based approaches that we have introduced previously accompanied with the preliminary semantic filtering and normaliza-

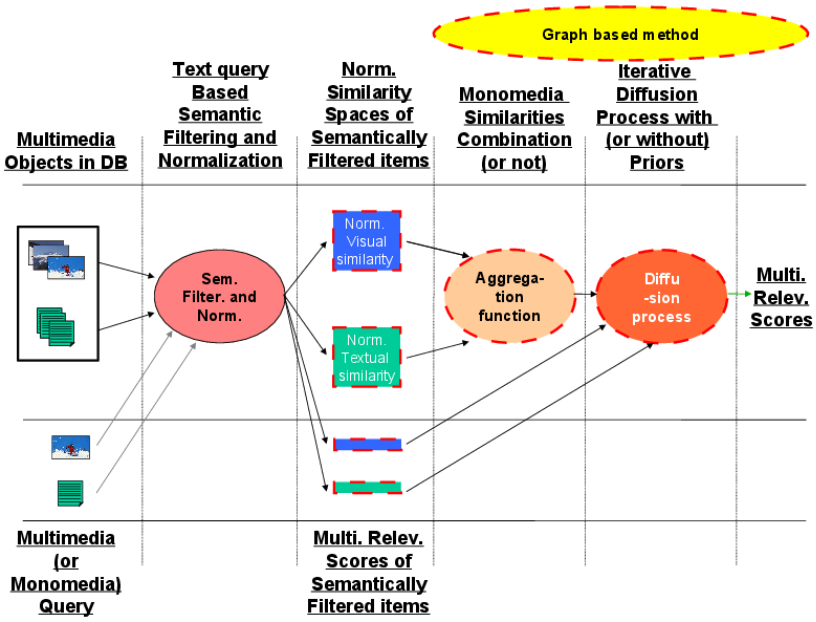

Fig. 5. Unified view of graph based methods.

tion steps. Overall, this schema represents our model to design graph based scores.

After having mixed visual and textual similarities through similarity diffusion processes using Eqs. 14 and 15, we propose to have an ultimate agregation which linearly combines monomedia similarities with the obtained graph based scores. As a result, the multimedia relevance score we propose is defined by:

$$rsv(q,.) = \alpha_t s_t(q,.) + \alpha_v s_v(q,.) + \alpha_{tv} x_{(i)}(q,.) + \alpha_{vt} y_{(i)}(q,.) \quad (16)$$

where $\alpha_t, \alpha_v, \alpha_{tv}, \alpha_{vt}$ are real numbers that sum to one and i depends on the type of diffusion process we want to use (typically $i = 1$ or $i = \infty$).

5. EXPERIMENTAL SETTINGS

Firstly, we describe the real-world datasets we applied the different techniques to. Then, we introduce the image and text representations and similarities used in our experiments.

5.1 Datasets

We conducted our experiments on real-world collections which are constituted of image/text items. The first two datasets were used in the ImageCLEF Photo or Wikipedia retrieval tasks[4] while the last one was constituted in order to assess web image search techniques[5]. We give below the description of these repositories and the tasks they were meant to address, according to the respective websites that present them.

[4] http://www.imageclef.org/datasets
[5] http://lear.inrialpes.fr/~krapac/webqueries/webqueries.html

- The IAPR dataset was used in the context of ImageCLEF 2008 [Grubinger et al. 2006]. It consists of 20,000 still natural images taken from locations around the world and comprising an assorted cross-section of still natural images. This includes pictures of different sports and actions, photographs of people, animals, cities, landscapes and many other aspects of contemporary life. Each image is associated with a text caption in up to three different languages (English, German and Spanish). It contains 60 query topics each with 3 example images.
- The Wikipedia collections WIKI10 and WIKI11 were used in the Wikipedia image retrieval task of ImageCLEF 2010 and 2011 [Popescu et al. 2010]. "The overall goal of the task is to investigate how well multi-modal image retrieval approaches that combine textual and visual evidence in order to satisfy a users multimedia information need could deal with larger scale image collections that contain highly heterogeneous items both in terms of their textual descriptions and their visual content. The aim is to simulate image retrieval in a realistic setting, such as the Web environment, where available images cover highly diverse subjects and have highly varied visual properties, while their accompanying textual metadata (if any) are user-generated and correspond to noisy and unstructured textual descriptions of varying quality and length."[6]. Both collections actually contain the same set of 237,434 images, but different topics in order to take into account several kinds of multimedia information needs. WIKI10 consists of 70 topics while WIKI11 contains 50 topics with 1-5 query images. "The ground truth for these topics was created by assuming binary relevance (relevant *vs*. non relevant) and by assessing only the images in the pools created by the retrieved images contained in the runs submitted by the participants each year."
- The Web Queries (WEBQ) repository was used as a benchmark in order to assess the research work described in [Krapac et al. 2010]. "The Web Queries dataset contains 71,478 images and meta-data retrieved by 353 web queries. For each retrieved image the relevance label is available. The relevance labels are obtained by manual labeling." Unlike the previous tasks, WEBQ contains only text topics, thus it is a case of asymmetric search scenario. Note that the actual collection contains only the thumbnails of the original images and a link to the web pages. As many of those links are not available any more we could use only the visual similarities S_v between the thumbnail images and no textual similarities S_t were computed. The "text scores" s_t were set to be 1 over the provided Google rank and the filtering was explicit in this case as for each of the 353 web query only a set of top retrieved documents (thumbnail images) are given. This shows that the WEBQ setting is very different from the other ones which might explain the difference in behaviour we observe in our experiments.

Though we use three different collections, our experiments concern four tasks : IAPR, WIKI10, WIKI11 and WEBQ. The tasks are all content based image/text multimedia data retrieval ones. On each topic given in each task, we tested different particular cases of the graph based approach introduced in section 4. A topic consists of an image/text query (except for the WEBQ as explained beforehand) and we were also provided with the binary ground truth (relevant *vs*. non relevant). We used the Mean Average Precision (MAP) in order to compare the obtained rankings and the ground truth in the goal of evaluating the different multimodal fusion techniques. We also computed if the results were statistically

[6]This is the description of the dataset as provided at `http://www.imageclef.org/wikidata`

different using paired t-tests at the 95% confidence level.

5.2 Monomodal Representation and Similarities

Standard preprocessing techniques were first applied to the textual part of the documents. After stop-word removal, words were lemmatized and the collection of documents indexed with Lemur[7]. We used a standard Dirichlet language model on IAPR and the Lexical Entailment (LE) information retrieval model [Clinchant et al. 2006], briefly introduced in appendix section A, on the Wikipedia datasets. These models were chosen to remain consistent with our previously published and state-of-the-art results [Ah-Pine et al. 2010; Ah-Pine et al. 2008; Ah-Pine et al. 2010; Csurka et al. 2011; Clinchant et al. 2010]. the "text scores" $s_t(q_t, d_t)$ were set to be 1 over the provided Google rank, as not only couldn't recompute distances between the actual query and web pages, but the filtering (provided set of images) would have not correspond to the filtering we would have obtained with our textual score.

As for image representations, we used the Fisher Vector (FV), proposed in [Perronnin and Dance 2007], an extension of the popular Bag-of-Visual word (BOV) image representation [Sivic and Zisserman 2003; Csurka et al. 2004] that describes an image by a histogram of quantized local features. In a nutshell, the Fisher Vector, described in appendix section B, consists in modeling the distribution of patches in any image with a Gaussian mixture model (GMM) and then in describing an image by its deviation from this average probability distribution.

Nevertheless, for the purpose of this paper, the choices of a particular textual and visual similarity are not of first importance. Our framework only requires as input a text ranking expert and a visual ranking expert. So, any visual/textual approaches could be employed and this is why we have moved the descriptions of our experts in the appendix. Our focus here is on the combination of visual and textual modalities. In fact, we did some preliminary experiments varying the visual and/or the textual features but the behavior concerning the combination and the conclusions we could draw were the same as for the monomodal experts used in the paper. Therefore, they do not bring new insights in our experiments and this is why we did not include these results in the paper.

6. EXPERIMENTAL RESULTS

This section contains an extended empirical analysis of our multimedia retrieval model. The experiments we conducted aim at studying the generalized graph based model we have proposed in Eq. 14 and Eq. 15 and which is depicted in Figure 5. Our goal is to establish some guidelines on the combination of visual and textual information in CBMIR using graph based methods. For convenience, we mention again the two principal formulas below:

$$x_{(i)} \propto \mathbf{K}(x_{(i-1)}, k) \cdot [(1-\gamma)D \cdot (\beta S_t + (1-\beta)S_v) + \gamma e \cdot s_t]$$
$$y_{(i)} \propto \mathbf{K}(y_{(i-1)}, k) \cdot [(1-\gamma)D \cdot (\beta S_v + (1-\beta)S_t) + \gamma e \cdot s_v]$$

Eq. 14 and Eq. 15 involve five different parameters:

- k, the number of nearest neighbors in the thresholding operator $\mathbf{K}$
- i, the number of steps in the similarity diffusion process

[7] http://www.lemurproject.org/

- $x_{(0)}$ and $y_{(0)}$, the initial distributions
- γ, if it is strictly positive it means that the model takes into account a prior distribution
- β, if it is strictly positive it means that the model linearly mixes the monomedia similarity matrices

We make the distinction between two types of parameters. On one hand, k and i are parameters that enable the way the similarity diffusion is performed to be monitored. On the other hand, the initial distributions, γ and β, are parameters that allow multimedia similarities to be combined in different ways. Since we have 2^5 possible combinations of parameters, we are not going to experiment with all possible cases. Instead we will conduct our experiments in order to underline the main differences between the underlying assumptions of cross-media similarities and of the random walk based scores.

In that regard, note that the cross-media similarities assume $k < l$ and $i = 1$ unlike the random walk technique which corresponds to the case $k = l$ and $i = \infty$. Therefore, the similarity diffusion processes assumed by the two techniques are opposite to each other. One interesting intermediate situation is given by the generalized diffusion process which uses $k < l$ and $i = \infty$. We did not formally define the case $k = l$ and $i = 1$ since this combination gave non optimal results after some preliminary tests.

Another dimension of our experiments study the impact of the second kind of parameters on the similarity diffusion process. Our aim is to examine if some strategies for mixing visual and textual similarities are beneficial or not with respect to the different ways the diffusion is carried out. In other words, is it useful to set $\gamma > 0$ and/or $\beta > 0$?

Eventually, we are interested in combining monomedia scores with graph based scores in an ultimate relevance score as described in Eq. 16.

6.1 Analysing the impact of different parameters of the proposed models

6.1.1 *Impact of the initialisations $x_{(0)}$ and $y_{(0)}$.* First of all, we comment on the impact of $x_{(0)}$ and $y_{(0)}$ on the similarity diffusion process. We typically assume two cases: either they are set to uniform distributions as suggested by random walk based approaches, or they are respectively set to $s_t(q,.)$ and $s_v(q,.)$ in order to apply the transmedia principle. Concerning, random walk based techniques which assume $i = \infty$ and $k = l$, this choice does not really matter because whatever the initialisation, the graph based scores eventually converge to the stationary distributions. On the contrary, for cross-media similarities which advocate for one step walks $i = 1$ and nearest neighbors $k < l$, this choice has an impact and preliminary results showed that search results were much better with $x_{(0)} = s_t(q,.)$ and $y_{(0)} = s_v(q,.)$. The generalized diffusion process which uses $i = \infty$ and $k < l$ present similar outcomes as cross-media similarities. As a consequence, we use $x_{(0)} = s_t(q,.)$ and $y_{(0)} = s_v(q,.)$ in the rest of our experiments. Note that after having used the text query based semantic filtering as a first level of combination between visual and textual information, this type of initialisation implies a second level of multimedia information combination.

6.1.2 *Impact of i and k when $\gamma = \beta = 0$.* In the first set of experiments, we use $i \in \{1, 2, 5, 10, 50, \infty\}$ and $k \in \{l, k^*\}$ where k^* is the best k value among the set $\{1, \ldots, l\}$ that we obtained after the first iteration and which was used in the subsequent iterations of the similarity diffusion process. We also set $\gamma = 0$ (no prior) and $\beta = 0$ (no late fusion of similarity matrices).

	IAPR		WIKI10		WIKI11		WEBQ
	s_v	s_t	s_v	s_t	s_v	s_t	s_t
	27.6	26.3	24	26.3	18	27.8	57
i	$y_{(i)}$	$x_{(i)}$	$y_{(i)}$	$x_{(i)}$	$y_{(i)}$	$x_{(i)}$	$x_{(i)}$
1	**28.7**	**20.8**	**18.9**	**15.7**	**12.6**	**6.9**	69.3
2	23.4†	17.5†	17.2†	13.8†	11.4†	5.3†	**69.5**†
5	18.7	12.3	17	13.5	11.3	5.1	68.4
10	16.8	9.6	17	13.5	11.3	5.1	68.4
50	15.4	8.5	17	13.5	11.3	5.1	68.4
∞	15.4	8.4	17	13.5	11.3	5.1	68.4

Table II. Varying the number of iterations i. Results obtained with $k = l$ and $\gamma = \beta = 0$. The symbol † indicates a statistical difference between $i = 1$ and $i = 2$ (which implies a statistical difference between $i = 1$ and $i > 1$).

	IAPR		WIKI10		WIKI11		WEBQ
	s_v	s_t	s_v	s_t	s_v	s_t	s_t
	27.6	26.3	24	26.3	18	27.8	57
i	$y_{(i)}$	$x_{(i)}$	$y_{(i)}$	$x_{(i)}$	$y_{(i)}$	$x_{(i)}$	$x_{(i)}$
1	**35.9**	**22.4**	**25.7**	**23.9**	**21.4**	**22.5**	69.3
2	32.5†	20.5†	23.9†	21.3†	19.1†	18.3†	**69.7**†
5	31.9	17.5	22.6	19.6	18.3	14.5	68.8
10	31.7	16.3	22.3	19.2	18.7	14.1	68.8
50	31.7	15.8	22.1	19.1	18.6	14	68.8
∞	31.6	15.2	22.1	19.1	18.6	14	68.8

Table III. Varying the number of iterations i. Results obtained with k^* (best k obtained after the first step $i = 1$) and $\gamma = \beta = 0$. The symbol † indicates a statistical difference between $i = 1$ and $i = 2$ (which implies a statistical difference between $i = 1$ and $i > 1$).

Our first goal is to examine the hypothesis of short versus long walks jointly with the impact of the nearest neighbor operator $\mathbf{K}$ in the similarity diffusion process. We use the initialisation discussed previously but we do not take into account any extra combination between visual and textual similarities for the moment. Accordingly, in Table II we show the MAP results we obtained when we set $k = l$ (no nearest neighbor operator) whereas in Table III, the evaluation measures are shown for k^*.

The comparison between Tables II and III leads us to the following observations:

- Whether $k = l$ or $k = k^*$, iterating the graph based formulas until convergence does not improve[8] the results in terms of MAP. For most tasks, $i = 1$ gives the best results. In the case of WEBQ, best performances were reached with $i = 2$.
- The results obtained with $k = k^*$ are much better than the ones given with $k = l$. In other words, in this current setting, it is better to make the similarity diffusion process rely on the nearest neighbors rather than all items of the collection.
- It is important to observe that k^*, which correspond to the best k after the first iteration, were in a rather small range (between 10 and 50) for all tasks except for WEBQ where k^* was a little bit greater.

[8]These results are to be contrasted with the ones obtained in [Craswell and Szummer 2007], where the authors found benefits in using long walks. However, the tasks addressed in [Craswell and Szummer 2007] are different since the graphs they deal with were sparse.

		IAPR		WIKI10		WIKI11		WEBQ
		s_v	s_t	s_v	s_t	s_v	s_t	s_t
		27.6	26.3	24	26.3	18	27.8	57
γ	k	$y_{(i)}$	$x_{(i)}$	$y_{(i)}$	$x_{(i)}$	$y_{(i)}$	$x_{(i)}$	$x_{(i)}$
0	10	35.5	19.3†	24†	23.5†	19.9†	22.5†	66.1†
0.3	10	36.5	25	28.1	**29.9**	22.9	**31**	66.2
γ^*	10	**36.6**	26.9*	**28.2**	**29.9**	23.1	**31**	67.4*
0	30	34.9	22.2†	25.7†	23.7†	21.4†	19.9†	68.3†
0.3	30	35.9	26.4	27.9	29.8	23.2	30.6	66
γ^*	30	35.9	**27.1***	28	29.8	23.2	30.7	67.9*
0	50	33.6	22.3†	25.7†	23†	21.3†	16.4†	68.7†
0.3	50	35.1	26.6	28	29.8	**23.3**	30.2	66
γ^*	50	35.1	26.9	**28.2**	**29.9**	**23.3**	30.2	68.2*
0	l	28.7†	20.8†	18.9†	15.7†	12.6†	6.9†	**69.3**†
0.3	l	33.4	26.6	25.9	28.3	19.6	27.8	66.1
γ^*	l	33.4	26.6	25.9	28.3	19.9	28.2	68.7*

Table IV. Varying the weight of the prior γ and the number of nearest neighbors k when $i = 1$. Results are shown for $i = 1$, $k \in \{10, 30, 50, l\}$ and $\gamma \in \{0, 0.3, \gamma^*\}$, where γ^* was the best γ found in the set $\{0.1, 0.2, , \cdots, 0.9\}$. Adding a prior always leads to significantly better results, as also shown by the symbol † indicating a statistical difference between $\gamma = 0.3$ (often best or close to best) and $\gamma = 0$. In contrast, there is rarely a statistical difference between $\gamma = 0.3$ and γ^* indicated by the symbol $\star$. Finally, if there is a statistical difference between $k = 10$ and other k values the results of $k > 10$ are colored in magenta.

- Assuming $k = k^*$, the cases $i = 1$ and $i = \infty$ respectively correspond to the cross-media similarities and the generalized diffusion process. We observe that, in this current setting, the former approach dominates the latter strategy.

When setting $\gamma = \beta = 0$, our experiments show that cross-media similarities ($i = 1$ and $k = k^*$) perform the best, then the generalized diffusion process ($i = \infty$ and $k = k^*$) and finally the random walk approach ($i = \infty$ and $k = l$). But, as mentioned previously, this first set of experiments does not exploit the different ways for mixing multimedia information that our generalized graph based model suggest. Therefore, in the next set of experiments, we study the impact of γ in our model.

6.1.3 *Impact of i and k when $\gamma > 0$ and $\beta = 0$.* In the second set of experiments, we use $i \in \{1, \infty\}$, $k \in \{10, 30, 50, l\}$ and $\beta = 0$. However, we now examine the hypothesis $\gamma = 0$ against $\gamma > 0$, *i.e.* where we assume a prior distribution in the similarity diffusion process. Note that in Eq. 14, the prior is the text based relevance score $s_t(q, .)$ while the similarity diffusion process is carried out using the image based similarity matrix S_v. In Eq. 15 it is the opposite. Thus, by setting $\gamma > 0$, we proceed to a third level of multimedia information fusion. Our goal is to study in what setting this fusion is beneficial.

To this end, we show in Table IV the results we obtain when γ and k vary together but assuming short walks with $i = 1$. By contrast, in Table V, we assume long walks with $i = \infty$ and we provide the MAP results with different combinations of γ and k values. In both tables, we tested with all values $\gamma \in \{0.1, 0.2, \ldots, 0.9\}$ but we only show the results for $\gamma = 0.3$ and $\gamma = \gamma^*$, the parameter value that provided the best performance.

We can make the following observations by comparing Tables IV and V:

- Whether $i = 1$ or $i = \infty$, assuming a prior in the similarity diffusion process by setting $\gamma > 0$ is a winning strategy for all tasks except for WEBQ, and $\gamma = 0.3$ appeared to

		IAPR		WIKI10		WIKI11		WEBQ
		s_v	s_t	s_v	s_t	s_v	s_t	s_t
		27.6	26.3	24	26.3	18	27.8	57
γ	k	$y_{(i)}$	$x_{(i)}$	$y_{(i)}$	$x_{(i)}$	$y_{(i)}$	$x_{(i)}$	$x_{(i)}$
0	10	31.4†	16.6†	22.2†	20.7†	19.1†	14†	66.4
0.3	10	35.1	24.2	28.4	29.2	**22.8**	**31.3**	66
γ^*	10	**35.4**	26.8*	**28.4**	29.2	**23.6**	**31.3**	66.9*
0	30	29†	14.6†	22.1†	18.3†	18.6	10.2†	69.1
0.3	30	33.4	24.1	27.8	**29.4**	22.1	30.9	67.3
γ^*	30	34.3*	26.9*	27.8	**29.4**	22.4	30.9	69.7
0	50	29.3	15.9†	19.9†	18.1†	15.1†	8.7†	69.5†
0.3	50	30.7	24.6	26.2	**29.4**	20.3	29.4	67.7
γ^*	50	33.1	**26.7***	27.3	**29.4**	21	30	**70.4***
0	100	15.4†	8.4†	17†	13.5†	11.3†	5.1†	68.4†
0.3	100	31.9	26.4	25.3	27.9	19.1	27.6	66.6
γ^*	l	32.1	26.6	25.5	27.9	19.8	28.2	69.5*

Table V. Varying the weight of the prior γ and the number of nearest neighbors k when $i = \infty$. Results are shown for $i = \infty$, $k \in \{10, 30, 50, l\}$ and $\gamma \in \{0, 0.3, \gamma^*\}$, where γ^* was the best γ found in the set $\{0.1, 0.2, , \cdots, 0.9\}$. Adding a prior always leads to significantly better results, as also shown by the symbol † indicating a statistical difference between $\gamma = 0.3$ (often best or close to best) and $\gamma = 0$. In contrast, there is rarely a statistical difference between $\gamma = 0.3$ and γ^* indicated by the symbol $\star$ (except WEBQ where we always found that the best γ was 0.1). Finally, if there is a statistical difference between $k = 10$ and other k values the results of $k > 10$ are colored in magenta.

be a stable default value for this parameter as compared to the best value γ^* among the set $\{0.1, 0.2, \ldots, 0.9\}$. As for WEBQ, we notice some improvements over the case $\gamma = 0$ only when taking $\gamma = \gamma^* = 0.1$ and assuming long walks with $i = \infty$. However, taking $\gamma > 0$ with $i = 1$ does not hurt much the performances for WEBQ.

- Regarding the nearest neighbor parameter, we notice that $k = 10$ gives the best or near-best performances whatever the value of γ. When it does not lead to the best results, it is often not statistically different from the latter scores. This result reinforces the previous observation that it is better to use nearest neighbors as proxies in the similarity diffusion process and $k = 10$ seems to be a good default parameter value in that respect.

- Regarding the number of iterations, the conclusions are less simple to make when $\gamma > 0$ in comparison with the previous case (when $\gamma = 0$). The MAP values shown in the two tables are close to each other. However, we can assume a light advantage for $i = 1$ for 3 out of 4 tasks. The WEBQ task gets better results when $i = \infty$ whereas all other tasks have similar or better MAP values when $i = 1$.

- To further compare $i = 1$ and $i = \infty$, we can restrict ourselves to the case $k < l$. In that case, we compare the cross-media similarities with the generalized diffusion process. It is interesting to mention that without any prior ($\gamma = 0$), the search performances are much better for $i = 1$. This highlights the fact that iterating the similarity diffusion process until convergence gives good results only when we consider a prior distribution. This is in line with [Hsu et al. 2007b] which underlines the importance of adding a prior in the random walk technique ($i = \infty$) in order to avoid the similarity

diffusion process getting trapped in local sub-optimal solutions[9].

If we add a prior in the similarity diffusion process by setting $\gamma > 0$, our experimental results show that cross-media similarities ($i = 1$, $k = 10$ and $\gamma = 0.3$) and the generalized diffusion process ($i = \infty$, $k = 10$, $\gamma = 0.3$) generally perform better than the random walk approach ($i = \infty$, $k = l$ and $\gamma = 0.3$). The two first strategies provide similar performances but we advocate for $i = 1$, $k = 10$ and $\gamma = 0.3$ because this approach enables the memory and time complexities to be reduced dramatically as we shall discuss later on in paragraph 6.3.

In the next paragraph, we investigate another possible way to mix visual and textual similarities in our generalized graph based framework by looking at the impact of the parameter β.

6.1.4 *Impact of i and k when $\gamma > 0$ and $\beta > 0$.* According to the previous experiments, we use $i \in \{1, \infty\}$, $k \in \{10, l\}$ and $\gamma = 0.3$, in the third set of experiments. We now focus on the parameter β and our objective is to test the hypothesis $\beta = 0$ against $\beta > 0$. By taking $\beta > 0$ we linearly combine the text query based semantically filtered similarity matrices S_v and S_t before applying the similarity diffusion process. This amounts to a fourth level of multimedia information fusion. However, unlike the previous cases, this fusion acts on the similarities between documents of the database and does not use the relevance scores except for the semantic filtering.

We test different values of β among the set $\{0.1, 0.2, \ldots, 0.9\}$ and in Table VI we show the MAP results obtained with $\beta = 0.5$, $\beta = 1$ and $\beta = \beta^*$. The latter case corresponds to the value which led to the best results. Setting $\beta = 1$ amounts to performing a pseudo-relevance feedback approach but using only one type of media instead of applying the transmedia principle. The case $\beta = 0.5$ corresponds to the simple arithmetic mean between S_v and S_t which assumes a uniform weight between both media. Note that we do not show the results for the WEBQ task because as we explained in section 5 we do no have the textual similarity matrix S_t in this case.

From Table VI we can have the following observations:

- Mixing S_v with S_t before the similarity diffusion process by setting $\beta > 0$, was significantly beneficial in only a few cases. More precisely it improves the MAP values in the case of IAPR for $x_{(i)}$ with parameters $i = 1$ and $k = 10$ (cross-media similarities) and $i = \infty$ and $k = 10$ (generalized diffusion process). Apart from IAPR, the task WIKI11 could also be improved by setting $\beta > 0$ but this time in the case of $y_{(i)}$ with parameters $i = \infty$ and 10. Note that in this latter case, the increase in the MAP value is only when $\beta = \beta^*$ while in the former case, the improvements are also valid with $\beta = 0.5$.
- The random walk approach with $i = \infty$ and $k = l$ is not improved when $\beta > 0$ since the obtained MAP scores are either lower or not statistically different from the case $\beta = 0$.
- Setting $\beta = 1$ generally hurts the performances. For $x(i)$ it correspond to the monomodal (text based) relevance feedback. This again shows the interest of exploiting both modalities using graph based techniques with a transmedia principle.

[9]Without the prior, (Q_{tv}) respectively (Q_{vt}) are independent from the query and hence are their eigenvectors.

			IAPR		WIKI10		WIKI11	
			s_v	s_t	s_v	s_t	s_v	s_t
			27.6	26.3	24	26.3	18	27.8
β	i	k	$y_{(i)}$	$x_{(i)}$	$y_{(i)}$	$x_{(i)}$	$y_{(i)}$	$x_{(i)}$
0	1	10	36.5	25	28.1	**29.9**	22.9	31
0	∞	10	35.1	24.2	28.4	29.2	22.8	**31.3**
0	∞	l	31.9	26.4	25.3	27.9	19.1	27.6
0.5	1	10	36.2	**27.3**	27.1	29.3	20.5	29.7
0.5	∞	10	34	27.2	27.3	29.2	21	29.5
0.5	∞	l	30.7	26	24.6	28	15.8	27.8
β^*	1	10	**36.6**	**27.3**	28.5	**29.9**	23	30.9
β^*	∞	10	34.9	27.2	**29.2**	29.3	**24**	31.2
β^*	∞	l	31.7	26.4	25.5	28	18.8	27.8
1	1	10	27.7	27.1	24.5	26.6	16.6	27.2
1	∞	10	26.3	25.9	24.4	26.3	15.2	27
1	∞	l	28.1	26	22.7	26.7	11.5	27.1

Table VI. Varying the weight of the matrices combination β with both i and k. Results are shown for $i \in \{1, \infty\}$, $k \in \{10, l\}$, $\gamma = 0.3$ and $\beta \in \{0, 0.5, 1, \beta^*\}$ where β^* was the best β found in the set $\{0.1, 0.2, \ldots, 0.9\}$. If there is a statistical difference between $\beta = 0$ and $\beta > 0$, the results of $\beta > 0$ are colored in magenta.

These experimental results do not encourage the mix of S_t and S_v in our graph based model by setting $\beta > 0$. Indeed, they indicate that limited gain could be obtained but only if we are able to detect the best β value among the set $\{0.1, 0.2, \ldots, 0.9\}$. Consequently, we state that no linear combination of S_t and S_v should be performed and this fourth level of information fusion should be discarded by setting $\beta = 0$ by default.

According to the experimental results we have presented so far, we support the use of our general graph based framework given by Eqs. 14 and 15 with the following recommendations:

- The initialisations of the similarity diffusion process should be based on the trans-media principle and should use the normalized text query based semantically filtered relevance scores. Therefore, we propose to set $x_{(0)} = s_t(q, .)$ and $y_{(0)} = s_v(q, .)$.
- The similarities should be diffused locally in terms of time which means that we recommend to carry out the stochastic process with very short walks. Accordingly, we suggest to set $i = 1$.
- Similarly, the similarities should be spread out locally in the similarity space. In other words, we suggest the propagation of similarities to rely on very few nearest neighbors only. Hence, one should apply the thresholding operator $\mathbf{K}(., k)$ with $k = 10$.
- It is also important to take into account a prior in the similarity diffusion process by adding a bias towards the text query based semantically filtered relevance scores. This strategy contributes to improve the search results. We found that $\gamma = 0.3$ was a good default setting in that regard.
- Working with mixed similarity matrices by setting $\beta > 0$ is not a good choice in general. This kind of information fusion has not proven to increase MAP scores for most of the tasks we have experimented with. Consequently, we propose to use $\beta = 0$ in our graph based model.

We show in the appendix the top retrieved results given by different methods we have tested for two different examples. These cases enable us to illustrate the advantages of

		IAPR	WIKI10	WIKI11	WEBQ
		s_t	s_t	s_t	s_t
		26.3	26.3	27.8	57
α	γ	rsv_{tv}	rsv_{tv}	rsv_{tv}	rsv_{tv}
0	$\hat{\gamma}$	$25^{\dagger}$	$29.9^{\dagger}$	$\mathbf{31}^{\dagger}$	$66.9^{\dagger}$
0.5	$\hat{\gamma}$	**27**	28.8	30	66.2
α^*	$\hat{\gamma}$	**27**	$29.9^{\star}$	$\mathbf{31}^{\star}$	$66.9^{\star}$
0	0	$19.3^{\dagger}$	$23.5^{\dagger}$	$22.5^{\dagger}$	$64.4^{\dagger}$
0.5	0	**27**	30	30.9	66.9
α^*	0	**27**	30	30.9	66.9

Table VII. Combining s_t with graph based scores. Results are shown for $i = 1$, $k = 10$, $\beta = 0$ and γ is either 0 or $\hat{\gamma}$, where $\hat{\gamma} = 0.3$ for IAPR, WIKI10, WIKI11 and $\gamma = \hat{} 0.1$ for WEBQ. We show results with $\alpha = 0$, $\alpha = 0.5$ or $\alpha = \alpha^*$ which corresponds to the best performing value in $\{0.1, 0.2, \ldots, 0.9\}$. The symbol † indicates a statistical difference between $\alpha = 0.5$ and $\alpha = 0$ and the symbol $\star$ indicates a statistical difference between $\alpha = 0.5$ and α^*. If there is a statistical difference between $\gamma = 0$ and $\gamma = \hat{\gamma}$, the results of $\hat{\gamma}$ are colored in magenta.

some parameter values over other ones.

6.2 Linear combination with s_t and s_v

In the next set of experiments, we analyze the linear combination between text query based semantic filtering relevance scores $s_t(q, .)$ and $s_v(q, .)$ on the one hand and graph based scores $x_{(i)}$ and $y_{(i)}$ on the other hand. In the latter case, we respectively use Eqs. 14 and 15 with $i = 1$, $k = 10$, $\beta = 0$ and $\gamma = 0.3$ for IAPR, WIKI 10 and WIKI11 respectively $\gamma = 0.1$ for WEBQ. The obtained graph based scores will be respectively denoted by $x_{(1)}$ and $y_{(1)}$ in the sequel.

6.2.1 *The asymmetric search scenario with a text query only.* We place ourselves in the asymmetric search scenario where it is supposed that the user only uses a text query. In that perspective, varying $\alpha \in \{0.1, 0.2, \ldots, 0.9\}$, we examine the following case:

$$rsv_{tv}(q, .) = \alpha s_t(q, .) + (1 - \alpha)x_{(1)} \tag{17}$$

Table VII the results we obtained for $\alpha = 0$, $\alpha = 0.5$ and for α^* which corresponds to the best performing combination. These results indicate that combining the text based relevance scores with the graph based scores using the parameter setting we discussed previously, does not bring a performance gain in terms of MAP values except for the IAPR task. The reason is that $s_t(q, .)$ already contributed to the graph based score through the initialisation setting ($x_{(0)} = s_t(q, .)$) and also as a prior ($\gamma > 0$). Therefore, an additional linear combination as suggested by Eq. 17 does not bring anything. In contrast, when we do not integrate a prior ($\gamma = 0$) it is important to recombine $x_{(1)}$ with $s_t(q, .)$ to obtain statistically similar performances. In addition, we can see that in this latter case uniform weighting is the best strategy. Finally, in both cases, we have significantly better results than the pure text based relevance scores $s_t(q, .)$.

6.2.2 *The symmetric search scenario with an image query and a text query.* In such a scenario we can exploit both image and text queries provided by the user. We thus study the linear combination of the two text query based semantically filtered relevance scores $s_v(q, .)$ and $s_t(q, .)$ with the two multimedia graph based scores $y_{(1)}$ and $x_{(1)}$ using Eq. 16 from section 4. We show the obtained results in Table VIII. In addition, we show the

	IAPR	WIKI10	WIKI11
$0.5s_v + 0.5s_t$	34.5	35.2	35.4
$\alpha_v^* s_v + \alpha_t^* s_t$	35.4	35.2	35.4
$rsv(q,.)$ (uniform weights)	37.3†	**36.1**	**36**
$rsv(q,.)$ (best weights)	**39.5**	**36.1**	**36**

Table VIII. Combining all relevance scores in a late fusion scheme. We show the results of rsv with uniform weights and best weights. We considered the case $i = 1$, $k = 10$, $\beta = 0$ and $\gamma = 0.3$. The symbol † indicates a statistical difference between uniform weights and tuned weights. We colored in magenta the values of the results with graph based methods where statistically better than the semantically filtered late fusion.

MAP values obtained by the late fusion of text query based semantically filtered relevance scores, $\alpha_v s_v(q,.) + \alpha_t s_t(q,.)$, which represents our baseline.

From this table, we can claim that in a symmetric search scenario, combining $s_t(q,.)$ and $s_v(q,.)$ with graph based scores $x_{(i)}$ and $y_{(i)}$ is beneficial since the MAP values increase in comparison with the baseline. In this case, the graph based measures are thus complementary to text query based semantic filtered relevance scores and yields improved retrieval performances.

6.3 Discussions about complexity and normalization of similarities

Our multimedia model detailed in section 4 generalizes cross-media similarities and random walk based scores, two popular graph based techniques that we have discussed in section 3. In order to compare, the two methods within our framework, we had to normalize all similarity profiles so as to manipulate probability distributions. This constraint is due to the stochastic nature of the random walk approach. Thereby, all the experimental results we have presented so far, have employed such a normalization.

Using this normalization, we found that if we mix visual and textual information in different but complementary manners by (i) applying the text query based semantic filtering, (ii) setting $x_{(0)} = s_t(q,.)$, $y_{(0)} = s_v(q,.)$ and (iii) setting $\gamma > 0$ (typically $\gamma = 3$), we obtain the best or near-best performances with the similarity diffusion process using $i = 1$ and $k = 10$. Furthermore, we demonstrate in the previous paragraph that these obtained graph based scores, $x_{(1)}$ and $y_{(1)}$, are complementary to $s_t(q,.)$ and $s_v(q,.)$ through a final linear combination given by Eq. 16.

This setting not only performs better on average, but it also enables the memory and time complexities to decrease dramatically. Firstly, the text query based semantic filtering step reduces the search space from n to $l \ll n$ documents[10]. In that case, the memory complexity decreases from $O(n^2)$ to $O(l^2)$.

Secondly, since $i = 1$, we iterate Eqs. 14 and 15 only once, this implies a time computation for the similarity diffusion process that reduces from $O(l^3)$ (in worst cases we need l iterations before stability) to $O(l^2)$. Moreover, since $k < l$, we only exploit the nearest neighbors and this further decreases the computation cost. Indeed, finding the k nearest neighbors requires $0(l \log(l))$ operations and spreading the similarity of k items requires $O(kl)$ operations. If we assume $l = 1,000$ and $k = 10$ then the overall cost for computing Eqs. 14, 15 and 16, has a $O(kl)$ complexity.

Another interesting fact to mention in regard to using $i = 1$, is that we do not need to ma-

[10]Typically $l = 1,000$ and n can be very large. The largest collection in our experiments is the Wikipedia task which contains 237,434 multimedia items.

	IAPR	WIKI10	WIKI11	WEBQ
$\alpha_v^* s_v + \alpha_t^* s_t$	35.4	35.2	35.4	57
$rsv(q,.)$ (best weights)	39.5	36.1	**36**	69.6
$rsv(q,.)$ (other normalization and uniform weights)	**40.2**	**36.2**	35.6	**70.7**

Table IX. Results with different score normalization using $i = 1$, $k = 10$, $\beta = 0$ and $\gamma = 0.3$. Note that for WEBQ we have shown results with the best performing parameters obtained without normalization ($i = 1$, $k = 50$, $\beta = 0$ and $\gamma = 0.3$) to show that effect of the normalization remains valid even if we fine tune the parameters.

nipulate probability distributions anymore. In that case, we could normalize the similarities differently from what we have done so far. We examine this hypothesis below. We computed the different graph based scores using another normalization: in the normalization step given in Figure 4, we replaced each relevance score or similarity $s(d,d')$ by $(s(d,d') - \min\{s(d,.)\})/(\max\{s(d,.)\} - \min\{s(d,.)\})$ instead of $(s(d,d') - \min\{s(d,.)\})/\sum_{d'} (s(d,d') - \min\{s(d,.)\})$. We show in the last row of Table IX, the performances of Eq. 16 with this other normalization procedure (all other parameters being equal). It is noteworthy to observe that this other normalization can outperform the previous best results in most cases. As a consequence, this suggests that our study of graph based methods could be further enriched with the impact of other kinds of normalization methods.

7. CONCLUSION

We have addressed the problem of multimedia information fusion in CBMIR and compared the cross-media similarities to the random walk methods. First of all, we have proposed an unifying framework that integrates the text query based semantic filtering on the one hand, and which generalizes both graph based techniques, on the other hand.

Furthermore, we have extensively studied many factors that impact the performances of these graph based methods. One of our goals was to provide some guidelines on how to best use those methods for two different multimodal search scenarios: the asymmetric and the symmetric cases. Our findings have been validated on three real-world datasets which are public and accessible to the research community.

All in all, we can summarize our findings about graph based methods as follows:

- Cross-media similarities and random walk based approaches can be seamlessly embedded into an unifying framework. The latter general graph based method is defined by Eq. 14 and Eq. 15 and it also allow one to take into account both the asymmetric and the symmetric search scenarios. The unifying graph based model exhibits transmedia diffusion processes with or without priors and bring to light the main differences between the two types of methods under study. But in a more general scope, it allows us to formalize the interesting features and parameters one should pay attention to when using an unsupervised graph based approach in content based image/text multimedia retrieval tasks.
- The experiments we conducted globally show that cross-media oriented diffusion processes ($i = 1$, $k < l$) outperform random walk based methods ($i = \infty$, $k = l$). Besides, the parameter γ suggested in the random walk method that the cross-media approach did not take into account has a positive impact in the latter model. Typically, we claim that the default setting for Eq. 14 and Eq. 15 should be:
  - $k = 10$ (a few nearest neighbors as proxies is better)

- $i = 1$ (one iteration is sufficient)
- $x_{(0)} = s_t(q, .)$ and $y_{(0)} = s_v(q, .)$ (the initialisation should rely on text query based semantic filtered relevance scores and a transmedia principle)
- $\gamma = 0.3$ (using a prior helps)
- $\beta = 0$ (no late fusion between similarity matrices).

- We have shown that the graph based scores using the previous setting are complementary to $s_t(q, .)$ and $s_v(q, .)$. A final linear combination between all these scores as suggested by Eq. 16 provide even better MAP values.
- We have also discussed the computational costs of our findings and we have shown that the suggested setting of our multimedia retrieval model can tackle large multimedia repositories in a scalable way.

The main limitation of our model is the importance it gives to the text based relevance scores since they have a strong impact on the search space because of the text query based semantic filtering step. Indeed, one particular drawback of our approach is its inability to retrieve relevant images which has no text description or whose associated text is very noisy. We could mitigate this problem in the symmetric search scenario, by applying a very performing (but costly) visual search given the image query. Indeed, if we can afford this step, the set obtained with the text query based semantic filtering could be enriched with the top results obtained from this visual search. Nevertheless, this strategy has to be handled carefully since even state-of-the-art unsupervised visual similarities do not perform semantic search or content-based retrieval perfectly. To illustrate this point, let us mention that pure visual similarities gave MAP measures of 6.2% and 2.7% for the tasks WIKI10 and WIKI11 respectively. Therefore, if we integrate the top list given by the image based retrieval systems, we may also add irrelevant multimedia items in our search space and thus propagate noise in our search results by performing the similarity diffusion process. In that case, the loss of performance could be greater than the gain. We conducted some preliminary tests that showed such kinds of behavior. Therefore, despite its inherent drawbacks, we still advocate for a semantic filtering step based on textual similarities in any CBMIR system, not only because it makes it possible to dramatically decrease the memory and the time computation costs but also because it enables a CBMIR system to bridge the semantic gap more effectively.

A. TEXT REPRESENTATION AND SIMILARITIES

We describe here the Lexical Entailment (LE) model used on the Wikipedia dataset as it is a less well-known model. [Berger and Lafferty 1999] addressed the problem of IR as a statistical translation problem with the well-known noisy channel model. This model can be viewed as a probabilistic version of the generalized vector space model. The analogy with the noisy channel is the following one: to generate a query word, a word is first generated from a document and this word then gets "corrupted" into a query word. The key mechanism of this model is the probability $p(v|u)$ that term u is "translated" by term v. These probabilities enable us to address a vocabulary mismatch, and also some kinds of semantic enrichments.

The problem lies in the estimation of such probability models. We refer here to a previous work [Clinchant et al. 2006] on LE models to estimate the probability that one term entails another but similar approach was proposed recently in [Karimzadehgan and Zhai

2010]. It can be understood as a probabilistic term similarity or as a unigram LM associated to a word (rather than to a document or a query). Let u be a term in the corpus, then LE models compute a probability distribution over terms v of the corpus denoted by $p(v|u)$. These probabilities can be used in IR models to enrich queries and/or documents and to give a similar effect to the use of a semantic thesaurus. However, LE is purely automatic, as statistical relationships are only extracted once from the considered corpus. In practice, a sparse representation of $p(v|u)$ is adopted, where we restrict v to be one of the 10 terms that are the closest to u using an information gain metric.

More formally, an entailment or similarity between words, expressed by a conditional probability $p(v|u)$, can be used to rank documents according to the following formula:

$$s_t(d_t, d'_t) = p(d_t|d'_t) = \prod_{v \in d_t} \sum_u p(v|u)p(u|d'_t). \tag{18}$$

where d_t (or q_t) and d'_t are two texts, $p(u|v)$ may be obtained by any of the methods described in [Clinchant et al. 2006] and $p(u|d'_t)$ is the LM of d'_t.

Note that this model give substantial improvements compared to standard retrieval models (language models, divergence from randomness, information models). For instance, the LE model obtains a MAP of 26.3% compared to 22.6% on the 2010 Wikipedia dataset. Similarly, on the 2011 dataset, the LE MAP is 27.82% compared to a 24.3% an information based model [Clinchant and Gaussier 2010].

B. IMAGE REPRESENTATION AND SIMILARITIES

As for image representations, we used the Fisher Vector (FV), proposed in [Perronnin and Dance 2007], an extension of the popular Bag-of-Visual word (BOV) image representation [Sivic and Zisserman 2003; Csurka et al. 2004] which represent an image as histogram of quantized local features. The Fisher Vector, similarly to the BOV, is based on an intermediate representation, the visual vocabulary, which is built on the the top of the low-level feature space. In our experiments we used two types of low-level features, the SIFT-like Orientation Histograms (ORH) and the local color (RGB) statistics (LCS) proposed in [Clinchant et al. 2007] and built an independent visual vocabulary for both of them.

The visual vocabulary was modeled by a Gaussian Mixture model (GMM) $p(u|\lambda) = \sum_{i=1}^{N} w_i \mathcal{N}(u|\mu_i, \Sigma_i)$, where $\lambda = \{w_i, \mu_i, \Sigma_i; i = 1, \ldots, N\}$ is the set of all parameters of the GMM and each Gaussian corresponds to a visual word. In the case of BOV representation, the low-level descriptors $\{u_t; t = 1, \ldots, T\}$ of an image d_v, are transformed into a high-level N dimensional descriptor, $\gamma(d_v)$, by accumulating over all low-level descriptors and for each Gaussian, the probabilities of generating a descriptor:

$$\gamma(d_v) = [\sum_{t=1}^{T} \gamma_1(u_t), \sum_{t=1}^{T} \gamma_2(u_t), \ldots, \sum_{t=1}^{T} \gamma_N(u_t)] \tag{19}$$

where

$$\gamma_i(u_t) = \frac{w_i \mathcal{N}(u_t|\mu_i, \Sigma_i)}{\sum_{j=1}^{N} w_j \mathcal{N}(u_t|\mu_j, \Sigma_j)}. \tag{20}$$

The Fisher Vector [Perronnin and Dance 2007] extends this BOV representation by going beyond counting measures (0-order statistics) and by encoding statistics (up to the second order) about the distribution of local descriptors assigned to each visual word. It

rather characterizes the low-level features $\{u_t\}_{t=1,...,T}$ of an image d_v by its deviation from the GMM distribution:

$$G_\lambda(d_v) = \frac{1}{T}\sum_{t=1}^{T} \nabla_\lambda \log \left\{ \sum_{j=1}^{N} w_j \mathcal{N}(u_t|\mu_j, \Sigma_j) \right\} \tag{21}$$

To compare two images d_v (or q_v) and d'_v from two multimedia documents d (or respectively the query q) and d', a natural kernel on these gradients is the Fisher Kernel [Perronnin and Dance 2007]:

$$s_v(d_v, d'_v) = G_\lambda(d_v)^\top F_\lambda^{-1} G_\lambda(d'_v), \tag{22}$$

where F_λ is the Fisher Information Matrix. As F_λ^{-1} is symmetric and positive definite, it has a Cholesky decomposition denoted by $L_\lambda^\top L_\lambda$. Therefore $s_v(d_v, d'_v)$ can be rewritten as a dot-product between normalized vectors using the mapping Γ_λ with:

$$\Gamma_\lambda(d_v) = L_\lambda \cdot G_\lambda(d_v) \tag{23}$$

which we refer to as the Fisher Vector (FV) of the image d_v.

As suggested in [Perronnin et al. 2010], we further used a square-rooted and $L2$ normalized versions of the FV and also built a spatial pyramid [Lazebnik et al. 2006]. Regarding this latter point, we repeatedly subdivide the image into 1, 3 and 4 regions: we consider the FV of the whole image (1x1); the concatenation of 3 FV extracted for the top, middle and bottom regions (1x3) and finally, the concatenation of four FV one for each quadrants (2x2). In other words, the spatial pyramid (SP) we obtained for each image considering both LCS and ORH features is given by $8 + 8 = 16$ FV. We used the dot product (linear kernel) to compute the similarity between the concatenation[11] of all FV for ORH and LCS.

C. SOME ILLUSTRATIONS OF SEARCH RESULTS OBTAINED WITH DIFFERENT GRAPH BASED SCORES

[11] Note that we do not need to explicitly concatenate all these vectors as $\langle [u, v], [u', v'] \rangle = \langle u, u' \rangle + \langle v, v' \rangle$.

WIKI10: "flying hot air balloon"

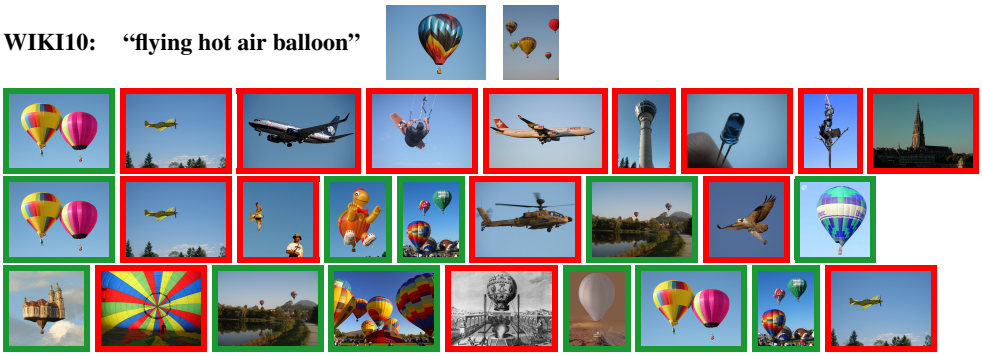

Fig. 6. Top retrieved images with pure visual similarity (second row), with text query based semantically filtered visual relevance scores s_v (third row) and with graph based scores $y_{(i)}$ using $k = 10$, $i = 1$ and $\gamma = 0$ (last row), for the topic 9 at ImageCLEF Wikipedia Challenge 2010 (shown in first row). Green means relevant, red non-relevant. Note that the first two non-relevant images in the last row are non-flying hot air balloons.

WIKI10: "lightning in the sky"

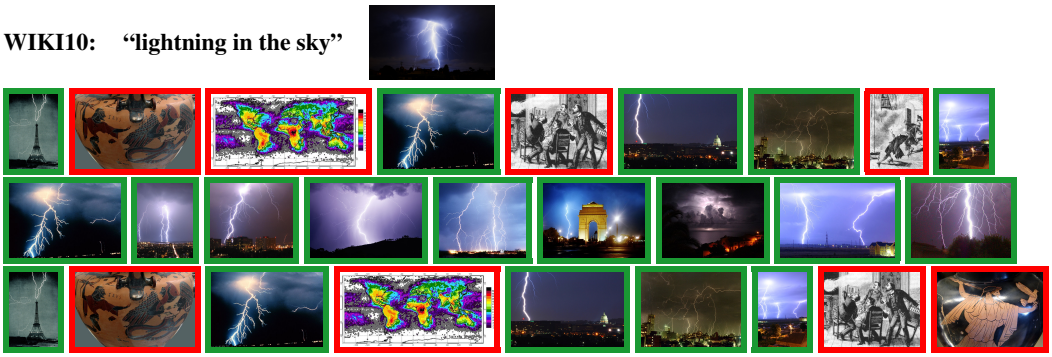

Fig. 7. Top retrieved images with textual similarity (second row), with multimedia relevance scores rsv_{tv} using $k = 10$ and $i = 1$ (third row) and with multimedia relevance scores rsv_{tv} using $k = l$ and $i = 1\infty$ (last row), for the topic 7 at ImageCLEF Wikipedia Challenge 2010 (shown in first row). Green means relevant, red non-relevant.

REFERENCES

AH-PINE, J., BRESSAN, M., CLINCHANT, S., CSURKA, G., HOPPENOT, Y., AND RENDERS, J. 2009. Crossing textual and visual content in different application scenarios. *Multimedia Tools and Applications 42,* 1, 31–56.

AH-PINE, J., CIFARELLI, C., CLINCHANT, S., CSURKA, G., AND RENDERS, J. 2008. XRCE's participation to ImageCLEF 2008. In *Working Notes of CLEF 2008*.

AH-PINE, J., CLINCHANT, S., AND CSURKA, G. 2010. Comparison of several combinations of multimodal and diversity seeking methods for multimedia retrieval. In *Multilingual Information Access Evaluation*. Lecture Notes in Computer Science (LNCS). Springer.

AH-PINE, J., CLINCHANT, S., CSURKA, G., AND LIU, Y. 2009. XRCE's participation to ImageCLEF 2009. In *Working Notes of the 2009 CLEF Workshop*.

AH-PINE, J., CLINCHANT, S., CSURKA, G., PERRONNIN, F., AND RENDERS, J.-M. 2010. *Leveraging image, text and cross-media similarities for diversity-focused multimedia retrieval*, Chapter 3.4. Volume INRE of etal Müller et al. [2010].

BERGER, A. L. AND LAFFERTY, J. D. 1999. Information retrieval as statistical translation. In *Proceedings of the*

22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 222–229.

BRIN, S. AND PAGE, L. 1998a. The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst. 30,* 1-7 (Apr.), 107–117.

BRIN, S. AND PAGE, L. 1998b. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems 30*, 107–117.

BRUNO, E., MOËNNE-LOCCOZ, N., AND MARCHAND-MAILLET, S. 2008. Design of multimodal dissimilarity spaces for retrieval of video documents. *PAMI 30,* 9, 1520–1533.

CAICEDO, J. C., MORENO, J. G., NIÑO, E. A., AND GONZÁLEZ, F. A. 2010. Combining visual features and text data for medical image retrieval using latent semantic kernels. In *Multimedia Information Retrieval*.

CLINCHANT, S., CSURKA, G., AH-PINE, J., JACQUET, G., PERRONNIN, F., SÁNCHEZ, J., AND MINOUKADEH, K. 2010. Xrce's participation in wikipedia retrieval, medical image modality classification and ad-hoc retrieval tasks of imageclef 2010. In *CLEF (Notebook Papers/LABs/Workshops)*.

CLINCHANT, S., CSURKA, G., PERRONNIN, F., AND RENDERS, J.-M. 2007. XRCE's participation to ImagEval. In *ImageEval Workshop at CVIR*.

CLINCHANT, S. AND GAUSSIER, E. 2010. Information-based models for ad hoc IR. In *SIGIR*. ACM.

CLINCHANT, S., GOUTTE, C., AND GAUSSIER, É. 2006. Lexical entailment for information retrieval. In *Advances in Information Retrieval, 28th European Conference on IR Research, ECIR 2006, London, UK, April 10-12*. 217–228.

CLINCHANT, S., RENDERS, J., AND CSURKA, G. 2007. XRCE's participation to ImageCLEF . In *CLEF Working Notes*.

CLINCHANT, S., RENDERS, J.-M., AND CSURKA, G. 2008. Trans–media pseudo–relevance feedback methods in multimedia retrieval. In *Advances in Multilingual and Multimodal Information Retrieval*. Lecture Notes in Computer Science (LNCS), vol. 5152. Springer, 569–576.

CRASWELL, N. AND SZUMMER, M. 2007. Random walks on the click graph. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. SIGIR '07. ACM, New York, NY, USA, 239–246.

CSURKA, G. AND CLINCHANT, S. 2012. An empirical study of fusion operators for multimodal image retrieval. In *CBMI*.

CSURKA, G., CLINCHANT, S., AND POPESCU, A. 2011. Xrce's participation at wikipedia retrieval of imageclef 2011. In *CLEF (Notebook Papers/Labs/Workshop)*.

CSURKA, G., DANCE, C., FAN, L., WILLAMOWSKI, J., AND BRAY, C. 2004. Visual categorization with bags of keypoints. In *ECCV Workshop on Statistical Learning for Computer Vision*.

ESCALANTE, H. J., HERNÁNDEZ, C. A., SUCAR, L. E., AND Y GÓMEZ, M. M. 2008. Late fusion of heterogeneous methods for multimedia image retrieval. In *MIR*.

FRANCESCHET, M. 2011. Pagerank: Standing on the shoulders of giants. *Communications of the ACM 54,* 6.

GAO, Y., WANG, M., ZHA, Z.-J., SHEN, J., LI, X., AND WU, X. 2013. Visual-textual joint relevance learning for tag-based social image search. *Image Processing, IEEE Transactions on 22,* 1, 363–376.

GRUBINGER, M., CLOUGH, P. D., MÜLLER, H., AND DESELAERS, T. 2006. The IAPR Benchmark: A New Evaluation Resource for Visual Information Systems. In *International Conference on Language Resources and Evaluation*. Genoa, Italy.

HSU, W. H., KENNEDY, L. S., AND CHANG, S.-F. 2006. Video search reranking via information bottleneck principle. In *ACM Multimedia*. 35–44.

HSU, W. H., KENNEDY, L. S., AND CHANG, S.-F. 2007a. Reranking methods for visual search. *IEEE MultiMedia 14,* 3, 14–22.

HSU, W. H., KENNEDY, L. S., AND CHANG, S.-F. 2007b. Video search reranking through random walk over document-level context graph. In *ACM Multimedia*. 971–980.

JARDINE, N. AND VAN RIJSBERGEN, C. 1971. The use of hierarchic clustering in information retrieval. *Information Storage and Retrieval 7,* 5, 217 – 240.

JEON, J., LAVRENKO, V., AND MANMATHA, R. 2003. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. SIGIR '03. ACM, New York, NY, USA, 119–126.

KARIMZADEHGAN, M. AND ZHAI, C. 2010. Estimation of statistical translation models based on mutual information for ad hoc information retrieval. In *SIGIR*, F. Crestani, S. Marchand-Maillet, H.-H. Chen, E. N. Efthimiadis, and J. Savoy, Eds. ACM, 323–330.

KLEINBERG, J. M. 1999. Authoritative sources in a hyperlinked environment. *J. ACM 46,* 5 (Sept.), 604–632.

KRAPAC, J., ALLAN, M., VERBEEK, J., AND JURIE, F. 2010. Improving web-image search results using query-relative classifiers. In *IEEE Conference on Computer Vision & Pattern Recognition (CVPR '10)*. IEEE Computer Society, San Francisco, United States, 1094–1101.

LANGVILLE, A. N. AND MEYER, C. D. 2005. A survey of eigenvector methods for web information retrieval. *SIAM Rev. 47,* 1 (Jan.), 135–161.

LAVRENKO, V., MANMATHA, R., AND JEON, J. 2003. A model for learning the semantics of pictures. In *NIPS*.

LAZEBNIK, S., SCHMID, C., AND PONCE, J. 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*.

LIU, T.-Y. 2009. Learning to rank for information retrieval. *Found. Trends Inf. Retr. 3,* 3 (Mar.), 225–331.

MA, H., ZHU, J., LYU, M. R., AND KING, I. 2010. Bridging the semantic gap between image contents and tags. *IEEE Transactions on Multimedia 12,* 5, 462–473.

MAGALHÃES, J. AND RÜGER, S. M. 2010. An information-theoretic framework for semantic-multimedia retrieval. *ACM Trans. Inf. Syst. 28,* 4, 19.

MAILLOT, N., CHEVALLET, J.-P., AND LIM, J.-H. 2006. Inter-media pseudo-relevance feedback application to imageclef 2006 photo retrieval. In *CLEF*, C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, M. de Rijke, and M. Stempfhuber, Eds. Lecture Notes in Computer Science, vol. 4730. Springer, 735–738.

MORI, Y., TAKAHASHI, H., AND OKA, R. 1999. Image–to–word transformation based on dividing and vector quantizing images with words.

MORIOKA, N. AND WANG, J. 2011. Robust visual reranking via sparsity and ranking constraints. In *ACM Multimedia*. 533–542.

MÜLLER, H., CLOUGH, P., DESELAERS, T., AND CAPUTO, B., Eds. 2010. *ImageCLEF- Experimental Evaluation in Visual Information Retrieval*. Vol. INRE. Springer.

NATSEV, A. P., HAUBOLD, A., TEŠIĆ, J., XIE, L., AND YAN, R. 2007. Semantic concept-based query expansion and re-ranking for multimedia retrieval. In *Proceedings of the 15th international conference on Multimedia*. MULTIMEDIA '07. ACM, New York, NY, USA, 991–1000.

PAN, J.-Y., YANG, H.-J., FALOUTSOS, C., AND DUYGULU, P. 2004. Automatic multimedia cross-modal correlation discovery. In *KDD*. 653–658.

PERRONNIN, F. AND DANCE, C. 2007. Fisher Kernels on visual vocabularies for image categorization. In *CVPR*. IEEE.

PERRONNIN, F., SÁNCHEZ, J., AND MENSINK, T. 2010. Improving the Fisher Kernel for large-scale image classification. In *ECCV*.

POPESCU, A., TSIKRIKA, T., AND KLUDAS, J. 2010. Overview Of The Wikipedia Retrieval Task At ImageCLEF 2010. In *Working notes of the 11th Workshop of the Cross-Language Evaluation Forum*. CLEF-campaign.

RASIWASIA, N., PEREIRA, J. C., COVIELLO, E., DOYLE, G., LANCKRIET, G. R. G., LEVY, R., AND VASCONCELOS, N. 2010. A new approach to cross-modal multimedia retrieval. In *ACM Multimedia*.

RODRIGUEZ-VAAMONDE, S., TORRESANI, L., AND FITZGIBBON, A. 2013. What can pictures tell us about web pages?: improving document search using images. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. SIGIR '13. ACM, New York, NY, USA, 849–852.

RUEGER, S. 2010. *Multimedia Information Retrieval*, 1st ed. Morgan and Claypool Publishers.

RUTHVEN, I. AND LALMAS, M. 2003. A survey on the use of relevance feedback for information access systems. *The Knowledge Engineering Review 18,* 02, 95–145.

SIVIC, J. S. AND ZISSERMAN, A. 2003. Video Google: A text retrieval approach to object matching in videos. In *ICCV*.

SMEATON, A. F., OVER, P., AND KRAAIJ, W. 2006. Evaluation campaigns and TRECVid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*. ACM Press, New York, NY, USA, 321–330.

SNOEK, C. G. M., WORRING, M., AND SMEULDERS, A. W. M. 2005. Early versus late fusion in semantic video analysis. In *ACM International Conference on Multimedia*. 399–402.

TIAN, X., YANG, L., WANG, J., YANG, Y., WU, X., AND HUA, X.-S. 2008. Bayesian video search reranking. In *ACM Multimedia*. 131–140.

VINOKOUROV, A., HARDOON, D. R., AND SHAWE-TAYLOR, J. 2003. Learning the semantics of multimedia content with application to web image retrieval and classification.

WANG, M., HUA, X.-S., HONG, R., TANG, J., QI, G.-J., AND SONG, Y. 2009. Unified video annotation via multigraph learning. *Circuits and Systems for Video Technology, IEEE Transactions on 19,* 5, 733–746.

WANG, M., HUA, X.-S., TANG, J., AND HONG, R. 2009. Beyond distance measurement: Constructing neighborhood similarity for video annotation. *Multimedia, IEEE Transactions on 11,* 3, 465–476.

WANG, M., LI, H., TAO, D., LU, K., AND WU, X. 2012. Multimodal graph-based reranking for web image search. *Image Processing, IEEE Transactions on 21,* 11, 4649–4661.

WANG, X.-J., MA, W.-Y., XUE, G.-R., AND LI, X. 2004. Multi-model similarity propagation and its application for web image retrieval. In *ACM Multimedia*. 944–951.

WILKINS, P., SMEATON, A. F., AND FERGUSON, P. 2010. Properties of optimally weighted data fusion in cbmir. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. SIGIR '10. ACM, New York, NY, USA, 643–650.

YANG, L. AND HANJALIC, A. 2010. Supervised reranking for web image search. In *ACM Multimedia*. 183–192.

ZHA, Z.-J., WANG, M., SHEN, J., AND CHUA, T.-S. 2012. Text mining in multimedia. In *Mining Text Data*. 361–384.