This is the accepted manuscript. The final publication is available at:
https://link.springer.com/article/10.1007%2Fs11042-017-4797-4

Noname manuscript No.
(will be inserted by the editor)

Multimedia retrieval based on non-linear graph-based fusion and Partial Least Squares Regression

Ilias Gialampoukidis · Anastasia Moumtzidou · Dimitris Liparas · Theodora Tsikrika · Stefanos Vrochidis · Ioannis Kompatsiaris

Received: date / Accepted: date

Abstract Heterogeneous sources of information, such as images, videos, text and metadata are often used to describe different or complementary views of the same multimedia object, especially in the online news domain and in large annotated image collections. The retrieval of multimedia objects, given a multimodal query, requires the combination of several sources of information in an efficient and scalable way. Towards this direction, we provide a novel unsupervised framework for multimodal fusion of visual and textual similarities, which are based on visual features, visual concepts and textual metadata, integrating non-linear graph-based fusion and Partial Least Squares Regression. The fusion strategy is based on the construction of a multimodal contextual similarity matrix and the non-linear combination of relevance scores from query-based similarity vectors. Our framework can employ more than two modalities and high-level information, without increase in memory complexity, when compared to state-of-the-art baseline methods. The experimental comparison is done in three public multimedia collections in the multimedia retrieval task. The results have shown that the proposed method outperforms the baseline methods, in terms of Mean Average Precision and Precision@20.

Keywords Multimedia retrieval · Non-linear fusion · Graph-based models

Ilias Gialampoukidis
Centre for Research and Technology-Hellas, Information Technologies Institute
6^{th} Km Charilaou-Thermi road
57001, Thessaloniki, Greece
Tel.: +30-2311-257-810
E-mail: heliasgj@iti.gr

Anastasia Moumtzidou, Dimitris Liparas, Theodora Tsikrika,
Stefanos Vrochidis, Ioannis Kompatsiaris
Centre for Research and Technology-Hellas, Information Technologies Institute
E-mail: {moumtzid,dliparas,theodora.tsikrika,stefanos,ikom}@iti.gr

1 Introduction

The abundance of multimedia content has highlighted the need to access efficiently and large and diverse multimedia collections, such as video collections (e.g. Youtube, Netflix) or annotated image collections (e.g. Facebook, Flickr). Searching in multimedia collections is a challenging problem, due to the heterogeneous sources of information, usually textual and visual, which need to be effectively combined in a scalable way. Modalities usually appear in different views, based on the nature of the features which are extracted, so the complexity increases dramatically when several modalities of multiple views appear in the multimedia collection.

Multimedia retrieval systems need to address those challenges, by means of multimodal fusion [3], either at the feature level (early fusion) or at the decision level (late fusion). Several modalities are merged into one unique source of information in order to support classic problems in supervised or unsupervised learning (eg. multimedia search, retrieval, summarization, recommendation, clustering and classification). The modalities are usually low-level visual descriptors (based on color, shape, texture, location, etc.), low-level textual features (raw text from webpages, video subtitles, or extracted from audio using automatic speech recognition, and from video using optical character recognition, etc.), metadata (time stamp, tags, source, position in a social graph) and high-level textual features [3]. All these sources of information formulate a multimedia item (multimodal object) and access to several modalities is possible through efficient multimedia indexing techniques, such as [25].

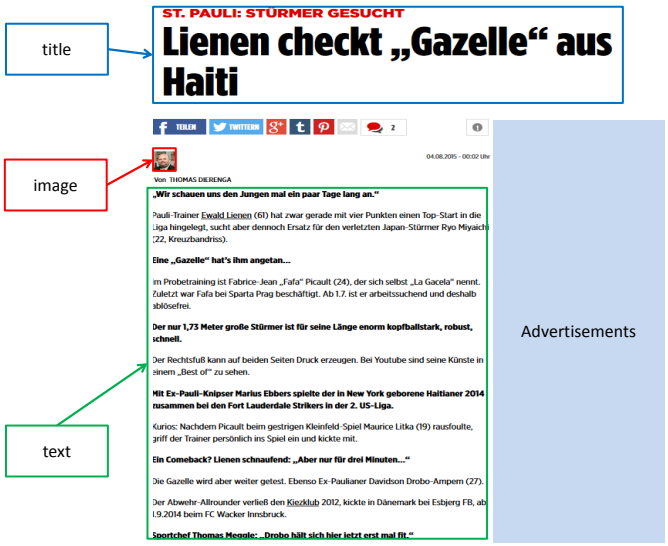


Fig. 1: A webpage with textual and visual information.

Textual and visual information is ubiquitous and the key problem is how to combine low- and high-level textual and visual information, in order to retrieve documents relevant to a given multimodal query, which also has several modalities. Towards this direction, we provide a novel framework for multimodal fusion of visual and textual similarities, which are based on visual features, visual concepts and textual concepts. Our method extends our previous work [8] using Partial Least Squares (PLS) Regression to combine multiple views of the same modality, such as SIFT descriptors and visual features based on Deep Convolution Neural Networks. The proposed method is motivated by the PLS approach [24], due to its effectiveness in multimodal hashing, and is compared to several baseline methods in unsupervised multimedia retrieval, such as weighted linear, non-linear, diffusion-based and advanced graph-based models. We also take into account complexity issues when handling several forms of data and metadata and our framework's memory complexity is comparable to a bi-modal multimedia retrieval framework. In brief, our multimedia retrieval framework:

- fuses multiple modalities, so as to retrieve multimedia objects in response to a multimodal query;
- has memory cost comparable to a bi-modal multimedia retrieval framework;
- integrates high-level and low-level information;
- uses unsupervised multimodal fusion and multimedia retrieval techniques

The structure of the paper is the following. In Section 2 we present the related work in multimodal fusion and multimedia retrieval. In Section 3 we describe the graph-based fusion techniques and the necessary background to present our framework in Section 4. In Sections 5 and 6 we evaluate the proposed multimedia retrieval framework. Finally, some concluding remarks are provided in Section 7.

2 Related Work

Over the years, many different approaches for multimedia retrieval have been introduced and compared. A critical challenge in this task is the combination of multiple heterogeneous feature sets (modalities) that can be extracted from collections of multimedia objects (e.g. low-level visual descriptors, high-level textual or visual features, etc.). The aforementioned combination process is known as multimodal fusion. An example of a study investigating multimodal fusion is the work of [22], in which a framework for video retrieval is presented. This framework extends conventional text-based search by fusing textual and visual similarity scores in a simple non-linear way. The former are obtained from video subtitles and the latter are based on visual concepts. Specifically in video retrieval systems, the possibility of exploiting user-generated relevance feedback as a way to improve video similarity has been investigated in [33].

In the context of multimodal fusion, there are three basic strategies with respect to the level, at which fusion is accomplished. The first strategy, called

early fusion, performs fusion at the feature level (e.g. [18,4]), where features from the considered modalities are combined into a common feature vector. The second strategy, known as late fusion, fuses information at the decision level, meaning that each modality is first learned separately and the individual results are aggregated into a final common decision (e.g. [34,15]). Finally, hybrid fusion aims at exploiting the advantages of both early and late fusion strategies (e.g. [16]). An overview of different studies regarding the fusion strategies described above can be found in [3]. Another interesting type of fusion is metric fusion [31], an approach aiming at fusing different "views" of the same modality, e.g. different types of low-level visual features for describing images.

Some multimedia and cross-modal retrieval studies have focused on specific methodologies. An example is the well-known Latent Dirichlet Allocation (LDA). In [32], a supervised multimodal mutual topic reinforce modeling approach for cross-media retrieval, called M3R, is proposed. Some other methodologies are Partial Least Squares (PLS) and correlation matching. With respect to the former, a PLS-based framework, which maps queries from multiple modalities to points in a common linear subspace, is introduced in [24]. Regarding the latter, [21] utilizes correlation matching between the textual and visual modalities of multimedia documents in the task of cross-modal document retrieval.

With respect to graph-based methods and random-walk approaches [2] present a unifying multimedia retrieval framework that incorporates two graph-based methods, namely cross-media similarities and random-walkbased scores. Specifically, the random-walk approach for multimodal fusion was introduced in [12], where the fusion of textual and visual information leads to improved performance in the video search task. The framework in [2] includes as special cases all well-known fusion models (e.g. early, late, diffusion-based, etc.) and does not require users' relevance feedback.

The recent advent of deep learning techniques has offered a compelling alternative to traditional approaches for solving multimedia retrieval problems. In this context, [7] makes use of deep auto-encoders to learn features from different modalities in the task of cross-modal retrieval. In a similar study, [30] propose a mapping mechanism for multimodal retrieval based on stacked auto-encoders. This mechanism learns one stacked auto-encoder for each modality in order to map the high-dimensional features into a common low-dimensional latent space. Finally, in [29], a model based on Convolutional Neural Networks (CNN) that can be used for modality-specific feature learning is introduced.

3 Graph-based fusion in Multimedia Retrieval

This section presents the necessary background in graph-based fusion for multimedia retrieval. The notation followed in this work is presented in Table 1.

Table 1: Notations and Definitions

Notations	Definitions
$\mathcal{M}$	multimedia collection
q	multimodal query
$s(q)$	fused similarity vector in response to the query q
S_m	Similarity (square) matrix for pairs of documents for the m-th modality
s_m	Query-based similarity vector for the m-th modality
$\mathbf{K}(.,k)$	k-th nearest neighbor thresholding operator acting on a similarity vector
C	multimodal contextual similarity matrix
P	row stochastic transition probability matrix
$p_{\kappa\lambda}$	transition probability between node κ and node λ
$c_{\kappa\lambda}$	the (κ,λ) element of the matrix C
$\alpha_m, \beta_m, \gamma$	parameters in [0,1]
$x_{(\infty)}, y_{(\infty)}$	steady state limiting distributions
T_m	Matrices containing the extracted latent vectors
Q_m	Matrices representing the loadings
E_m	Error matrices
W_m	Weight matrices

In multimedia retrieval, the task is to retrieve from a multimedia collection $\mathcal{M}$ a ranked list of multimedia items, relevant to a multimodal query q of M modalities. The pairwise similarities between the query q and the items of the collection $\mathcal{M}$ formulate a vector of similarity scores s_m per modality. The classic late fusion $s^w(q)$ of similarity vectors is a weighted linear combination of $s_m, m = 1, 2, \ldots, M$ [3]:

$$s^w(q) = \sum_{m=1}^{M} \alpha_m s_m \tag{1}$$

Alternative to the linear fusion method of Equation (1), the non-linear analogue has been considered in multimedia retrieval tasks [22]:

$$s^{nl}(q) = \sum_{m=1}^{M} (s_m)^{\alpha_m} \tag{2}$$

Cross-media similarities have been defined in the case of two modalities, where the similarity vector of one modality s_1 is propagated to the similarities of the other modality, formally written as:

$$s^{cm}_{1\to 2}(q) = \mathbf{K}(s_2, k) \cdot S_1 \tag{3}$$

where $\mathbf{K}(.,k)$ is the operator that takes as input a vector and gives zero value to elements whose score is strictly lower than the k^{th} highest value. The regular matrix multiplication operation is denoted by "$\cdot$". The linear combination of unimodal similarities with cross-media similarities is [5, 1]:

$$s^{cm} = \alpha_1 s_1 + \alpha_2 s_2 + \alpha_3 s^{cm}_{1\to 2} + \alpha_4 s^{cm}_{2\to 1} \tag{4}$$

under the condition that $\sum_{i=1}^{4} \alpha_i = 1$.

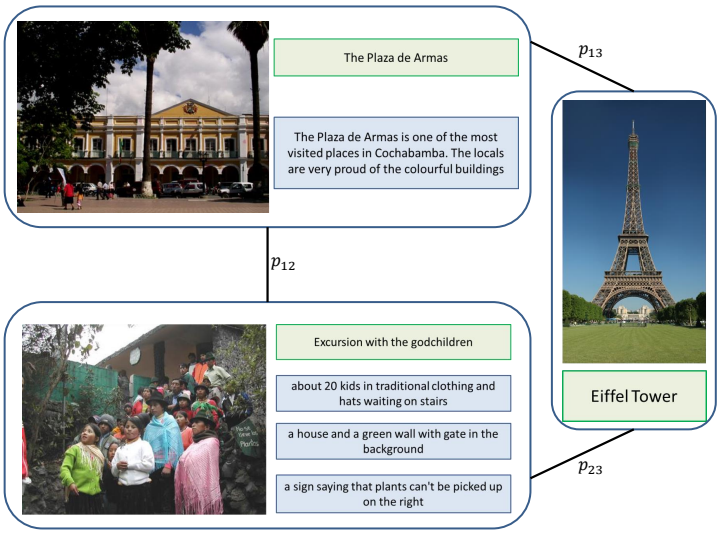


Fig. 2: A simple graph of three multimedia items and the corresponding transition probabilities $p_{\kappa\lambda}$, $\kappa, \lambda = 1, 2, 3$.

Graph-based models in multimedia retrieval rely on random walks on graphs of multimedia items. A graph is formulated by nodes, which are multimedia items from the collection $\mathcal{M}$, and all links are weighted by transition probabilities from node κ to node λ. Consider for example n multimedia items with two modalities and S_1 and S_2. A multimodal contextual similarity matrix C is defined as:

$$C = \beta_1 S_1 + \beta_2 S_2, \quad \beta_1 + \beta_2 = 1 \tag{5}$$

where $\beta_1, \beta_2 \in [0, 1]$. The matrix C becomes row stochastic P when multiplied with the diagonal matrix D of size $n \times n$, with diagonal elements $d_{\kappa\kappa} = 1/\sum_{\lambda=1}^{n} d_{\kappa\lambda}$, hence $P = D \cdot C$. The stochastic matrix P has elements $p_{\kappa\lambda}$ which are transition probabilities from multimedia item κ to item λ. The transition probability $p_{\kappa\lambda}$ between two multimodal items is also depicted in Figure 2 and provides the weight on the link from node κ to node λ. The graph-based approach has been proposed in [12] in the context of video retrieval, but is directly applicable to any pairs of modalities.

Since P is a stochastic transition probability matrix, the future evolution of a state vector $x_{(i)}$ of size $1 \times n$ is given by $x_{(i+1)} = x_{(i)} \cdot P$, having stationary (steady state) distribution $x_{(\infty)} = \lim_{i \to \infty} x_{(i)}$ after many transitions (iterations). However, the addition of a "personalization" vector [17], such as the query-based similarity vector s_1 on the textual modality [12] would introduce a perturbation towards the results of a text search:

$$x_{(i+1)} = (1 - \gamma) x_{(i)} \cdot P + \gamma s_1 \tag{6}$$

where $\gamma \in [0,1]$ and after many iterations ($i \to \infty$):

$$x_{(\infty)} = (1-\gamma)x_{(\infty)} \cdot P + \gamma s_1 \tag{7}$$

Moreover, an image is also available as a part of the multimodal query, so a similar graph-based process with a perturbation s_2 towards the results of an image search is:

$$y_{(i+1)} = (1-\gamma)y_{(i)} \cdot P + \gamma s_2 \tag{8}$$

Therefore, a random walk of i iterations on a graph of multimedia items, linearly combined with the query-based similarity vectors s_1 and s_2, provides the fused graph-based similarity score [2]:

$$s^{rw}(q) = \alpha_1 s_1 + \alpha_2 s_2 + \alpha_3 x_{(i)} + \alpha_4 y_{(i)} \tag{9}$$

under the restriction $\sum_{m=1}^{4} \alpha_m = 1$.

A unifying graph-based model has been proposed [2], combining the aforementioned approaches, in the case of two modalities:

$$\begin{aligned} x_{(i)} &\propto \mathbf{K}(x_{(i-1)}, k) \cdot [(1-\gamma)D \cdot (\beta_1 S_1 + \beta_2 S_2) + \gamma e \cdot s_1] \\ y_{(i)} &\propto \mathbf{K}(y_{(i-1)}, k) \cdot [(1-\gamma)D \cdot (\beta_1 S_2 + \beta_2 S_1) + \gamma e \cdot s_2] \end{aligned} \tag{10}$$

where e is the $l \times 1$ vector of ones, i is the number of iterations and the model sets: $x_{(0)} = s_1$ and $y_{(0)} = s_2$. The number $l < n$ is fixed, usually set to $l = 1000$ and is defined as the number of initially filtered multimedia items, with respect to the dominant modality (usually textual metadata). After the initial filtering stage, l items are left, so the similarity matrices S_1 and S_2 are $l \times l$. The final relevance score vector is given by:

$$s^{graph}(q) = \alpha_1 s_1 + \alpha_2 s_2 + \alpha_3 x_{(i)} + \alpha_4 y_{(i)} \tag{11}$$

under the restriction $\sum_{m=1}^{4} \alpha_m = 1$ and $x_{(i)}, y_{(i)}$ are given by Equation 10.

In case of a large number of iterations ($i \to \infty$), the graph-based model of Equation 11 becomes a (generalized) diffusion process [2]:

$$s^{dif}(q) = \alpha_1 s_1 + \alpha_2 s_2 + \alpha_3 x_{(\infty)} + \alpha_4 y_{(\infty)} \tag{12}$$

In Table 2 we present the linear fusion, the random walk fusion, the cross-media similarities fusion and a general diffusion process as special cases of the unifying framework of Equation (10).

Table 2: Some special cases of the unifying unsupervised fusion model of Equation (10)

Fusion Model	Reference	Equation	Conditions
Linear	[3]	(1)	$\alpha_3 = \alpha_4 = 0$
Random walk	[12]	(7)	$x_{(0)}, y_{(0)}$ uniform, $k = l, i = \infty, \beta_1 > 0, \gamma > 0,$
Cross-media	[5]	(4)	$x_{(0)} = s_1, y_{(0)} = s_2, \beta_1 = 0, \gamma = 0, k < l, i = 1$
Diffusion	[2]	(12)	$x_{(0)} = s_1, y_{(0)} = s_2, \beta_1 > 0, \gamma > 0, k < l, i = \infty$

4 A hybrid non-linear graph-based fusion with PLS Regression

In this section we present the non-linear graph-based fusion of M modalities and our adopted method for PLS Regression. Our framework combines elements from PLS Regression and non-linear graph-based fusion and is presented in Figure 3 in the general case of M modalities.

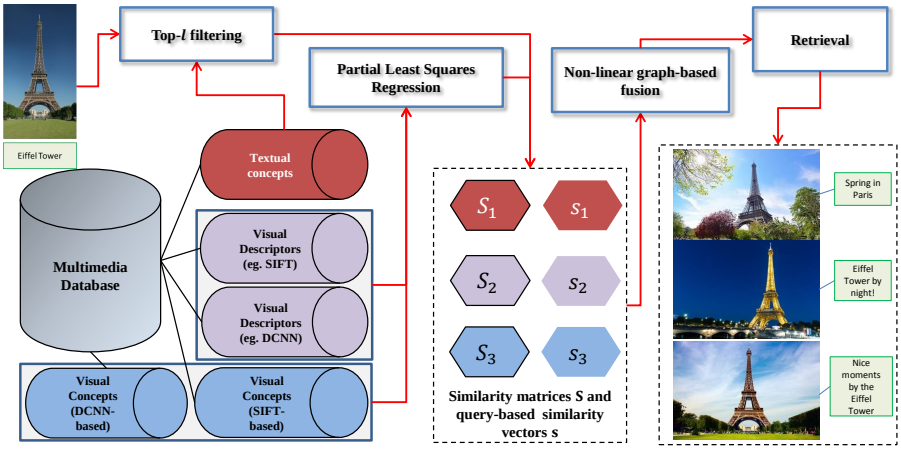


Fig. 3: A hybrid multimedia retrieval framework based on non-linear graph-based fusion and Partial Least Squares Regression.

4.1 Non-linear graph-based fusion of M modalities

The unifying model of Equation (10) has several parameters and a direct extension to M modalities is not an easy task. If we consider the general case of M modalities, the parameters α_m are M^2, the free parameters β_m are $M-1$ and the free parameters γ are $M-1$, so for a direct generalization of the graph-based model of Equation 10, the number of parameters, for M modalities, is M^2+2M-2. Even for $M=3$ modalities, the number of involved parameters is $3^2+2*3-2=13$ and for $M=4$ modalities is $4^2+2*4-2=22$. Therefore, we propose an extension to multiple modalities, in which the number of involved parameters increases with the number of modalities in a linear way, in contradiction to the quadratic increase of the form M^2+2M-2.

Given M modalities, we follow the notation of Table 1 and we initially construct a contextual similarity matrix:

$$C=\sum_{m=1}^{M}\beta_m S_m, \qquad \sum_{m=1}^{M}\beta_m=1 \tag{13}$$

The matrix C of Equation 13 is row-normalized so as to obtain the row-stochastic matrix P, such that $P = D \cdot C$, where D has diagonal elements $d_{\kappa\kappa} = 1/\sum_{\lambda=1}^{n} d_{\kappa\lambda}$. Therefore, the matrix P is row-stochastic, i.e. P has row sums equal to one: $\sum_{\lambda=1}^{n} p_{\kappa\lambda} = 1$.

The formulation of a transition probability matrix P involves a random walk approach on the graph of multimedia items, similar to Equation 5 after row-normalization. Motivated by the graph-based model of Equation 10 we set $x_{(0)}^{m} = s_m$ for all modalities $m = 1, 2, \ldots, M$, and we define the update rule:

$$x_{(i)}^{m} \propto \mathbf{K}(x_{(i-1)}^{m}, k) \cdot \left[\left(1 - \sum_{w \neq m} \gamma_w \right) P + \sum_{w \neq m} \gamma_w e \cdot s_w \right] \tag{14}$$

In our proposed graph-based model, which fuses M modalities, the vector of relevance score in response to the query q, is given by:

$$s^{l\_graph}(q) = \sum_{m=1}^{M} \alpha_m s_m + \sum_{m=1}^{M} \alpha'_m x_{(i)}^{m} \tag{15}$$

or the non-linear analogue:

$$s^{nl\_graph}(q) = \sum_{m=1}^{M} (s_m)^{\alpha_m} + \sum_{m=1}^{M} \alpha'_m x_{(i)}^{m} \tag{16}$$

under the restriction

$$\sum_{m=1}^{M} \alpha_m + \sum_{m=1}^{M} \alpha'_m = 1 \tag{17}$$

The model of Equation (16) has $M - 1$ free parameters a_m, $M - 1$ free parameters β_m and $M - 1$ free parameters γ_m, thus $3M - 3$ parameters in total, hence the increase in the number of parameters is linear in the number of modalities.

The need to extend the model of Equation (10) to multiple modalities has been highlighted in [9] and the non-linear graph-based fusion approach of Equation (16) has been presented in [8] and has been integrated in multimedia search engines [20]. In this context, we further elaborate our non-linear graph-based fusion of M modalities, combining also Partial Least Squares (PLS) Regression in the overall multimedia retrieval framework. Before the brief presentation of our adopted approach in PLS Regression, we present the non-linear graph-based fusion of $M = 3$ modalities.

$M = 3$ modalities. The contextual similarity matrix of Equation (13) becomes:

$$C = \beta_1 S_1 + \beta_2 S_2 + \beta_3 S_3, \quad \beta_1 + \beta_2 + \beta_3 = 1 \tag{18}$$

The matrix C is row normalized so as to get the corresponding transition probability matrix P as follows:

$$p_{\kappa\lambda} = \frac{c_{\kappa\lambda}}{\sum_{\lambda=1}^{l} c_{\kappa\lambda}} \tag{19}$$

The model of Equation (14) reduces to:

$$\begin{aligned} x^1_{(i)} &\propto \mathbf{K}(x^1_{(i-1)}, k) \cdot [(1-\gamma_2-\gamma_3)\, P + \gamma_2 e \cdot s_2 + \gamma_3 e \cdot s_3] \\ x^2_{(i)} &\propto \mathbf{K}(x^2_{(i-1)}, k) \cdot [(1-\gamma_1-\gamma_3)\, P + \gamma_1 e \cdot s_1 + \gamma_3 e \cdot s_3] \\ x^3_{(i)} &\propto \mathbf{K}(x^3_{(i-1)}, k) \cdot [(1-\gamma_2-\gamma_1)\, P + \gamma_2 e \cdot s_2 + \gamma_1 e \cdot s_1] \end{aligned} \tag{20}$$

The vectors of relevance scores $s^{l\_graph}(q)$ (or $s^{nl\_graph}(q)$), in response to the query q, linearly (or non-linearly) combine the similarity vectors $s_m, m = 1, 2, 3$ and the vectors $x^m_{(i)}, m = 1, 2, 3$. In the case of graph-based linear fusion:

$$s^{l\_graph}(q) = \alpha_1 s_1 + \alpha_2 s_2 + \alpha_3 s_3 + \alpha'_1 x^1_{(i)} + \alpha'_2 x^2_{(i)} + \alpha'_3 x^3_{(i)} \tag{21}$$

and in the non-linear case:

$$s^{nl\_graph}(q) = (s_1)^(\alpha_1) + (s_2)^(\alpha_2) + (s_3)^(\alpha_3) + \alpha'_1 x^1_{(i)} + \alpha'_2 x^2_{(i)} + \alpha'_3 x^3_{(i)} \tag{22}$$

4.2 Partial Least Squares Regression

The Partial Least Squares (PLS) model has been used in multimodal fusion [24] to efficiently combine two modalities. Given two matrices X and Y, PLS decomposes X and Y as follows:

$$\begin{aligned} X &= T \cdot P^T + E \\ Y &= U \cdot Q^T + F \end{aligned} \tag{23}$$

where T and U are projections of X and Y, respectively, to latent spaces containing the extracted latent vectors. P and Q are orthogonal "loading" matrices and, finally, E and F are error matrices. The aim of PLS is to maximize the covariance between T and U. PLS Regression leads to models that are able to fit the response variable with fewer components than the Principal Components Regression (PCR) and moreover takes into account the response variable, contrary to the PCR model. We use the NIPALS[1] algorithm of PLS, which is adapted to our problem as follows:

$$\begin{aligned} S_1 &= T_1 \cdot Q_1^T + E_1 \\ S_2 &= T_2 \cdot Q_2^T + E_2 \end{aligned} \tag{24}$$

[1] http://www.eigenvector.com/Docs/Wise_pls_properties.pdf

for any two similarity matrices S_1, S_2. Initially we choose u_1 as one column of S_2 and iteratively construct (normalized) projection vectors:

$$w_1 = \frac{S_1^T u_1}{\|S_1^T u_1\|} \rightarrow t_1 = S_1 w_1 \rightarrow q_1 = \frac{u_1^T t_1}{\|u_1^T t_1\|} \rightarrow u_1 = S_2 q_1 \rightarrow \pi_1 = \frac{S_1^T t_1}{\|t_1^T t_1\|} \tag{25}$$

The regression coefficient for the first stage is defined as: $b_1 = u_1^T t_1 (t_1^T t_1)^{-1}$. After calculating scores and loadings for the first latent variable, the S_1 and S_2 block residuals are calculated:

$$\begin{aligned} E_1 &= S_1 - t_1 \pi_1^T \\ E_2 &= S_2 - u_1 q_1^T \end{aligned} \tag{26}$$

The entire procedure is repeated by replacing S_1 and S_2 with their residuals, to get the fused similarity S_{fused} defined as:

$$S_{fused} = \sum_i t_i b_i, \qquad b_i = u_i^T t_i (t_i^T t_i)^{-1} \tag{27}$$

4.3 Memory Complexity of the non-linear graph-based fusion

The memory complexity is $\mathcal{O}(l^2)$ for the computation of each similarity matrix $S_m, m = 1, 2, \ldots, M$, $\mathcal{O}(l)$ for each similarity vector $s_m(q, .)$ and $\mathcal{O}(kl)$ for each $x_{(i)}^m, m = 1, 2, \ldots, M$, thus the overall memory complexity is quadratic in l: $\mathcal{O}(Ml^2 + Mkl + Ml)$.

In order to compare directly the baseline method with our retrieval framework with M modalities, under the same memory complexity, we seek for the number of filtered documents l', such that:

$$Ml'^2 + Mkl' + Ml' = 2l^2 + 2kl + 2l \tag{28}$$

The non-negative solution of Equation (28) with respect to l' is:

$$l' = \sqrt{\frac{(k+1)^2}{4} + \frac{2l^2 + 2kl + 2l}{M}} - \frac{k+1}{2} \tag{29}$$

For example, in the case of $M = 3, k = 10, l = 1000$, we find $l' \cong 815$. In Table 3 we report the l' values (for $k = 10, l = 1000$), so as to avoid significant memory increase in the implementation of our framework.

In Figure 4 we observe that even for 15 modalities, the number of the top-l filtered documents remains higher than 300 (red line). The increase in the number of modalities does not imply linear decrease in the number filtered documents, hence a critical number of documents are involved in the multimodal fusion, even in the case of several modalities.

Table 3: The proposed values for the top-l filtering step, in order to ensure a reasonable memory cost.

Modalities	l	**Modalities**	l
2	1000	9	468
3	815	10	444
4	704	11	423
5	630	12	405
6	575	13	389
7	531	14	374
8	497	15	361

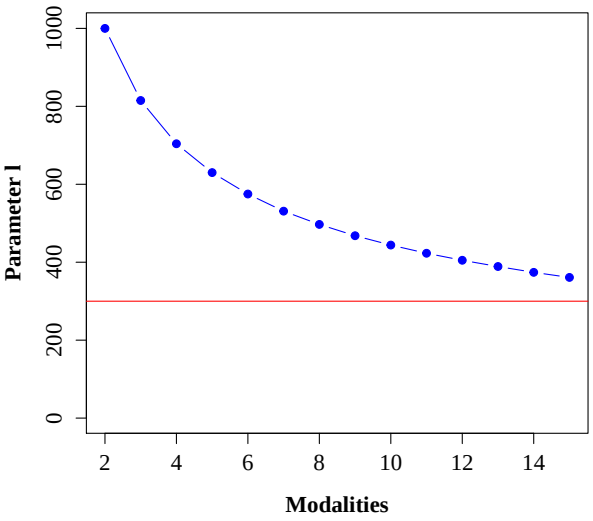


Fig. 4: Selection of the parameter l for 2 to 15 modalities.

5 Experimental Setup

5.1 Dataset Description

The proposed multimedia retrieval framework is evaluated in three datasets, namely the the WIKI11 [26], the IAPR-TC12 [10] and the MediaEval dataset from the diversity task of 2015 [13]. The WIKI11 dataset has 237,434 images with descriptions in one to three languages and 50 topics with one to five query images with caption. The 20,000 images of IAPR-TC12 include pictures of sports, actions, people, animals, cities, landscapes and many other topics. The IAPR-TC12 dataset has 60 queries with 3 examples per query. The MediaEval2015 of the diversity task has been based on the corresponding task of 2014 [13] and has 36,452 images and 123 topics.

The WIKI11 and IAPR-TC12 datasets have been annotated by means of the ImageCLEF campaign [10,26]. A title and a short description correspond to each image of both datasets, thus formulating the textual modality. The reliability and reusability of the Wikipedia collection has also been tested [27].

In each topic the datasets have up to five exemplary images and in our experiments we get the image which has the most detected visual concepts. Apart from the visual concepts, we have also extracted visual descriptors and textual concepts for each topic and for all multimedia items in the collections. In the following, we report the low- and high-level information we have extracted from the aforementioned datasets.

5.2 Feature Extraction

The features, which are employed in the evaluation of the proposed hybrid multimedia retrieval framework, are listed as follows:

Visual descriptors: The scale-invariant local descriptors RGB-SIFT [28] are extracted and are then locally aggregated into one vector representation (4000-dimensional) using VLAD encoding [14]. In addition Deep Convolution Neural Networks (DCNN) are used for the extraction of DCNN descriptors [19].

Visual concepts: The images of the multimedia objects are indexed by 346 high-level concepts (TRECVID), which are detected by multiple independent concept detectors. The locally aggregated features (VLAD encoding for RGB-SIFT descriptors) serve as input to Logistic Regression classifiers and their output is averaged and further refined. Similarly, the extracted DCNN visual descriptors [19] provide DCNN-based visual concepts, in addition to the ones provided by RGB-SIFT.

Textual concepts: The textual concepts used in evaluation of the multimedia retrieval task are extracted using the DBpedia Spotlight annotation tool, which is an open source project for automatic annotation of DBpedia entities in natural language text [6].

5.3 Settings

The textual features (tf-idf scores) are compared using the cosine similarity and the similarities of the visual features are calculated as [11]:

$$S_{\kappa\lambda} = 1 - \frac{d_{\kappa\lambda}}{\max_{\lambda} d_{\kappa\lambda}} \tag{30}$$

where $d_{\kappa\lambda}$ is the Euclidean distance between item κ and item λ. The initial filtering stage involves Lucene[2] text search, where we keep the top-1K similar-to-the-query documents, i.e. $l = 1000$. One iteration ($i = 1$) is used as in [2],

[2] https://lucene.apache.org/

since more than one ($i > 1$) iterations adds the noise of one modality to the others. The parameter k in the operator $\mathbf{K}$ is set to $k = 10$ and we used the Python implementation[3] of PLS Regression.

In all three datasets we mark all retrieved results which are not annotated as not relevant, a fact with a strong impact to the overall performance in MAP and P@20 scores.

In the MediaEval dataset, the topics "Machu Piccu" appears in the collection as "machupiccu", therefore we modified the text query in the initial filtering stage. Similar modifications have also been done in the queries: "sights, bad_weather, bad_weather tourist_destinations, South_Korea, oxidised" for the IAPR dataset, "siegessaeule" for the MediaEval dataset and "boxing_match" for the Wikipedia dataset, respectively. Four queries of the Wikipedia dataset have no text and they are skipped.

6 Results

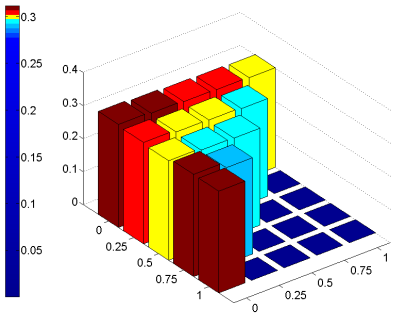
The overall multimedia retrieval framework is evaluated using the Mean Average Precision (MAP) and the Precision at the top-20 retrieved results (P@20) as two of the most well-established measures in Information Retrieval tasks.

We compare our model with other unsupervised multimedia retrieval methods. Firstly, the majority vote over all modalities (rule-based fusion) determines the modality with the highest performance [23]. Secondly, we use the cross-media fusion [5] of three modalities and thirdly the random-walk approach of [12]. Fourth baseline method is the non-linear fusion [22] of all modalities and finally we compare our framework with the extension of the unifying fusion framework of [2] in the case of three modalities [9] in two cases: first with the SIFT visual descriptors and second with the state-of-the-art DCNN visual features. Our proposed framework combines SIFT with DCNN using PLS Regression, using non-linear graph-based fusion of all three modalities.

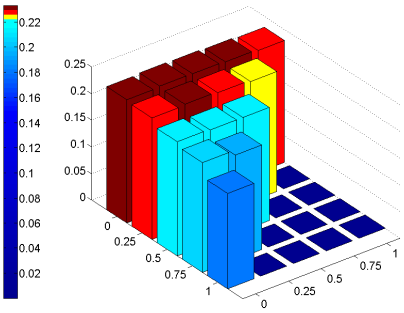
Initially, the parameters $\alpha_m, \alpha'_m, \beta_m, m = 1, 2, 3$ are kept constant while the parameters $\gamma_1, \gamma_2, \gamma_3$ ($\gamma_3 = 1 - \gamma_1 - \gamma_2$) change. We find that the optimal choice of parameters $\gamma_1, \gamma_2, \gamma_3$ is (1,0,0) for the WIKI dataset, (0.25,0,0.75) for the IAPR-TC12 dataset and (0.5,0,0.5) for the MediaEval dataset. Modifying the parameters $\beta_m, m = 1, 2, 3$, while the other parameters are kept constant, did not led to significant changes in the performance. Therefore, we proceed by changing the parameters α_m, α'_m, taking also into account that many well-known fusion techniques appear as special cases. For example, if $\alpha_1 = 1, \alpha_2 = 0, \alpha_3 = 0$, $\alpha'_m = 0$, only one modality is involved (text). Similarly, in case $\alpha_1 = \alpha_2 = \alpha_3 = 0$ and all other $\alpha'_m, m = 1, 2, 3$ are tuned, then we get a random-walk based fusion.

The final results are presented in Table 4. The best performance appears for two components in the PLS Regression for the WIKI11 and the MediaEval dataset, while a peak appears in the performance for three components in the

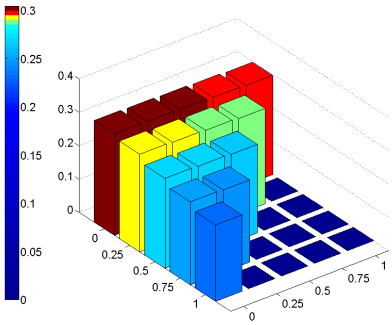
[3] http://scikit-learn.org/stable/modules/generated/sklearn.cross_decomposition.PLSRegression.html



(a) WIKI



(b) IAPR



(c) MediaEval

Fig. 5: Variation in MAP scores by changing the parameters γ_m

IAPR-TC12 dataset. The best parameter selection are $\alpha_1 = 0.5, \alpha_2 = 0.0, \alpha'_1 = 0, \alpha'_2 = 0.25$ for the WIKI11 dataset, $\alpha_1 = 0.5, \alpha_2 = 0.25, \alpha'_1 = 0.5, \alpha'_2 = 0.0$ for the IAPR-TC12 dataset and $\alpha_1 = \alpha_2 = \alpha'_1 = \alpha'_2 = 0.0$ for the MediaEval

Table 4: Evaluation results.

Method	WIKI11		IAPR-TC12		MediaEval	
	MAP	**P@20**	**MAP**	**P@20**	**MAP**	**P@20**
Majority vote (best modality)	0.3325	0.1630	0.2385	0.2050	0.3154	0.3760
Cross-media fusion	0.3325	0.1630	0.2392	0.2084	0.3144	0.3764
Random walk based fusion	0.3344	0.1640	0.2401	0.2083	0.3218	0.4020
Non-linear fusion	0.3341	0.1730	0.2418	0.2200	0.3220	0.4020
Graph-based fusion (SIFT)	0.3341	0.1730	0.2418	0.2200	0.3214	0.4028
Graph-based fusion (DCNN)	0.3958	0.2003	0.2716	0.2400	0.3646	0.4439
Graph-based fusion with PLS	**0.4013**	**0.2040**	**0.2771**	**0.2417**	**0.3667**	**0.4581**

dataset, respectively. Our method outperforms all considered baseline methods in both MAP and P@20 scores.

7 Conclusion

We presented a multimedia retrieval framework, which combines graph-based and non-linear fusion along with Partial Least Squares Regression for the fusion of several modalities of multiple views. The hybrid framework has been presented, in general, for M modalities and has been evaluated in two public datasets with text-image multimodal objects. The experimental results demonstrate that our proposed approach outperforms other baseline multimedia retrieval approaches in terms of two measures in three multimedia collections. In the future, we plan to employ multimodal non-linear and graph-based fusion techniques in multimodal classification and clustering when multiple modalities of multiple views appear, taking into account computational and memory complexity issues.

Acknowledgements This work was partially supported by the European Commission by the projects MULTISENSOR (FP7-610411) and KRISTINA (H2020-645012).

References

1. Ah-Pine, J., Bressan, M., Clinchant, S., Csurka, G., Hoppenot, Y., Renders, J.M.: Crossing textual and visual content in different application scenarios. Multimedia Tools and Applications **42**(1), 31–56 (2009)
2. Ah-Pine, J., Csurka, G., Clinchant, S.: Unsupervised visual and textual information fusion in cbmir using graph-based methods. ACM Transactions on Information Systems (TOIS) **33**(2), 9 (2015)
3. Atrey, P.K., Hossain, M.A., El Saddik, A., Kankanhalli, M.S.: Multimodal fusion for multimedia analysis: a survey. Multimedia systems **16**(6), 345–379 (2010)
4. Caicedo, J.C., BenAbdallah, J., González, F.A., Nasraoui, O.: Multimodal representation, indexing, automated annotation and retrieval of image collections via non-negative matrix factorization. Neurocomputing **76**(1), 50–60 (2012)
5. Clinchant, S., Csurka, G., Perronnin, F., Renders, J.M.: Xrces participation to imageval. In: ImageEval workshop at CVIR (2007)
6. Daiber, J., Jakob, M., Hokamp, C., Mendes, P.N.: Improving efficiency and accuracy in multilingual entity extraction. In: Proceedings of the 9th International Conference on Semantic Systems, pp. 121–124. ACM (2013)

7. Feng, F., Wang, X., Li, R.: Cross-modal retrieval with correspondence autoencoder. In: Proceedings of the 22nd ACM international conference on Multimedia, pp. 7–16. ACM (2014)
8. Gialampoukidis, I., Moumtzidou, A., Liparas, D., Vrochidis, S., Kompatsiaris, I.: A hybrid graph-based and non-linear late fusion approach for multimedia retrieval. In: Content-Based Multimedia Indexing (CBMI), 2016 14th International Workshop on, pp. 1–6. IEEE (2016)
9. Gialampoukidis, I., Moumtzidou, A., Tsikrika, T., Vrochidis, S., Kompatsiaris, I.: Retrieval of multimedia objects by fusing multiple modalities. In: Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, pp. 359–362. ACM (2016)
10. Grubinger, M., Clough, P., Müller, H., Deselaers, T.: The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In: International Workshop OntoImage, pp. 13–23 (2006)
11. Hafner, J., Sawhney, H.S., Equitz, W., Flickner, M., Niblack, W.: Efficient color histogram indexing for quadratic form distance functions. Pattern Analysis and Machine Intelligence, IEEE Transactions on **17**(7), 729–736 (1995)
12. Hsu, W.H., Kennedy, L.S., Chang, S.F.: Video search reranking through random walk over document-level context graph. In: Proceedings of the 15th international conference on Multimedia, pp. 971–980. ACM (2007)
13. Ionescu, B., Popescu, A., Lupu, M., Gînsca, A.L., Müller, H.: Retrieving diverse social images at mediaeval 2014: Challenge, dataset and evaluation. In: MediaEval (2014)
14. Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pp. 3304–3311. IEEE (2010)
15. Kitanovski, I., Trojacanec, K., Dimitrovski, I., Loskovska, S.: Multimodal medical image retrieval. In: ICT Innovations 2012, pp. 81–89. Springer (2013)
16. Lan, Z.z., Bao, L., Yu, S.I., Liu, W., Hauptmann, A.G.: Multimedia classification and event detection using double fusion. Multimedia tools and applications **71**(1), 333–347 (2014)
17. Langville, A.N., Meyer, C.D.: A survey of eigenvector methods for web information retrieval. SIAM review **47**(1), 135–161 (2005)
18. Magalhães, J., Rüger, S.: An information-theoretic framework for semantic-multimedia retrieval. ACM Transactions on Information Systems (TOIS) **28**(4), 19 (2010)
19. Markatopoulou, F., Mezaris, V., Patras, I.: Cascade of classifiers based on binary, non-binary and deep convolutional network descriptors for video concept detection. In: Image Processing (ICIP), 2015 IEEE International Conference on, pp. 1786–1790. IEEE (2015)
20. Moumtzidou, A., Gialampoukidis, I., Mironidis, T., Liparas, D., Vrochidis, S., Kompatsiaris, I.: A multimedia interactive search engine based on graph-based and non-linear multimodal fusion. In: Content-Based Multimedia Indexing (CBMI), 2016 14th International Workshop on, pp. 1–4. IEEE (2016)
21. Rasiwasia, N., Costa Pereira, J., Coviello, E., Doyle, G., Lanckriet, G.R., Levy, R., Vasconcelos, N.: A new approach to cross-modal multimedia retrieval. In: Proceedings of the international conference on Multimedia, pp. 251–260. ACM (2010)
22. Safadi, B., Sahuguet, M., Huet, B.: When textual and visual information join forces for multimedia retrieval. In: Proceedings of International Conference on Multimedia Retrieval, p. 265. ACM (2014)
23. Sanderson, C., Paliwal, K.K.: Identity verification using speech and face information. Digital Signal Processing **14**(5), 449–480 (2004)
24. Siddiquie, B., White, B., Sharma, A., Davis, L.S.: Multi-modal image retrieval for complex queries using small codes. In: Proceedings of International Conference on Multimedia Retrieval, p. 321. ACM (2014)
25. Tsikrika, T., Andreadou, K., Moumtzidou, A., Schinas, E., Papadopoulos, S., Vrochidis, S., Kompatsiaris, I.: A unified model for socially interconnected multimedia-enriched objects. In: MultiMedia Modeling, pp. 372–384. Springer (2015)
26. Tsikrika, T., Kludas, J.: The wikipedia image retrieval task. In: ImageCLEF, pp. 163–183. Springer (2010)
27. Tsikrika, T., Kludas, J., Popescu, A.: Building reliable and reusable test collections for image retrieval: The wikipedia task at imageclef. IEEE MultiMedia **19**(3), 0024 (2012)

28. Van De Sande, K.E., Gevers, T., Snoek, C.G.: Evaluating color descriptors for object and scene recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on **32**(9), 1582–1596 (2010)
29. Wang, J., He, Y., Kang, C., Xiang, S., Pan, C.: Image-text cross-modal retrieval via modality-specific feature learning. In: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, pp. 347–354. ACM (2015)
30. Wang, W., Ooi, B.C., Yang, X., Zhang, D., Zhuang, Y.: Effective multi-modal retrieval based on stacked auto-encoders. Proceedings of the VLDB Endowment **7**(8), 649–660 (2014)
31. Wang, Y., Lin, X., Zhang, Q.: Towards metric fusion on multi-view data: a cross-view based graph random walk approach. In: Proceedings of the 22nd ACM international conference on Conference on information & knowledge management, pp. 805–810. ACM (2013)
32. Wang, Y., Wu, F., Song, J., Li, X., Zhuang, Y.: Multi-modal mutual topic reinforce modeling for cross-media retrieval. In: Proceedings of the ACM International Conference on Multimedia, pp. 307–316. ACM (2014)
33. Xu, S., Li, H., Chang, X., Yu, S.I., Du, X., Li, X., Jiang, L., Mao, Z., Lan, Z., Burger, S., et al.: Incremental multimodal query construction for video search. In: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, pp. 675–678. ACM (2015)
34. Younessian, E., Mitamura, T., Hauptmann, A.: Multimodal knowledge-based analysis in multimedia event detection. In: Proceedings of the 2nd ACM International Conference on Multimedia Retrieval, p. 51. ACM (2012)