A hybrid graph-based and non-linear late fusion approach for multimedia retrieval

Ilias Gialampoukidis, Anastasia Moumtzidou, Dimitris Liparas, Stefanos Vrochidis, Ioannis Kompatsiaris
Information Technologies Institute
Centre for Research and Technology Hellas
{heliasgj, moumtzid, dliparas, stefanos, ikom}@iti.gr

Abstract—**Nowadays, multimedia retrieval has become a task of high importance, due to the need for efficient and fast access to very large and heterogeneous multimedia collections. An interesting challenge within the aforementioned task is the efficient combination of different modalities in a multimedia object and especially the fusion between textual and visual information. The fusion of multiple modalities for retrieval in an unsupervised way has been mostly based on early, weighted linear, graph-based and diffusion-based techniques. In contrast, we present a strategy for fusing textual and visual modalities, through the combination of a non-linear fusion model and a graph-based late fusion approach. The fusion strategy is based on the construction of a uniform multimodal contextual similarity matrix and the non-linear combination of relevance scores from query-based similarity vectors. The proposed late fusion approach is evaluated in the multimedia retrieval task, by applying it to two multimedia collections, namely the WIKI11 and IAPR-TC12. The experimental results indicate its superiority over the baseline method in terms of Mean Average Precision for both considered datasets.**

Keywords—Multimedia retrieval; non-linear fusion; unsupervised multimodal fusion

I. Introduction

The need for efficient and rapid access to large and diverse multimedia collections has led to multimedia retrieval systems gaining much attention from the relevant community. The main challenge in searching for multimedia content is how to effectively combine heterogeneous sources of information, in most cases textual and visual. This problem is known as multimodal fusion [1], with applications to several related problems, such as multimodal search, retrieval, summarization, recommendation etc.

Multimodal fusion utilizes information from several modalities. The modalities can be, for example, low-level visual descriptors (based on color, shape, texture, location, etc.), low-level textual features (raw text from webpages, video subtitles, features extracted from audio files using automatic speech recognition or from video files using optical character recognition, etc.), metadata (time stamp, tags, source, position in a social graph) and high-level textual features [1].

All sources of information need to be indexed smartly, in order to ensure fast access to the modalities of a multimedia object. Recently an efficient representation and indexing technique for multimodal objects [2], named Socially Interconnected MultiMedia-enriched Objects (SIMMO) has been proposed. For the multimedia retrieval task, the key problem is to combine all available information in order to retrieve multimodal documents, relevant to a given multimodal query. Towards this direction, we provide a novel hybrid framework for the multimodal fusion of three modalities. Our method is compared to all prominent baseline methods of unsupervised fusion (early, late linearly weighted, random-walk and graph-based fusion). Our motivation arises from the unifying framework presented in [3], which exploits two modalities. Our contribution can be summarized as follows:

- We propose a novel multimodal fusion approach, which combines graph-based and non-linear fusion.
- We present a multimedia retrieval method based on our hybrid graph-based and non-linear late fusion technique.
- We evaluate our hybrid approach using two public datasets with text-image multimodal objects.

Section II presents the state-of-the-art techniques in multimodal fusion and multimedia retrieval. Section III describes the necessary background, the notation that we adopt, as well as our novel hybrid framework. In Section IV, our approach is compared to other baseline multimodal fusion methods in the multimedia retrieval task. Finally, some concluding remarks are provided in Section V.

II. Related Work

One important challenge in multimedia retrieval is the fusion of heterogeneous modalities (e.g. visual, textual). In [4], a video retrieval framework is proposed, in which a text-based similarity score is provided by means of Lucene[1] text indexing on the video subtitles, visual concepts provide visual-based similarity scores and the aforementioned similarity scores are fused through a simple non-linear approach. Video search systems are often interactive. Specifically, the video similarity is improved by user-generated relevance feedback

[1] https://lucene.apache.org/core/

978-1-4673-8695-1/16/$31.00 ©2016 IEEE

and the query is progressively refined by the user [5]. Otherwise, the returned results of the multimedia retrieval module are re-ranked in a post-process of the core search [6]. In contrast to the aforementioned works, our framework performs multimodal retrieval by combining high-level textual and visual features with a novel fusion method.

One important issue in multimodal fusion is the level, at which fusion is realized. The three basic approaches involve information fusion at the feature level (also known as early fusion), at the decision level (also known as late fusion) or at both levels (hybrid fusion) [1]. Most approaches involve the combination of textual and visual features. Metric fusion [7] is a random walk approach, whose aim is to fuse different “views” of the same modality, such as SIFT, GIST and LBP visual features and has been evaluated in the image retrieval task. A co-training approach for fusing multiple views of a modality has been applied in the clustering task [8]. However, in this paper, we focus on the multimedia retrieval task.

Other multimedia and cross-modal retrieval tasks [9] employ Latent Dirichlet Allocation (LDA) probabilistic approaches, such as in [10], which either generate a joint topic probability distribution for all modalities [11] or combine the topic distribution per modality [12]. In some studies, Convolutional Neural Networks (CNN) are used to learn high-level features from two modalities (text-image pairs) in cross-modal retrieval (e.g. [13]). Another approach that aims at efficient cross-modal retrieval is by using correlation matching [14] between the two modalities. In addition, a Partial Least Squares (PLS) based approach is followed in [15], in which different modalities of the data are mapped into a common latent space. This methodology is evaluated in the image retrieval task. Contrary to the aforementioned approaches for cross-modal retrieval, our proposed framework does not involve a training stage, but proposes an unsupervised fusion of all features. In addition, the query in multimedia retrieval is a multimodal object, thus all modalities need to be exploited.

Graph-based methods and random-walk approaches have been used in a unifying framework [3] for fusing visual and textual information in Content-Based Multimedia Information Retrieval. The random-walk approach for multimodal fusion was first presented in [16], where the fusion of textual and visual information leads to improved video search results. The unifying framework described in [3] does not require user’s relevance feedback and is unsupervised. Moreover, it includes as special cases all well-known early, late, linearly weighted, diffusion and graph-based models and finally, it is evaluated in the multimedia retrieval task. In this paper, we extend the aforementioned unsupervised unifying framework to multiple modalities by means of a hybrid framework, which combines graph-based and non-linear fusion.

III. Multimedia Retrieval with Non-Linear Fusion

The necessary background in unsupervised fusion of two modalities and the notation that we adopt are presented in subsection A below. In subsection B, we present a novel hybrid framework, which fuses multiple modalities.

A. *Unsupervised graph-based fusion of two modalities*

In order to compare the similarity between any two multimodal objects, it is necessary to fuse the similarities of all involved modalities. Given two modalities (one textual and one visual) of n objects and a query q, we denote by s_t the textual-based similarity score vector, in response to a query q and by s_v the corresponding visual-based similarity score vector.

The classic late fusion method provides a fused similarity score vector $s(q)$, by averaging the two similarity vectors:

$$s(q) = a \cdot s_t + (1-a) s_v \tag{1}$$

Another way to fuse two similarity vectors is the following non-linear way [4]:

$$s(q) = (s_t)^a + (s_v)^{1-a} \tag{2}$$

Eq. 2 is an alternative approach for the simple linear (weighted mean) fusion of Eq. 1. In a more general and unifying fusion approach [3], random-walk based scores and cross-media similarities are employed:

$$s(q) = a_t s_t(q,.) + a_v s_v(q,.) + a_{tv} x_{(i)} + a_{vt} y_{(i)} \tag{3}$$

In Eq. 3, the fusion of similarity vectors $s_t, s_v, x_{(i)}, y_{(i)}$ is linear, under the restriction $a_t + a_v + a_{tv} + a_{vt} = 1$, and the vectors $x_{(i)}, y_{(i)}$ are defined as follows:

$$x_{(i)} \propto \mathbf{K}\big(x_{(i-1)}, k\big) \cdot [(1-\gamma) D \cdot (\beta S_t + (1-\beta) S_v) + \gamma e \cdot s_t] \tag{4}$$

$$y_{(i)} \propto \mathbf{K}\big(y_{(i-1)}, k\big) \cdot [(1-\gamma) D \cdot (\beta S_v + (1-\beta) S_t) + \gamma e \cdot s_v] \tag{5}$$

where D is a row-normalizing matrix so that, for any matrix A, the matrix $D \cdot A$ is row-stochastic. The $l \times 1$ vector of ones is denoted by e. The operator $\mathbf{K}(x, k)$ takes as input a vector x and assigns a zero value to elements whose score is strictly lower than the k-th highest score in x. The number of iterations is $i = 1$ and as initial condition, the model of Eq. 4 and Eq. 5 sets $x_{(0)} = s_t$ and $y_{(0)} = s_v$. The default value for k is 10. The number $l < n$ is fixed, usually set to $l = 1000$ and is defined as the number of initially semantically filtered objects, with respect to the textual modality, assuming mainly that “the text query is the main semantic source with regard to the user information”. We also denote by S_t and S_v the $l \times l$ similarity matrices for all pairs of objects, with respect to the textual and visual modality, respectively. The dot product is denoted by " $\cdot$ " and the (i, j) element of a matrix A is denoted by $A[i, j]$. The similarity matrices S_t and S_v are normalized, as in Eq. 6 below, in order to ensure that all elements are numbers in the interval [0, 1].

$$A[i,j] \leftarrow \frac{A[i,j] - \min_j A[i,j]}{\max_j A[i,j] - \min_j A[i,j]} \tag{6}$$

The model of Eq. 4 and Eq. 5 is a unifying framework because it includes, as special cases, many well-known late fusion methods. Indicatively, for $a_{tv} = a_{vt} = 0$, the model reduces to the classic linear weighted average of textual and visual similarities, for $a_t = a_v = a_{vt} = 0$, $\gamma = 0$ and sufficiently large number of iterations i, the model is the random-walk based approach [16] and for $a_v = a_{tv} = 0, \beta = 0, \gamma = 0$ the model is the cross-media approach, known for performing well in the ImageCLEF tasks [17, 18].

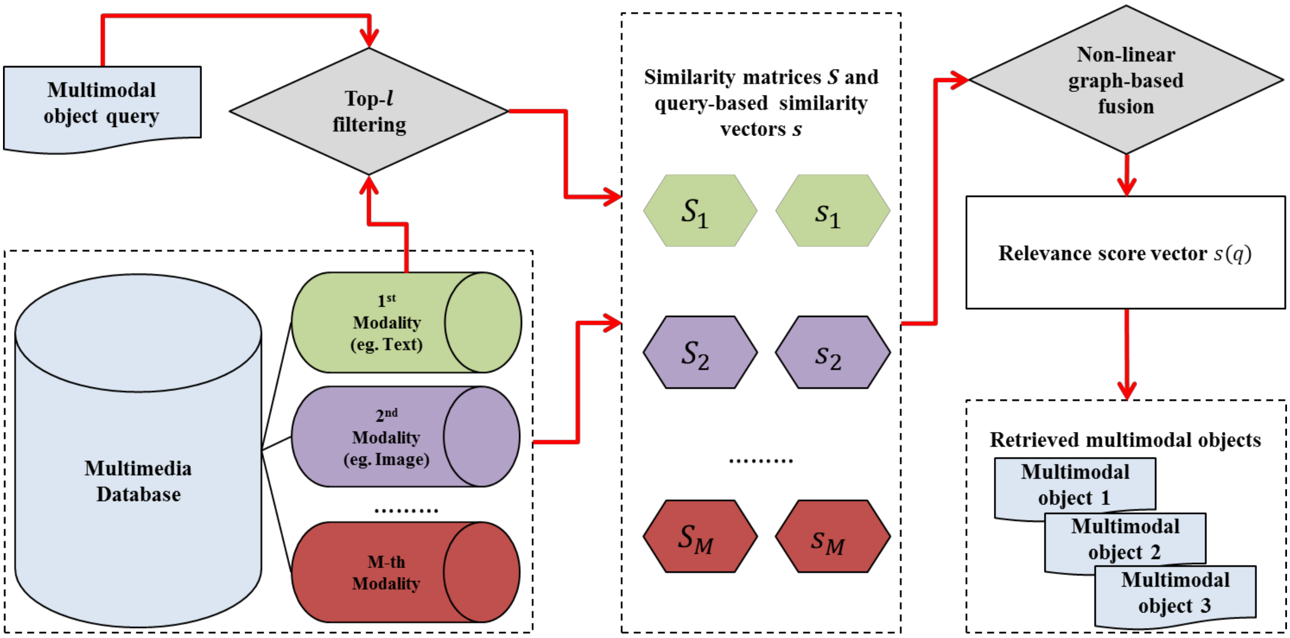

Fig. 1. A hybrid graph-based and non-linear late fusion model for multimedia retrieval. The first modality is involved in the semantic filtering stage, and therefore, the dimensions of all similarity matrices and query-based similarity vectors are reduced. On the reduced similarity matrices and vectors, the fusion of all modalities obtaines one relevance score vector, allowing for the retrieval of the most relevant-to-the-query multimodal objects.

B. *Unsupervised graph-based fusion of M modalities*

The unifying model of Eq. 4 and Eq. 5 cannot be easily extended to several modalities, due to the prohibitively increasing number of involved parameters. In the general case of M modalities, the parameters α are M^2, the free parameters β are $M-1$ and the free parameters γ are also $M-1$. For a direct generalization of the considered model of Eq. 4 and Eq. 5, the number of parameters, for M modalities, is M^2+2M-2. Even for $M=3$ modalities, the number of parameters is $3^2+2+2=13$ and for $M=4$ modalities is $4^2+3+3=22$.

Motivated by the model of Eq. 4 and Eq. 5, 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 number of parameters of the unifying framework [3]. We introduce the notation S_m for the similarity matrix with respect to the m-modality and s_m for the corresponding query-based similarity vectors, in order to define the following multimodal contextual similarity matrix:

$$C = \sum_{m=1}^{M} \beta_m S_m\,, \qquad \sum_{m=1}^{M} \beta_m = 1 \tag{7}$$

The matrix C of Eq. 7 is row-normalized so as to obtain the row-stochastic matrix P:

$$P[i,j] = \frac{C[i,j]}{\sum_{j=1}^{l} C[i,j]} \tag{8}$$

For all $m = 1,2,\ldots,M$ modalities we set $x^m_{(0)} = s_m(q,.)$, and we define the update rule:

$$x^m_{(i)} \propto \boldsymbol{K}\left(x^m_{(i-1)}, k\right) \cdot \left[\left(1 - \sum_{w \neq m} \gamma_w\right) P + \sum_{w \neq m} \gamma_w e \cdot s_w(q,.)\right] \tag{9}$$

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

$$s(q) = \sum_{m=1}^{M} a_m s_m + \sum_{m=1}^{M} a'_m x^m_{(i)} \tag{10}$$

under the restriction

$$\sum_{m=1}^{M} a_m + \sum_{m=1}^{M} a'_m = 1 \tag{11}$$

Alternatively, under the same restriction of Eq. 11, we propose the combination of the non-linear fusion, already presented in Eq. 2, and the fusion of Eq. 10 as follows:

$$s(q) = \sum_{m=1}^{M} (s_m)^{a_m} + \sum_{m=1}^{M} a'_m x^m_{(i)} \tag{12}$$

For M modalities, the models of Eq. 10 and Eq. 12 have $M-1$ free parameters a, $M-1$ free parameters β and $M-1$ free parameters γ, thus $3M-3$ parameters in total. Our hybrid graph-based and non-linear late fusion approach for multimedia retrieval is illustrated in Fig. 1 for the general case of M modalities. Our hybrid graph-based and non-linear late fusion approach for multimedia retrieval is illustrated in Fig. 1 for the general case of M modalities. The similarities S_m are linearly combined in Eq. 7, to formulate the uniform contextual similarity matrix C, which is row-normalized in Eq. 8, so as to provide the row-stochastic matrix P. The $l \times l$ matrix P is regarded as a transition probability matrix on a graph of l nodes, where $P[i,j]$ is the probability to have a transition from object i to object j on the graph random walk of l objects. Assuming there is a $1 \times l$ vector z, which describes the initial state (node) of the random walker, the probability vector describing the current position (node) of the random walker after i-steps is given by $z \cdot P^i$. In our case, the initial state vector z is given by the result of $\boldsymbol{K}\left(x_{(i-1)}^m, k\right)$, multiplied by the modified transition probability matrix according to Eq. 9. Finally, the graph-based scores per modality $x_{(i)}^m$ are combined with the similarity vectors s_m in a non-linear way, as described in Eq. 12.

Example with $\boldsymbol{M = 3}$: The model, which we have introduced in Eq. 10 and its variation in Eq. 12 with non-linear fusion are presented for $M = 3$ modalities, so as to be tested in the following section (Section IV). In particular, the multimodal contextual similarity matrix of Eq. 7 becomes:

$$C = \beta_1 S_1 + \beta_2 S_2 + (1 - \beta_1 - \beta_2) S_3 \qquad (13)$$

The matrix C of Eq. 13 is row-normalized, as done in Eq. 8, so as to obtain the row-stochastic matrix P, which is substituted in the following model:

$$\begin{aligned} x_{(i)}^1 &\propto \mathbf{K}\left(x_{(i-1)}^1, k\right) \cdot [(1 - \gamma_2 - \gamma_3)P + \gamma_2 e \cdot s_2(q,.) + \gamma_3 e \cdot s_3(q,.)] \\ x_{(i)}^2 &\propto \mathbf{K}\left(x_{(i-1)}^2, k\right) \cdot [(1 - \gamma_1 - \gamma_3)P + \gamma_1 e \cdot s_1(q,.) + \gamma_3 e \cdot s_3(q,.)] \\ x_{(i)}^3 &\propto \mathbf{K}\left(x_{(i-1)}^3, k\right) \cdot [(1 - \gamma_2 - \gamma_1)P + \gamma_2 e \cdot s_2(q,.) + \gamma_1 e \cdot s_1(q,.)] \end{aligned} \qquad (14)$$

The vector of relevance score $s(q)$, in response to the query q, linearly combines the similarity vectors $s_m, m = 1,2,3$ and the vectors $x_{(i)}^m, m = 1,2,3$ of Eq. 14. Alternatively, the relevance score $s(q)$ is obtained by non-linear fusion of all similarity vectors $s_m, m = 1,2,3$ and $x_{(i)}^m, m = 1,2,3$. The proposed models are summarized as follows:

- Graph-based with linear fusion:

$$s(q) = a_1 s_1 + a_2 s_2 + a_3 s_3 + a'_1 x_{(i)}^1 + a'_2 x_{(i)}^2 + a'_3 x_{(i)}^3 \qquad (15)$$

- Graph-based with non-linear fusion:

$$s(q) = (s_1)^{a_1} + (s_2)^{a_2} + (s_3)^{a_3} + a'_1 x_{(i)}^1 + a'_2 x_{(i)}^2 + a'_3 x_{(i)}^3 \qquad (16)$$

The memory complexity is $\mathcal{O}(l^2)$ for the computation of each similarity matrix $S_m, m = 1,2,\dots,M$, $\mathcal{O}(l)$ for each similarity vector $s_m(q,.)$ and $\mathcal{O}(kl)$ for each $x_{(i)}^m$, thus the overall memory complexity is quadratic in l: $\mathcal{O}(Ml^2 + Mkl + Ml)$. The filtering stage of $l < n$ multimodal objects allows for tuning the memory complexity and involves the multimodal fusion only l objects. The filtering step reduces the complexity of the models in Eq. 15 and Eq. 16, with significant reduction in the processing time of the multimedia retrieval task.

IV. Experiments

In the following, we describe the datasets that are used for the evaluation of the proposed hybrid multimedia retrieval framework, the features that are extracted from all multimodal objects and the corresponding results.

A. Datasets

The proposed multimedia retrieval framework is evaluated in two datasets, namely the IAPR-TC12[2] dataset and the WIKI11[3] dataset. The 20,000 images of IAPR-TC12 include pictures of sports, actions, people, animals, cities, landscapes and many other topics. 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 IAPR-TC12 dataset has 60 queries with 3 examples per query. The IAPR-TC12 and the WIKI11 datasets have been annotated by means of the ImageCLEF campaign [19, 20]. A title and a short description correspond to each image of both datasets, thus formulating the textual modality.

B. 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 [21] are extracted and are then locally aggregated into one vector representation (4000-dimensional) using VLAD encoding [22].

Visual concepts: The images of the multimedia objects are indexed by 346 high-level concepts (e.g. water, aircraft), 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 [23].

Textual concepts: The textual concepts used in evaluation of the multimedia retrieval task are extracted using the DBpedia Spotlight[4] annotation tool, which is an open source project for automatic annotation of DBpedia entities in natural language text [24].

The similarity between any two objects, for each modality, strongly depends on the distance between any two objects. Given the Euclidean distances for all pairs (i,j) of objects $D[i,j]$, the corresponding similarity $S[i,j]$ is [25]:

$$S[i,j] = 1 - \frac{D[i,j]}{\max D[i,j]} \qquad (17)$$

The visual similarities are provided by means of the vector representation for each visual modality and the Euclidian distances for all pairs (i,j) of objects are transformed into a similarity by (17). For the textual concepts, Lucene indexing provides the similarity score between any two text documents, where Lucene-based similarity[5] scores are obtained by "Lucene's Practical Scoring Function".

C. Results

The proposed hybrid multimedia retrieval framework is evaluated using the Mean Average Precision (MAP). We selected MAP scores as the most well-established measure in

[2] http://imageclef.org/photodata

[3] http://www.imageclef.org/wikidata

[4] http://dbpedia-spotlight.github.io/demo/

[5] https://lucene.apache.org/core/3_0_3/api/core/org/apache/lucene/search/Similarity.html

Information Retrieval tasks. In brief, for a given query, the average precision is initially computed, defined as the area under the precision-recall curve and, then, averaged for all queries to obtain the MAP score for each dataset.

The proposed hybrid multimedia retrieval framework is compared to other well-known baseline methods. First, the majority vote over all modalities is a rule-based fusion [26], in which the highest MAP score over all modalities determines which modality to fuse. Second, the classical linear weighted fusion method [1] which fuses only the similarity vectors $s_m, m = 1,2,3$. Third, the non-linear fusion method of the similarity vectors $s_m, m = 1,2,3$ [4]:

$$s(q) = (s_1)^{a_1} + (s_2)^{a_2} + (s_3)^{a_3} \tag{18}$$

Fourth, the random-walk based approach [16] for multimodal fusion, fifth, the cross-media approach [17, 18] and, finally, the unifying fusion framework [3], in which we keep the best two modalities.

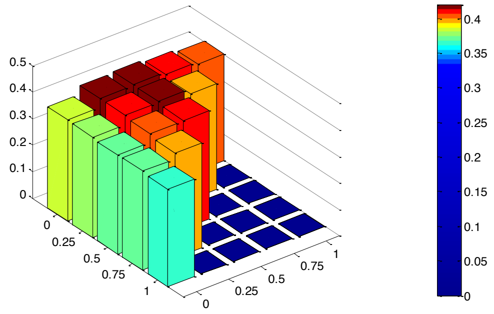

Fig. 2. Mean Average Precision for the WIKI11 dataset for the pairs of the parameters γ_1, γ_2

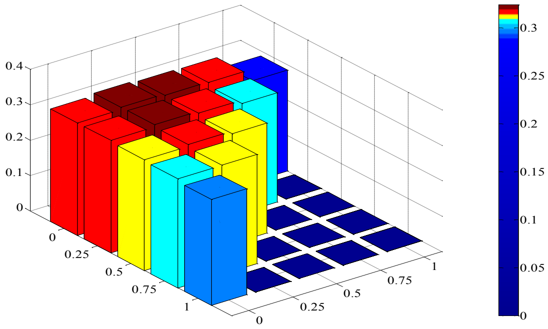

Fig. 3. Mean Average Precision for the IAPR-TC12 dataset for the pairs of the parameters γ_1, γ_2

Fig. 2 and Fig. 3 show the results of the proposed hybrid framework, when the parameters γ_1, γ_2 $(\gamma_3 = 1 - \gamma_1 - \gamma_2)$ are tuned, while the other parameters are kept fixed and the best MAP scores are marked with a " $*$ ". In particular, our hybrid fusion method, presented in Eq. 16, is tuned using the defaults $k = 10, i = 1$, while $a_1 = a_2 = a_3 = a'_1 = a'_2 = a'_3 = 1/6$ and $\beta_1 = \beta_2 = \beta_3 = 1/3$. We observe that the best MAP scores appear when $\gamma_1 = 0.5, \gamma_2 = 0, \gamma_3 = 0.5$ for the WIKI11 dataset and when $\gamma_1 = 0.25, \gamma_2 = 0, \gamma_3 = 0.75$ for the IAPR-TC12 dataset. For the best values of $\gamma_i, i = 1,2,3$ we then tuned the parameters $\beta_i, i = 1,2,3$ under the assumption $\beta_1 + \beta_2 + \beta_3 = 1$ and we did not observe any triplet $(\beta_1, \beta_2, \beta_3)$ that significantly improves Mean Average Precision. For the best values of $\gamma_i, i = 1,2,3$ we then tuned the parameters $a_i, a'_i, i = 1,2,3$ and the best MAP values are reported in Table I.

In Table I, the proposed graph-based fusion method, as in Eq. 15, outperforms the main state of the art approaches as explicitly obtained by the unifying unsupervised fusion framework in Section III.A. The hybrid graph-based fusion method with non-linear fusion (Eq. 16) further increases MAP.

TABLE I. RESULTS

Ref.	Fusion Method	Mean Average Precision	
		WIKI11	***IAPR-TC12***
[26]	Majority vote (best modality)	0.3479	0.2342
[3]	Unifying fusion framework (best two modalities)	0.3637	0.2769
[1]	Linear weighted fusion	0.4227	0.3214
[17]	Cross-media fusion	0.4254	0.3068
[16]	Random walk based fusion	0.4239	0.3224
[4]	Non-linear fusion	0.4231	0.3211
	Graph-based fusion	0.4290	0.3261
	Graph-based + non-linear fusion	**0.4302**	**0.3308**

For the IAPR-TC12 dataset, the proposed hybrid graph-based framework with non-linear fusion outperforms the simple non-linear fusion by 3.02%, the random walk based fusion by 2.61%, the cross-media fusion by 7.82%, the linear weighted fusion by 2.92%, the unifying approach for the best two modalities by 19.46% and the majority vote scheme by 41.24%.

For the WIKI11 dataset, the proposed hybrid graph-based framework with non-linear fusion outperforms the simple non-linear fusion by 1.68%, the random walk based fusion by 1.49% , the cross-media fusion by 1.12%, the linear weighted fusion by 1.77%, the unifying approach for the best two modalities by 18.28% and the majority vote scheme by 23.65%.

Given a multimodal query with textual concepts, visual descriptors and visual concepts, in Fig. 1, we demonstrate the top-retrieved results for the proposed hybrid graph-based framework with non-linear fusion and the top-retrieved results for two baseline methods.

V. Conclusion

In this work we presented a novel hybrid multimodal fusion approach, which combines graph-based and non-linear fusion. The hybrid fusion approach is utilized, in order to retrieve multimodal objects relevant to a multimodal query. 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 vividly the fact that our proposed approach significantly outperforms other baseline multimedia retrieval approaches. In the future, we will apply similar multimodal non-linear fusion techniques in multimodal classification and clustering by employing several modalities, without significant increase in the computational and memory cost.

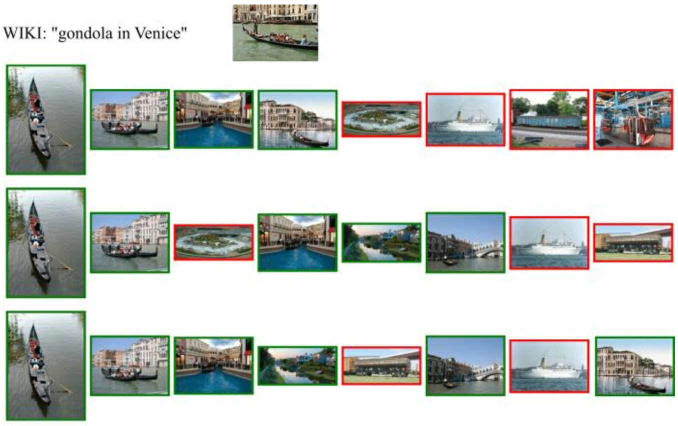

Fig. 2. The results for the query “gondola in Venice” with its associated image. In the first row, our hybrid framework retrieves 4 relevant objects, contrary to the non-linear fusion (second row) which retrieves 3 relevant objects. The unifying framework with two modalities (third row) retrieves only two relevant images out of seven retrieved results.

Acknowledgements

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

References

[1] P. K. Atrey, M. A. Hossain, A. El Saddik, & M. S. Kankanhalli, “Multimodal fusion for multimedia analysis: a survey”, Multimedia systems, 16(6), pp. 345-379, 2010

[2] T. Tsikrika, K. Andreadou, A. Moumtzidou, E. Schinas, S. Papadopoulos, S. Vrochidis, & I. Kompatsiaris, “A Unified Model for Socially Interconnected Multimedia-Enriched Objects”, In MultiMedia Modeling, pp. 372-384, Springer International Publishing, January 2015.

[3] J. Ah-Pine, G. Csurka, and S. Clinchant, “Unsupervised Visual and Textual Information Fusion in CBMIR Using Graph-Based Methods”, ACM Transactions on Information Systems (TOIS), 33(2), 9, 2015

[4] B. Safadi, M. Sahuguet, and B. Huet, “When textual and visual information join forces for multimedia retrieval”, In Proceedings of International Conference on Multimedia Retrieval, p. 265, April, 2014.

[5] S. Xu, H. Li, X. Chang, S. I. Yu, X. Du, X. Li, and A. Hauptmann, “Incremental Multimodal Query Construction for Video Search”. In Proceedings of the 5th ACM on International Conference on Multimedia Retrieval (pp. 675-678). ACM, June, 2015.

[6] T. Mei, Y. Rui, S. Li, and Q. Tian, “Multimedia search reranking: A literature survey”, ACM Computing Surveys (CSUR), 46(3), 38, 2014.

[7] Y. Wang, X. Lin, and Q. Zhang, “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.

[8] A. Kumar, and H. Daumé, “A co-training approach for multi-view spectral clustering”. In Proceedings of the 28th International Conference on Machine Learning (ICML-11), pp. 393-400, 2011.

[9] J. Jeon, V. Lavrenko, and R. Manmatha, “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, pp. 119-126, ACM, July, 2003.

[10] Y. Wang, F. Wu, J. Song, X. Li, and Y. Zhuang, 2014, “Multi-modal mutual topic reinforce modeling for cross-media retrieval”, In Proceedings of the ACM International Conference on Multimedia, pp. 307-316, ACM, November, 2014.

[11] D. M. Blei, and M. I. Jordan, “Modeling annotated data”, In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 127-134, ACM, July, 2003.

[12] J. Costa Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. R. Lanckriet, R. Levy, and N. Vasconcelos, “On the role of correlation and abstraction in cross-modal multimedia retrieval”, Pattern Analysis and Machine Intelligence, IEEE Transactions on, 36(3), pp. 521-535, 2014.

[13] J. Wang, Y. He, C. Kang, S. Xiang and C. Pan, “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, June, 2015.

[14] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos, “A new approach to cross-modal multimedia retrieval”. In Proceedings of the international conference on Multimedia, pp. 251-260, ACM, October, 2010.

[15] B. Siddiquie, B. White, A. Sharma, and L. S. Davis, “Multi-Modal Image Retrieval for Complex Queries using Small Codes”. In Proceedings of International Conference on Multimedia Retrieval, p. 321, ACM, April, 2014.

[16] W. H. Hsu, L. S. Kennedy, and S. F. Chang, “Video search reranking through random walk over document-level context graph”, In Proceedings of the 15th international conference on Multimedia, pp. 971-980, ACM, September, 2007.

[17] S. Clinchant, G. Csurka, F. Perronnin, and J. M. Renders, “XRCE’s participation to ImagEval”, In ImageEval workshop at CVIR, Vol. 4, 2007.

[18] S. Clinchant, G. Csurka, J. Ah-Pine, G. Jacquet, F. Perronnin, J. Sánchez, and K. Minoukadeh, “XRCE’s participation in wikipedia retrieval, medical image modality classification and ad-hoc retrieval tasks of ImageCLEF 2010”. In CLEF (Notebook Papers/LABs/Workshops), September, 2010.

[19] M. Grubinger, P. Clough, H. Müller, and T. Deselaers, “The IAPR-TC12 benchmark: A new evaluation resource for visual information systems”, In International Workshop OntoImage, pp. 13-23, May, 2006.

[20] T. Tsikrika, and J. Kludas, “The Wikipedia image retrieval task. In ImageCLEF, pp. 163-183, Springer Berlin Heidelberg, 2010.

[21] V. De Sande, E. A. Koen, T. Gevers, and G. M. Snoek. “Evaluating color descriptors for object and scene recognition”, Pattern Analysis and Machine Intelligence, IEEE Transactions on, 32, no. 9, pp. 1582-1596, 2010.

[22] H. Jégou, M. Douze, C. Schmid, and P. Pérez, “Aggregating local descriptors into a compact image representation”, In Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pp. 3304-3311. IEEE, 2010.

[23] B. Safadi, and G. Quénot, “Re-ranking by local re-scoring for video indexing and retrieval”, In Proceedings of the 20th ACM international conference on Information and knowledge management, pp. 2081-2084,. ACM, October, 2011.

[24] J. Daiber, M. Jakob, C. Hokamp, and P. N. Mendes, “Improving efficiency and accuracy in multilingual entity extraction”, In Proceedings of the 9th International Conference on Semantic Systems, pp. 121-124, ACM, September, 2013.

[25] J. Hafner, H. S. Sawhney, W. Equitz, M. Flickner, and W. Niblack, “Efficient color histogram indexing for quadratic form distance functions”, Pattern Analysis and Machine Intelligence, IEEE Transactions on, 17(7), pp. 729-736, 1995.

[26] C. Sanderson, and K. K. Paliwal, “Identity verification using speech and face information”, Digital Signal Processing, 14(5), pp. 449-480, 2004.