Retrieval of Multimedia Objects by Fusing Multiple Modalities

Ilias Gialampoukidis
ITI-CERTH
Thessaloniki, Greece
heliasgj@iti.gr

Anastasia Moumtzidou
ITI-CERTH
Thessaloniki, Greece
moumtzid@iti.gr

Theodora Tsikrika
ITI-CERTH
Thessaloniki, Greece
theodora.tsikrika@iti.gr

Stefanos Vrochidis
ITI-CERTH
Thessaloniki, Greece
stefanos@iti.gr

Ioannis Kompatsiaris
ITI-CERTH
Thessaloniki, Greece
ikom@iti.gr

ABSTRACT

Searching for multimedia objects with heterogeneous modalities is critical for the construction of effective multimedia retrieval systems. Towards this direction, we propose a framework for the multimodal fusion of visual and textual similarities, based on visual features, visual concepts and textual concepts. Our method is compared to the baseline method that only fuses two modalities but integrates all early, late, linearly weighted, diffusion and graph-based models in one unifying framework. Our framework integrates more than two modalities and high-level information, so as to retrieve multimedia objects enriched with high-level textual and visual concepts, in response to a multimodal query. The experimental comparison is done under the same memory complexity, in two multimedia collections in the multimedia retrieval task. The results have shown that we outperform the baseline method, in terms of Mean Average Precision.

1. INTRODUCTION

Multimedia retrieval systems are becoming more and more popular as there is a need for effective and efficient access to very large and diverse collections of multimedia objects, such as videos (e.g. YouTube and Netflix) and images (e.g. Flickr). Searching in such collections is challenging due to the heterogeneous media that each item in the collection may contain (e.g. text, images, and videos). Therefore, these multiple media and the different features that can be extracted from them, e.g. low-level visual descriptors (based on color, shape, location, etc.), low-level textual features (image captions, video subtitles, etc.), high-level textual or visual features (named entities, concepts, etc.), or metadata (timestamps, tags, etc.), need to be combined to support various multimedia analysis tasks, such as retrieval, summarization, clustering, and classification; this combination of multiple modalities is referred to as *multimodal fusion*.

Multimodal fusion has been widely investigated and is typically performed at the feature level (*early fusion*), at the decision level (*late fusion*), or in a *hybrid* manner (see [3] for a survey). This work focuses on the decision level or late fusion of multiple modalities for the multimedia retrieval task; to this end, several multimedia and cross-media approaches have been proposed. Metric fusion [13] is a random walk approach designed to fuse different "views" of the same modality, such as SIFT, GIST and LBP visual features; our focus though is on the combination of diverse modalities, such as textual and visual similarity scores. A recent video retrieval framework [9] proposes to fuse textual similarity scores based on video subtitles with visual similarity scores based on visual concepts in a simple non-linear way.

Other approaches have been motivated by Latent Dirichlet Allocation (LDA) and either generate a joint topic probability distribution for all modalities, or combine the topic distribution per modality [4]. Each query is related to a topic and the retrieved documents are assigned a topic distribution. If the topic distribution of a retrieved document is maximized at the query's topic, the document is considered to be relevant. Convolutional Neural Networks (CNN) have also been used to learn high-level features and combine two modalities (text-image pairs) for cross-modal retrieval [12]. A Partial Least Squares (PLS) based approach [10] that maps different modalities into a common latent space has also been used in an image retrieval task. Contrary to these approaches that require training, our focus is on unsupervised multimodal fusion. Furthermore, many of these approaches can only support monomodal queries, whereas our goal is to also cater for queries consisting of multiple modalities. It should be noted that multimedia search systems are often interactive and user feedback is incorporated for progressively refining the query (e.g. [14]); such relevance feedback approaches are beyond the scope of this work.

One unsupervised multimedia retrieval approach that has been recently proposed [2] combines textual and visual similarities by integrating into a unifying graph-based framework (i) a cross-media approach that not only considers the similarity of the query to the objects in the collection, but also the similarities among them [1] and (ii) a random walk approach for multimodal fusion, and in particular a video retrieval approach that links two objects (i.e. nodes in the

ICMR'16, June 06-09, 2016, New York, NY, USA
© 2016 ACM. ISBN 978-1-4503-4359-6/16/06…$15.00
DOI: http://dx.doi.org/10.1145/2911996.2912068

graph) with a weighted edge if there exists a multimodal similarity between them [6]. To decrease the complexity, the framework assumes that the textual part of a multimodal query "is the main semantic source with regard to the user information" [2] and thus it first filters out any object not in the top-l retrieved based on their textual similarity scores, and then applies graph-based techniques only to these l selected items. This graph-based framework includes as special cases all well-known early, late, weighted, diffusion-based, as well as graph-based fusion models, does not require users' relevance feedback, and has been evaluated in the context of multimedia retrieval tasks. Thus far, though, this framework has only considered two modalities.

2. METHODOLOGY

First, we briefly describe the graph-based framework proposed in [2] that supports the fusion of two modalities and, then, we present its extension to M modalities.

2.1 Background

As described above, the framework first selects the top-l multimedia documents based on their textual similarity to query q. All subsequent operations are performed on these l selected documents. First, the $1 \times l$ query-based similarity vectors on the textual and visual modalities, $s_t(q,.)$ and $s_v(q,.)$, respectively, are computed and are normalized so that their elements sum to one. Then, the $l \times l$ textual and visual similarity matrices, S_t and S_v, respectively, are computed for these documents and are normalized using: $(s(d,d') - \min s(d,.))/(\max s(d,.) - \min s(d,.))$, where $s(d,d')$ denotes the similarity between two documents d and d'. By denoting the regular matrix multiplication operation as "$\cdot$" and the (i,j) element of a matrix A as $A[i,j]$, this graph-based framework sets $x_{(0)} = s_t(q,.)$, $y_{(0)} = s_v(q,.)$, and defines the following update rule:

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

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

where D is the row-normalizing matrix, e is the $l \times 1$ vector of ones, and the operator $\mathbf{K}(x,k)$ takes as input a vector x and assigns a zero value to elements with score strictly lower than the k-th highest score in x. Following i iterations, the final ranking with respect toquery q is given by the linear combination of $s_t, s_v, x_{(i)}$ and $y_{(i)}$:

$$\mathit{score}(q) = \alpha_t s_t(q,.) + \alpha_v s_v(q,.) + \alpha_{tv} x_{(i)} + \alpha_{vt} y_{(i)} \quad (1)$$

under the restriction that: $\alpha_t + \alpha_v + \alpha_{tv} + \alpha_{vt} = 1$.

This framework includes all well-known weighted, graph-based and diffusion-based fusion techniques, as special cases of its parameters. For $a_{tv} = a_{vt} = 0$, Equation (1) becomes the weighted mean fusion model. For $a_t = a_v = a_{vt} = 0, \gamma = 0$ and sufficiently large number of iterations i, the model is the random walk approach [6]. For $a_v = a_{tv} = 0, \beta = 0, \gamma = 0$, and $i = 1$, the model is the cross-media approach [1].

In the experiments reported in [2], $\beta = 0, \gamma = 0.3, i = 1$ and $k = 10$ are recommended as the default parameters when fusing the top-l ($l = 1000$) results returned by text-based search. The weights in the linear combination of $s_t, s_v, x_{(1)}$ and $y_{(1)}$ are tuned in $\{0.1, 0.2, \ldots, 0.9\}$ and the best values are compared to the uniform weighting strategy ($\alpha_t = \alpha_v = \alpha_{tv} = \alpha_{vt} = 0.25$). The results show only an incremental increase in Mean Average Precision (MAP) in one dataset and no increase in the other datasets, indicating the potential effectiveness of this uniform weighting scheme.

2.2 Multimedia Retrieval Using M Modalities

Our aim is to extend the aforemonetioned graph-based framework to more than two modalities. Assuming that there are M modalities with corresponding $1 \times l$ query-based similarity vectors $s_m(q,.)$ and $l \times l$ similarity matrices S_m, $m = \{1, 2, \ldots, M\}$, we compute the following contextual similarity matrix for each modality m:

$$C_m = \left(1 - \sum_{w=1}^{M-1} \beta_w\right) S_m + \sum_{w=1}^{M-1} \beta_w S_{w \neq m} \quad (2)$$

The matrices C_m of Equation (2) are row-normalized so as to obtain the corresponding row-stochastic transition probability matrices P_m with elements:

$$P_m[i,j] = \frac{C_m[i,j]}{\sum_{j=1}^{l} C_m[i,j]} \quad (3)$$

For all modalities m, we set $x_{(0)}^m = s_m(q,.), m = \{1, \ldots, M\}$, and motivated by [2], we define the following update rule:

$$x_{(i)}^m \propto \mathbf{K}(x_{(i-1)}^m, k) \cdot \left[\left(1 - \sum_{\substack{w=1 \\ w \neq m}}^{M} \gamma_w\right) P_m + \sum_{\substack{w=1 \\ w \neq m}}^{M} \gamma_w s_w(q,.)\right] \quad (4)$$

Inspired by the model of Equation (1), we finally propose the vector of relevance score in response to query q:

$$\mathit{score}(q) = \sum_{m=1}^{M} \alpha_m s_m(q,.) + \sum_{m=1}^{M} \alpha'_m x_{(i)}^m \quad (5)$$

where

$$\sum_{m=1}^{M} \alpha_m + \sum_{m=1}^{M} \alpha'_m = 1 \quad (6)$$

For three modalities, for example, Equation (2) becomes:

$$C_1 = (1 - \beta_1 - \beta_2)S_1 + \beta_1 S_2 + \beta_2 S_3$$
$$C_2 = (1 - \beta_1 - \beta_2)S_2 + \beta_1 S_1 + \beta_2 S_3$$
$$C_3 = (1 - \beta_1 - \beta_2)S_3 + \beta_1 S_1 + \beta_2 S_2$$

The contextual similarity matrices $C_m, m = \{1, 2, 3\}$ are row-normalized to obtain $P_m, m = \{1, 2, 3\}$ using Equation (3), and the update rule (Equation (4)) becomes:

$$x_{(i)}^1 \propto \mathbf{K}(x_{(i-1)}^1, k) \cdot [(1-\gamma_2-\gamma_3)P_1 + \gamma_2 e \cdot s_2(q,.) + \gamma_3 e \cdot s_3(q,.)]$$

$$x_{(i)}^2 \propto \mathbf{K}(x_{(i-1)}^2, k) \cdot [(1-\gamma_1-\gamma_3)P_2 + \gamma_1 e \cdot s_1(q,.) + \gamma_3 e \cdot s_3(q,.)]$$

$$x_{(i)}^3 \propto \mathbf{K}(x_{(i-1)}^3, k) \cdot [(1-\gamma_1-\gamma_2)P_3 + \gamma_1 e \cdot s_1(q,.) + \gamma_2 e \cdot s_2(q,.)]$$

Finally, $\mathit{score}(q)$ is computed as in Equation (5), which linearly combines $s_m(q,.)$ and $x_{(i)}^m$ for $m = \{1, 2, 3\}$.

Figure 1 depicts our multimedia retrieval framework in the particular case of fusing three modalities, namely visual features, visual concepts and textual concepts. The top-l documents in the filtering step are obtained by using the textual concepts to index and retrieve each document in response to the q using the open-source Apache Lucene[1] system. Then,

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

the $l \times l$ similarity matrices: S_1 on visual descriptors, S_2 on visual concepts, S_3 on textual concepts and the corresponding $1 \times l$ query-based similarity vectors: $s_1(q,.)$ on visual descriptors, $s_2(q,.)$ on visual concepts and $s_3(q,.)$ on textual concepts are conputed. Finally, we fuse these similarity matrices and the query-based similarity vectors to get a single relevance score vector in response to query q: $score(q)$.

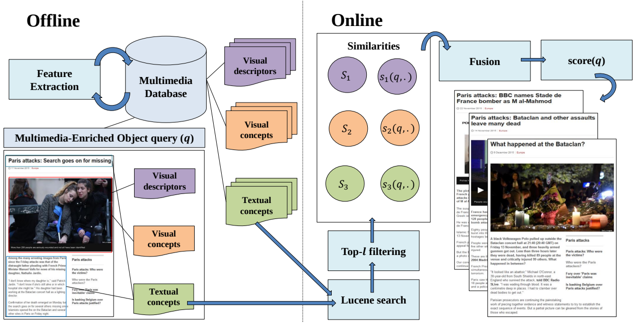

Figure 1: Multimedia Retrieval Framework by fusing 3 modalities

Memory Complexity. The memory complexity is $\mathcal{O}(l^2)$ for the computation of each similarity matrix S_m, $\mathcal{O}(l)$ for each similarity vector $s_m(q,.)$ and $\mathcal{O}(kl)$ for each $x^m_{(i)}, m = 1, 2, \dots, M$, 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 the number of filtered documents l', such that $Ml'^2 + Mkl' + Ml' = 2l^2 + 2kl + 2l$. The non-negative solution is:

$$l' = \sqrt{\frac{(k+1)^2}{4} + \frac{2l^2 + 2kl + 2l}{M}} - \frac{k+1}{2} \qquad (7)$$

For example, for $M = 3, k = 10, l = 1000$, we find $l' \cong 815$. We also observe that even for 15 modalities, the number of the top-l filtered documents remains > 300, hence a significant number of documents is involved in the fusion, even in the case of several modalities. We shall examine whether the fusion of three modalities in this framework outperforms the baseline approach, without additional memory cost.

3. EXPERIMENTS

In this section, we describe the datasets used for evaluation, the features extracted from their multimedia objects, the employed similarities, and the experimental results.

3.1 Evaluation Datasets

We evaluate our framework using two test collections: the IAPR-TC12[2] and the WIKI11[3] both created in the context of the ImageCLEF benchmarking activities. The IAPR-TC12 collection consists of (i) 20,000 images, each annotated with a title and a description, and also various metadata (e.g. date, location, etc.) not considered in this work and (ii) the 60 topics created in ImageCLEF 2007, each with a title and three image examples. The WIKI11 collection consists of (i) 237,434 images extracted from Wikipedia and their user-generated captions/descriptions and (ii) 50 topics, each with a title and one to five query images.

3.2 Features and Monomodal Similarities

We use the following state-of-the-art features and similarity functions for each modality in the documents and queries; it should be noted though that our method is capable of fusing any similarity score obtained by any kind of features. As visual descriptors (VD), we extract the scale-invariant local descriptors RGB-SIFT [11], which are then locally aggregated into one vector representation using VLAD encoding [7], and employ a similarity function based on the Euclidean distance [5]. We also employ the 346 high-level visual concepts (VC) introduced in the TRECVID 2011 semantic indexing task[4], which are detected by multiple independent concept detectors that use the aforementioned visual descriptors as input to Logistic Regression classifiers that have their output averaged and further refined [8]. Finally, we use the title/caption of each image so as to extract textual concepts (TC) using the DBpedia Spotlight[5] annotation tool. For the cases of visual and textual concepts, similarities are computed based on Lucene's retrieval function.

3.3 Experimental Setup and Results

We evaluate the MAP of our framework that fuses three modalities against the baseline (Section 2.1) that fuses two modalities. As the baseline models all well-known weighted, graph-based and diffusion-based fusion techniques as special cases of its parameters α, β, and γ, the best performance among all these fusion techniques coincides with the best performance of the weight parameters α, β and γ. Therefore, we tune these parameters so as to present, in Table 1, the best MAP scores of the fusion using two modalities. First, we combine textual concepts with visual descriptors (TC & VD), and then, we combine textual with visual concepts (TC & VC); the combination of visual descriptors and concepts (VD & VC) is not considered as it reduces the problem to the classic image retrieval task and no initial filtering can be performed with respect to the textual modality. We adopt the default parameters reported in [2], i.e. $k = 10$, one iteration ($i = 1$) and uniform weights ($a_t = a_v = a_{tv} = a_{vt} = 1/4$).

To compare directly the baseline method with our framework with three modalities under the same memory complexity, we use $l' = 815$ (see Equation (7)). For $m = \{1, 2, 3\}$, we adopt a uniform weighting strategy ($\alpha_m = \alpha'_m = 1/6$) and we tune the parameters γ_m, while the parameters β_m are kept constant (and equal to $1/3$); the results for different values of γ_m are reported in Table 1.

We observe that our framework outperforms the baseline method for several values of the parameters $\gamma_m, m = \{1, 2, 3\}$ under the same memory cost. In particular, MAP increases up to 13.44% for WIKI11 and up to 15.71% for IAPR-TC12. We further tuned the parameters α_m, α'_m and $\beta_m, m = \{1, 2, 3\}$ and we did not observe any further increase in MAP, which implies that there is no other weighted, graph-based or diffusion-based fusion techniques that outperforms our framework. The small differences in MAP when employing the best parameters $\gamma_m, m = \{1, 2, 3\}$ compared to the uniform ones ($\gamma_m = 1/3$) allows for setting $\gamma_m = 1/3, m = \{1, 2, 3\}$, without significant decrease in MAP (less than 2%).

[2] http://imageclef.org/photodata
[3] http://www.imageclef.org/wikidata
[4] http://www-nlpir.nist.gov/projects/tv2011/tv11.sin.346.concepts.simple.txt
[5] http://dbpedia-spotlight.github.io/demo/

Table 1: MAP values for both datasets; in bold the values which outperform the baselines.

M	γ_1	γ_2	γ_3	WIKI11	IAPR
2	TC & VD (best α, β, γ)			0.3654	0.2769
2	TC & VC (best α, β, γ)			0.3472	0.2729
3	0.00	0.00	1.00	0.3637	**0.3065**
3	0.00	1.00	0.00	0.3433	**0.2858**
3	1.00	0.00	0.00	**0.3855**	0.2518
3	0.50	0.50	0.00	**0.4083**	**0.2912**
3	0.5	0.25	0.25	**0.4145**	**0.2970**
3	0.25	0.50	0.25	**0.4029**	**0.3136**
3	0.25	0.25	0.50	**0.4048**	**0.3204**
3	0.34	0.33	0.33	**0.4105**	**0.3148**

4. CONCLUSION

We presented an unsupervised graph-based framework that fuses M modalities for multimedia retrieval. In this work, we consider the fusion of three modalities (textual concepts, visual descriptors, and visual concepts), but the overall framework is directly applicable to any number of modalities. We also presented a theoretical formula which provides the optimal number of documents that need to be initially filtered, so that the memory cost in the case of M modalities remains the same as in the case of two modalities. The experiments have shown that the MAP improves in the case of three modalities (up to 15.71% in some cases). We also observed that the results of a uniform weighting strategy do not significantly differ from those obtained using the best weights. In the future, we plan to evaluate our framework in multilingual settings using language agnostic features, such as textual concepts either in the language of the query or mapped to a common language using multilingual ontologies.

5. ACKNOWLEDGMENTS

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

6. REFERENCES

[1] J. Ah-Pine, S. Clinchant, and G. Csurka. Comparison of several combinations of multimodal and diversity seeking methods for multimedia retrieval. In *Multilingual Information Access Evaluation II. Multimedia Experiments*, pages 124–132. Springer, 2009.

[2] 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.

[3] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16(6):345–379, 2010.

[4] 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):521–535, 2014.

[5] 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):729–736, 1995.

[6] 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*, pages 971–980. ACM, 2007.

[7] 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*, pages 3304–3311. IEEE, 2010.

[8] 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*, pages 2081–2084. ACM, 2011.

[9] 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*, page 265. ACM, 2014.

[10] 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*, page 321. ACM, 2014.

[11] K. E. Van De Sande, T. Gevers, and C. G. Snoek. Evaluating color descriptors for object and scene recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1582–1596, 2010.

[12] 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*, pages 347–354. ACM, 2015.

[13] 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*, pages 805–810. ACM, 2013.

[14] S. Xu, H. Li, X. Chang, S.-I. Yu, X. Du, X. Li, L. Jiang, Z. Mao, Z. Lan, S. Burger, et al. Incremental multimodal query construction for video search. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 675–678. ACM, 2015.