LINKING TEXT AND VISUAL CONCEPTS SEMANTICALLY FOR CROSS MODAL MULTIMEDIA SEARCH

Bahjat Safadi, Mathilde Sahuguet and Benoit Huet

EURECOM Sophia Antipolis, France

ABSTRACT

Currently, popular search engines retrieve documents on the basis of text information. However, integrating the visual information with the text-based search for video and image retrieval is still a hot research topic. In this paper, we propose and evaluate a video search framework based on using visual information to enrich the classic text-based search for video retrieval. With the proposed framework, we endeavor to show experimentally, on a set of real world scenarios, that visual cues can effectively contribute to significant quality improvement of video retrieval. Experimental results show that mapping text-based queries to visual concepts improves the performance of the search system. Moreover, when appropriately selecting the relevant visual concepts for a query, a very substantial improvement of the system's performance is achieved.

Index Terms— Multimedia retrieval, video search and visual cues

1. INTRODUCTION

Since the last decade, multimedia documents continue to grow in a phenomenal way. In particular, videos constitute an increasingly popular mean to convey information, due to the ease of both capturing and sharing them. Hence, searching for relevant content is a crucial issue, as one may be overwhelmed by the amount of available information. Popular search engines retrieve documents on the basis of text information. This is especially the case for textual documents, but also for images and videos. Although videos are visually very rich, they cannot be directly exploited when searching for specific contents, due to the so called *semantic gap* [10]. Several research works attempt to integrate textual and visual information based on input images and/or on relevance feedback for multimedia retrieval [11, 13, 14, 15].

In this paper, we propose and evaluate a video search framework using high-level visual concepts in complementing text for video retrieval. We intend to report how much improvement this information can provide to the pure text-based search, and how we can tune this system to get better results. Indeed, we want to explore cross modality between textual and visual features: we know text is able to give valuable results, but looses the specificity of the information in videos, while visual features exploit this visual part but are not descriptive enough by themselves. We argue that improved retrieval can be achieved by combining textual and visual information to create an enriched query.

The originality of our work lies in the fact that we start from a text query to perform the visual search: we attempt to overcome the semantic gap by automatically mapping input text to semantic concepts. Text and visual concepts scores are calculated separately and results are combined by a late fusion method. This paper investigates the following questions: to which extent can visual concepts add information when retrieving videos? How can we cope with the confidence in visual concept detection? Recently, several research works have addressed these questions. We review below some of these methods.

Hauptmann et al. [6] have analyzed the use of visual concepts for video retrieval in the scenario of a news collection. The authors studied the impact of different factors: the number, the type and the accuracy of concept detectors. They concluded that it is possible to reach valuable results within a collection with fewer than 5000 concepts of modest quality. In their evaluation, they started from a query directly constituted of concepts, while we propose to automate the concept mapping from a text query.

The work of Habibian et al. [5] focuses on creating a concept detectors vocabulary for *event* recognition in videos. In order to derive useful concepts, they have studied the words used to describe videos and events. Their resulting recommendations were that concepts should be diverse, both specific and general. Moreover, vocabularies should have more than 200 concepts, and it is better to increase the number of concepts than the accuracy of the detectors.

The authors in [1] have addressed the question "how much can different features contribute to multimedia retrieval?". They have studied the impact of using different descriptors (textual and visual) for video hyperlinking. They concluded that textual features perform the best for this task, while visual features by themselves cannot predict reliable hyper-

This work was supported by the European Commission under contracts FP7-287911 LinkedTV and FP7-318101 MediaMixer.

links. Nevertheless, they suggest that using visual features for reranking results of a text-based search slightly improves the performance. In this paper, we endeavor to estimate how visual concepts can improve a search, depending on the way they are used.

Another aspect of our framework is the automatic linking of a textual query to visual concepts through a semantic mapping. Several works achieve this step by exploiting ontologies. In [12], the authors developed an OWL ontology of concept detectors that they have aligned with WordNet [4]. They question whether semantically enriching detectors helps in multimedia retrieval tasks. Similarly, an ontology based on LSCOM taxonomy has been developped [1], and has been aligned with ontologies such as DBpedia[2].

2. THE PROPOSED FRAMEWORK

This work focuses on the search of a known video segment in a video dataset, using a query provided by a user in the form of text. Indeed, writing text is the most straightforward mean for a user to formulate a query: the user doesn't need any input image when searching documents. We follow an approach in which a user provides such query that was recently taken by the MediaEval Search task [2]. In this situation, a query is constituted of two parts: the first part gives information for a text search, while the other part provides *cues* on visual information using words. Here, we give an example of such a query:

- **Query:** Medieval history of why castles were first built; **Visual cues:** Castle.

For the text-based search, the state-of-the-art methods perform sufficiently well. However, the visual cues are not straightforwardly understandable by a computer, since some queries are not so easy to interpret. As these visual cues can be any text words, it is a challenging task to have a visual model for every word of the text query. Thus, a basic candidate solution is to have a set of models for predefined visual concepts (the maximum it covers, the better it is), and to map each word to its closest predefined concept. Then, the models of the mapped concepts will be used as visual content models for each query.

Ideally, this mapping process should be done manually to avoid any intent gap between the query and the mapped concepts. However as it has a direct correlation to the number of the available concepts, it can be a very time consuming process that may be subject to personal interpretation and therefore error prone, especially when the number of concepts is of large-scale. Strong of these facts, this process should be automated, even knowing that it will provide some noise in the mapping.

[1]http://vocab.linkeddata.es/lscom/

[2]http://www.eurecom.fr/ atemezin/def/lscom/lscom-mappings.ttl

2.1. Our framework

Our proposed framework operates on any provided video collection with associated subtitles (or automatic speech recognition). First, we need to pre-process the video collection in order to extract and index features (i.e. text, concepts, scenes), which are needed by our work. Text search is straightforward with a search engine such as Lucene/Solr[3]. Nevertheless, it is different for a search based on visual features: incorporating visual information in the search task requires to design a complex framework that maps queries to a vocabulary of concepts and that is able to rank the videos segments accordingly. More details about the framework are available in [7].

In this work, we search for segments inside a video collection given a text query. Videos are pre-segmented into *scenes* and we extract textual and visual features (visual concepts) in order to give grounds to the search.

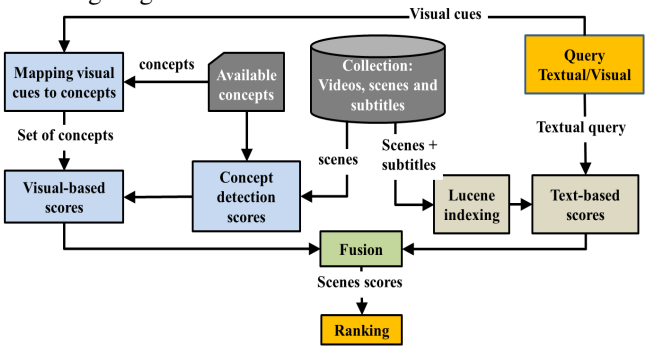

Fig. 1. Our multimodal video retrieval framework

2.1.1. *Text-based scores computation: T*

We have used the search platform Lucene/Solr for indexing textual features. We temporally aligned text from the subtitles to the scenes, performed base processing (converting to lower-case, stop-words removal, etc) and indexed each scene in Lucene/Solr together with its corresponding text. Then, we compute the text-based scores by using Lucene's default text search based on TF-IDF representation and cosine similarity.

2.1.2. *Visual-based scores computation: V*

In the visual cues description, the user provides a textual description of what are the visual characteristics of the video segment (s)he is looking for. As we propose to enhance the text-based search using visual concepts, we need a mapping between the text-based query and the concepts that should be found in the video, among the set of concepts that were computed.

For this mapping, we use the work reported in [8]. Keywords are extracted from the "visual cues" using the Alchemy API[4], and then some of those keywords are mapped with concepts for which a detector is available. This was done by

[3]http://lucene.apache.org/solr/

[4]http://www.alchemyapi.com/

Table 1. An example of concepts mapping with their associated confidence scores β to the visual query: "Castle"

c	Windows	Plant	Court	Church	Building
β	0.4533	0.4582	0.5115	0.6123	0.701

computing a semantic distance between the keyword and the names of the concepts, based on WordNet synsets. Hence, each keyword is aligned to several concepts with a confidence score: this score gives a clue on the proximity between the keyword and the concept.

In this work, we will study the impact of the *confidence score* β on the set of concepts C^q associated to each query q, through its text-based visual cues. We plan to compute the performance of the system with different thresholds θ that will, automatically, define the set of visual concepts, which should be included with each q. Given the set of concepts C^q for query q and a threshold θ, the selected concepts C'^q are those having $\beta \geq \theta$.

An example of concept mapping is given in table 1, where, the term *Castle* was mapped to five concepts (form the predefined set of concepts) with different associated confidence scores β-values.

For each query q, we compute the visual score v_i^q associated to every scene i as the following: $v_i^q = \sum_{c \in C'^q} w_c \times v_i^c$, where w_c is the valid detection rate of concept c, which is used as a weight for the corresponding concept detection score. v_i^c is the score of scene i to contain the concept c. The sum is made over the selected concepts C'^q. Notice that when $\theta = 0$, all the set of C^q is included. Therefore, evaluating the threshold θ is the main objective of this paper and this will be compared with two baseline methods: i) using only text-based search and ii) using text-based search with all available visual concepts C (e.g. the 151 visual concepts '$\theta = 0$').

2.1.3. Fusion between text-based and visual-based scores

Scores of the scenes (T) based on the text feature are computed for each query. Independently, we compute scores (V) based on visual attributes and apply late fusion between both in order to obtain the final ranking of items. After these scores are calculated, the score of each scene is updated according to its t_i and v_i scores. Many alternative fusion methods are applicable to such situation [3]. Here, we chose a simple weighting fusion function as follows:

$$s_i = t_i^{\alpha} + v_i^{1-\alpha} \tag{1}$$

where α is a parameter in a range of [0,1] that controls the "strength" of the fusion method. There are two critical values of α: $\alpha = 0$ and $\alpha = 1$. $\alpha = 1$ gives the baseline (i), which corresponds to the initial text-based scores only. $\alpha = 0$ uses the visual scores of the corresponding concepts only, which are expected to be very low on the considered task. However, this parameter has to be tuned by cross-validation within a development set or different subsets.

3. EXPERIMENTS

We conducted and evaluated our work on the MediaEval 2013 Search task [2]. The used dataset contains 2323 videos from the BBC, amounting to 1697 hours of television content of all sort: news shows, talk shows, series, documentaries, etc. The collection contains not only the videos and audio tracks, but also some additional information such as subtitles, transcripts or metadata. Along with this dataset comes a set of queries that matches exactly the description we gave in section 2. Those queries were created by 29 users who defined 50 search queries related to video segments watched inside the whole collection. To each query is associated the video segment seeked by the user, described by the name of the video, the beginning and end time of the segment inside the video. The used measures for the considered search task are: the Mean Reciprocal Rank (MRR), Mean Generalized Average Precision (mGAP), and Mean Average Segment Precision (MASP).

The visual scores were produced by applying the approach presented in [9], using a sub-set of ten different low-level descriptors. Each detector was used to train a linear SVM on 151 semantic concepts of TRECVid 2012 SIN task, which resulted in ten SVM-models for each concept. The same descriptors were computed on the considered dataset (i.e. Mediaeval 2013) and the models for each concept were used to predict the presence of the concepts at each video-shot of the dataset. A simple late fusion approach was applied to produce one score for each concept per shot. These scores are then normalized by the min-max function. We have no information about the quality of the trained models trained. Thus, all the mapped concepts are used with the same confidence w for each concept. In table 2, we report the Maximum and Mean number of concepts per query at each confidence level θ, as well as the number of queries that have at least one concept at each θ ($\#Q(\#c'^q > 0)$). It is clear that when θ increases, the number of associated concepts decreases, and when $\theta > 0.7$ very few concepts will be included for each query. Furthermore, there are only 24 queries that have at least one concept with a strong confidence score $\beta \geq 1.0$.

Table 2. Number of concepts associated to queries and the number of queries for each value of θ ($Q' = \#Q(\#c'^q > 0)$.

θ	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
max	94	86	53	40	23	17	9	5	3
mean	80	39	22	13	8	5	2	1	1
$\#Q'$	49					45	35	29	24

3.1. Optimizing the α parameter of the fusion function

MediaEval does not provide a relevant development set for the search task. However, we chose to tune the α parameter (equation 1) using different subsets of 20 queries. We have randomly chosen ten different subsets, each includes 20

queries out of the 50. The optimal value of α is likely to depend on the collection and the queries themselves. We run the evaluations with different values of α, including the two following cases: $\alpha = 1$ which is the baseline when using only text-based search, and $\alpha = 0$ that means only visual contents were used. The aim of tuning is to get the value of α that enables to obtain the best performance of our system.

Table 3 reports the optimal values of α for each threshold θ. These values were chosen after applying the majority vote on the ten selected subsets of different 20 queries each.

Table 3. The optimal α- values with different concepts selection thresholds θ

θ	≤ 0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
α	0.9	0.9	0.5	0.5	0.7	0.7	0.7	0.7

3.2. Evaluation on MediaEval test queries

We have evaluated the proposed method to find the best combination of visual concepts scores with text-based scores, in function of the confidence threshold (θ) of concept mapping. We have set the values of the α parameter as obtained by cross-validation (see Table 3).

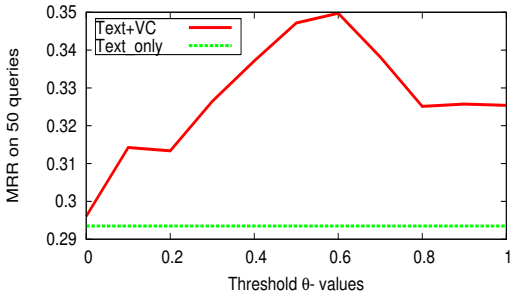

Fig. 2. MRR values on MediaEval queries with different θ-values, performance computed on all 50 queries.

Figure 2 shows the system performance (with MRR measure) when combining the visual content (selected using threshold θ) with the text-based search approach (i.e. Text+VC). When $\theta = 0$, all the predefined visual concepts are selected, and as the θ value increases, the number of selected concepts decreases. In other words, the θ values perform as a noise remover in the concept mapping. The system performance with the evaluation of θ is compared to the two considered baselines. As we can see, combining the visual scores of all concepts (i.e. $\theta = 0$) does not improve the text-based approach, while significant improvement can be achieved by combining only mapped concepts with $\theta \geq 0.1$ to each query. However, best performance is obtained when $\theta \geq 0.6$ and the gain comparing to the baseline approaches is about $19 - 20\%$.

We believe that the real improvement should be computed only on the 24 queries that contain at least one mapped visual concept when $\beta \geq 1.0$, since in other cases the visual information cannot be employed. The results on these 24 queries,

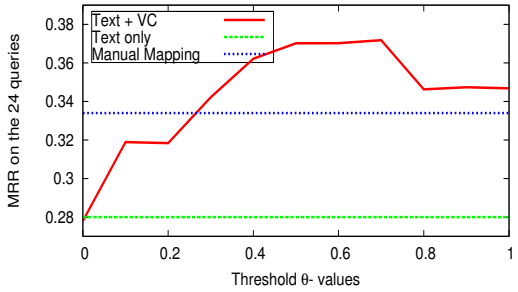

Fig. 3. MRR values on only 24 queries.

which have atleast one relevance visual concept, are reported in Figure 3. In addition to the two baselines, we compare the performance of our framework to performance with manual concepts mapping, where we have manually mapped the visual cues to concepts of TRECVid. For the all manually mapped concepts, we have set the confidence score $\beta = 1$. We report that the maximum number of concepts that were manually mapped to a query is 37 and the mean average is 22. This is almost equal to the values of $\theta = 0.4$ as shown in table 2. As we can see, concept mapping improves significantly the performance of the text-based search task on these queries. Moreover, the best performance was achieved with $\theta \geq 0.6$, with gain of about $32 - 33\%$ comparing to the text-based search system. Interestingly, the manual mapping improves the performance, and it is almost equal to the performance with the automated mapping between $\theta = 0.3$ and $\theta = 0.4$. We believe this is due to the fact that both cases have almost the same number of mapped concepts in average. However, as $\theta \geq 0.4$ the system with automated mapping outperforms the manual mapping. This concludes that using only concepts with high confidence values $\beta \geq 0.6$ (only few concepts per query) leads to better performance.

4. CONCLUSION

While popular search engines retrieve documents on the basis of text information only, this paper aimed at proposing and evaluating an approach to include high-level visual features in the search of video segments. A novel video search framework using visual information in order to enrich a text-based search for video retrieval has been presented. With the proposed framework, we endeavor to show experimentally, on a set of real world scenarios, that visual cues can effectively contribute to the quality improvement of video retrieval. We conducted our evaluations on the MediaEval 2013 search task. Experimental results show that mapping text-based queries to visual concepts is not a straightforward task, but when performed correctly it improves the performance of the search system. Moreover, with an appropriate concept mapping ($\beta \geq 0.6$) a significant improvement of about $32 - 33\%$ in MRR measure of the system's performance was achieved.

5. REFERENCES

[1] S. Chen, M. Eskevich, G. J. F. Jones, and N. E. O'Connor. An Investigation into Feature Effectiveness for Multimedia Hyperlinking. In *MMM14, 20th International Conference on MultiMedia Modeling*, pages 251–262, Dublin, Ireland, January 2014.

[2] M. Eskevich, R. Aly, R. Ordelman, S. Chen, and G. J. F. Jones. The search and hyperlinking task at mediaeval 2013. In *MediaEval*, Barcelona, Spain, October 2013.

[3] S. Essid, M. Campedel, G. Richard, T. Piatrik, R. Benmokhtar, and B. Huet. *Machine learning techniques for multimedia analysis*. Book chapter in "Multimedia Semantics: Metadata, Analysis and Interaction", July 2011.

[4] C. Fellbaum, editor. *WordNet: an electronic lexical database*. MIT Press, 1998.

[5] A. Habibian, K. E. van de Sande, and C. G. Snoek. Recommendations for Video Event Recognition Using Concept Vocabularies. In *Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval*, ICMR '13, pages 89–96, Dallas, Texas, USA, April 2013.

[6] A. Hauptmann, R. Yan, W.-H. Lin, M. Christel, and H. Wactlar. Can High-Level Concepts Fill the Semantic Gap in Video Retrieval? A Case Study With Broadcast News. *Multimedia, IEEE Transactions on*, 9(5):958–966, 2007.

[7] B. Safadi, M. Sahuguet, and B. Huet. When textual and visual information join forces for multimedia retrieval. In *International Conference on Multimedia Retrieval, ICMR '14*, Glasgow, United Kingdom, April 2014.

[8] M. Sahuguet, B. Huet, B. Cervenkova, E. Apostolidis, V. Mezaris, D. Stein, S. Eickeler, J. L. Redondo Garcia, and L. Pikora. LinkedTV at MediaEval 2013 search and hyperlinking task. In *MEDIAEVAL 2013, Multimedia Benchmark Workshop*, Barcelona, Spain, October 2013.

[9] P. Sidiropoulos, V. Mezaris, and I. Kompatsiaris. Enhancing Video concept detection with the use of tomographs. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, Melbourne, Australia, September 2013.

[10] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

[11] J. R. Smith and S.-F. Chang. VisualSEEk: a fully automated content-based image query system. In *Proceedings of the fourth ACM international conference on Multimedia*, MULTIMEDIA '96, pages 87–98, New York, NY, USA, 1996. ACM.

[12] C. G. M. Snoek, B. Huurnink, L. Hollink, M. de Rijke, G. Schreiber, and M. Worring. Adding Semantics to Detectors for Video Retrieval. *Multimedia, IEEE Transactions on*, 9(5):975–986, 2007.

[13] D. M. Squire, W. Müller, H. Müller, and J. Raki. Content-based query of image databases, inspirations from text retrieval: inverted files, frequency-based weights and relevance feedback. In *the 10th Scandinavian Conference on Image Analysis (SCIA'99)*, Kangerlussuaq, Greenland, June 1999.

[14] F. Thollard and G. Quénot. Content-Based Re-ranking of Text-Based Image Search Results. In *ECIR13,35th European Conference on IR Research*, pages 618–629, Moscow, Russia, March 2013.

[15] Z.-J. Zha, L. Yang, T. Mei, M. Wang, and Z. Wang. Visual query suggestion. In *Proceedings of the 17th ACM International Conference on Multimedia*, MM '09, pages 15–24, New York, NY, USA, 2009. ACM.