When textual and visual information join forces for multimedia retrieval

Bahjat Safadi
EURECOM
Sophia Antipolis, France
safadi@eurecom.fr

Mathilde Sahuguet
EURECOM
Sophia Antipolis, France
sahuguet@eurecom.fr

Benoit Huet
EURECOM
Sophia Antipolis, France
huet@eurecom.fr

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. The framework extends conventional text-based search by fusing together text and visual scores, obtained from video subtitles (or automatic speech recognition) and visual concept detectors respectively. We attempt to overcome the so called problem of semantic gap by automatically mapping query text to semantic concepts. 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. 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 significant improvement of the system's performance is achieved.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Abstracting methods, Indexing methods*; I.2.6 [**Artificial Intelligence**]: Learning—*Concept learning*

General Terms

Algorithms, Experimentation

Keywords

Multimedia retrieval, video search and visual cues

1. INTRODUCTION

Since the last decade, more and more multimedia documents are being published and consumed in many forms, mainly over the Internet. In particular, videos constitute an increasingly popular mean to convey information, due to the ease of both capturing and sharing them: it has become very common to record a video on a mobile phone and to upload it to a social sharing platform such as YouTube.

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. Several research works attempt to include visual information based on input images and/or on relevance feedback [19, 21, 14, 22].

In this work, we focus on the use of visual information to improve content retrieval in a video collection. Indeed, videos are visually very rich and it is not straightforward to exploit such data when searching for specific content. This phenomena is commonly called *semantic gap*: there is no direct or easy match between the meaning of a situation or an object, a concept, and the representation that can be made of it, in particular by a computer [18]. Indeed, there is a gap between the low-level features extracted from an image, and the high-level semantics that can be understood from it.

In this paper, we propose and evaluate a video search framework using visual information, in the form of visual concepts, for video retrieval. We intend to report how much improvement this information can provide to the 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.

Hence, we aim at designing a system that is able to query not only the textual features but also the visual ones. 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.

This work proposes to evaluate the use of high-level semantic concepts in complementing text for video retrieval. Text and visual concepts' scores are calculated separately and we apply a late fusion function to combine the results. We investigate the following two questions: i) to which ex-

ICMR '14, Apr 01-04 2014, Glasgow, United Kingdom
ACM 978-1-4503-2782-4/14/04.
http://dx.doi.org/10.1145/2578726.2578760.

tent can visual concepts add information when retrieving videos? ii) How can we cope with the confidence in visual concept detection? This paper contributes to answering the above questions by first, studying an effective approach for combining visual and textual information, then, investigating how reliable visual concept detectors should be to achieve better improved performance, of multi-modal search on video database.

The outline of the paper continues as follows: in section 2 we introduce the related work; then the task at hand and the framework used are presented in section 3; in section 4 we evaluate our proposed system and show the results that answer the main questions; finally, we give some conclusions and future works in section 5.

2. RELATED WORK

In this work, we perform an in-depth study of the use of visual features, which are represented here by the response of visual concept detectors, in multimedia retrieval. The study reported here, relates and extends the state-of-the-art in the following ways.

The work of Hauptmann et al.[9] analyses the use of visual concepts only for video retrieval in the scenario of a news collection. The authors study the impact of different factors: the number, the type and the accuracy of concept detectors. They conclude that it is possible to reach valuable results within a collection with fewer than 5000 concepts of modest quality. In their evaluation, they start from a query directly constituted of concepts, while we propose to automate the concept mapping from a text query. Nevertheless, they suggest the use of semi-automated methods for creating concepts-based queries.

Such work inspired the study of [7], although their focus is slightly different: they want to represent *events*. They aim at creating a concept detectors vocabulary for event recognition in videos. In order to derive useful concepts, they study the words used to describe videos and events. The resulting recommendations on the concepts are the following: concepts should be diverse, both specific and general. They also have results on the number of concepts to be used: vocabularies should have more than 200 concepts, and it is better to increase the number of concept than the accuracy of the detectors.

Hamadi et al. [8] proposed a method, denoted as 'conceptual feedback', to improve the overall detection performance that implicitly takes into account the relations between concepts. The descriptor of normalized detection scores was added to the pool of available descriptors, then a classification step was applied on this descriptor. The resulting detection scores are finally fused with the already available scores obtained with the original descriptors. They have concluded that significant improvement on the indexing system's performance can be achieved, when merging the classification scores of the conceptual feedback with their original descriptors. However, they have evaluated their approach on TRECVID 2012 semantic indexing task, which is based only on detecting semantic visual concepts, and no text-based queries was used.

How much can different features (textual, low-level descriptors and visual concepts) contribute to multimedia retrieval? The authors in [2] have addressed this question by studying the impact of different descriptors, both textual and visual ones, for video hyperlinking. They concluded that the textual features (in this case transcripts) performs the best for this task, while visual features by themselves (both low level and high level) cannot predict reliable hyperlinks, due to a great variability in the results. Nevertheless, they suggest that using visual features for reranking results obtained from a text 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 [20], the authors developed an OWL ontology of concept detectors that they have aligned with WordNet [6]. They question whether semantically enriching detectors helps in multimedia retrieval tasks. Similarly, an ontology based on LSCOM taxonomy [12] has been developped[1], and has been aligned with ontologies such as DBpedia[2].

3. TASK DEFINITION

3.1 Motivation

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 (for which (s)he would need to perform a preliminary image search or need drawing skills). We follow an approach in which a user provides such query that was recently taken by the MediaEval Search task [3]. 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 in the searched segments, using words. Here, we give two examples of such a query:

(1)
- *Text query:* Medieval history of why castles were first built
- *Visual cues:* Castle

(2)
- *Text query:* Best players of all time; Embarrassing England performances; Wake up call for English football; Wembley massacre;
- *Visual cues:* Poor camera quality; heavy looking football; unusual goal celebrations; unusual crowd reactions; dark; grey; overcast; black and white;

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 concept in the list. Then, the models of the mapped concepts will be used as visual content models for each query.

[1] `http://vocab.linkeddata.es/lscom/`
[2] `http://www.eurecom.fr/~atemezin/def/lscom/lscom-mappings.ttl`

Ideally, this mapping process should be done manually to avoid any intent gap between the query and the mapped concepts. However, this is a very time consuming process, which may be subject to personal interpretation and therefore error prone. Strong of these facts, this process should be automated, even knowing that it will provide some noise in the mapping. Our framework uses a predefined mapping between keywords from the visual cues and the visual concepts automatically computed using WordNet distances. The construction of such mapping is outside the scope of this paper and is detailed in [15].

Instead, we want to study how to perform a joint query combining text and visual concepts for video segment search. Using visual concepts relies on the accuracy of concept detectors, which can vary from one concept to the other. Hence, the query used should be carefully designed and take into account the confidence in different modules: concept mapping, concept detectors; It should also balance the part given to text and visual concepts in the search.

3.2 Evaluation

The main objective of our work is to study the impact of combining visual concepts with text-based queries in the context of video retrieval. We want to evaluate if concepts improve the text-based search and how to best combine them. We chose to use the MediaEval 2013 Search and Hyperlinking dataset, in order to enable replication and comparison of our work with other related works. The measures used for the considered search task are:

- the Mean Reciprocal Rank (MRR) assesses the ranks of the relevant segment returned for the queries. It averages the multiplicative inverse of the ranks of the correct answers (within a given time windows, here 60s).
- the Mean Generalized Average Precision (mGAP) is a variation of the previous that takes into account the distance to the actual relevant jump-in point. Hence, this measure also takes into account the start time of the segment returned.
- the Mean Average Segment Precision (MASP) assesses of the search in term of both precision of the retrieved segments and the length of the segments that should be watched before reaching the relevant content [4]. It takes into account the length of overlap between the returned segments and the relevant segment. It hence favors segments whose boundaries are close to the expected ones.

3.3 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. Figure 1 illustrates this framework.

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

3.3.1 Pre-processing

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.

Scenes segmentation.

As a video by itself is too long to present to the user, and it may not be relevant as a whole, we want to retrieve meaningful segments of video. Shots are too short segments, hence we define scenes as combinations of adjacent shots, that have temporal and visual consistency. We use the work proposed in [17]. This algorithm, based on an extension of the Scene Transition Graph (STG) [24], groups video shots by taking into account visual similarity (using HSV histogram comparisons) between temporally adjacent shots represented by keyframes.

Concepts detection.

For visual concept detection, we follow the approach presented in [16], which is based on the state-of-art for content-based multimedia indexing (CBMI). CBMI systems consists of two main phases: the modeling and indexing phases. In the modeling, the system should be extract different low-level features form a training set (the labeled set) to build different descriptors based on the content, such as Color-histograms, SIFT [11], Opponent-SIFT [23], bag-of-visual-words, etc. Then, for each concept a classifier (e.g. SVM) should be trained on each type of these descriptors to obtain a classification model. This model will be used to assign scores for new unlabeled samples (e.g. video-shots) as containing an instance of the learned concept.

The indexing phase is achieved by extracting the same descriptors on the test set, and using the learned model (on each descriptor) to predict the presence of the learned concept in these samples. Then, for each sample per concept, the system assigns a predicted score by fusing its scores from all the different models.

In this paper, we will directly use scores on the key-frame level, which are computed using a set of pre-calculated classification models. These models were trained on 151 predefined concepts from the complete list of concepts provided by TRECVid 2012 Semantic Indexing task (SIN) [13].

3.3.2 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.

3.3.3 Visual-based scores computation: V

Concept detector scores for each scene: v.

The concept scores extracted from the videos express the confidence that the corresponding concepts appears in the main frame of each shot. By extension, we assume that they represent the confidence of appearance for the entire shot.

We first normalize all the visual scores on a scale from

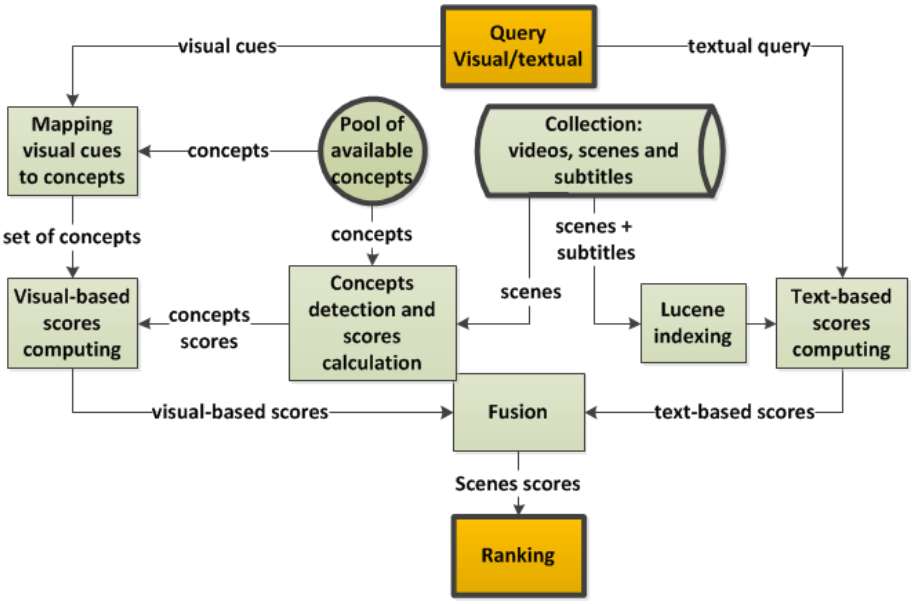


Figure 1: Our multimodal video retrieval framework

0 to 1 by a min-max normalization function. This function aims to scale the scores for each concept, so that they all fall in a range of l *to* u bounds. Thus, the visual scores values are normalized by subtracting the minimum and maximum score for each concept and then applying the following equation on each bin value:

$$v'_{ij} = l + \frac{(u-l) \times (v_{ij} - min_j)}{max_j - min_j} \quad (1)$$

where v_{ij} is the score of the j^{th} concept for the i^{th} frame, min_j and max_j are respectively the minimum and maximum score of the j^{th} concept, and u and l are the new dimension space. Results in v' are often normalized to the $[0, 1]$ range. Then, the visual score v of each scene is obtained by the mean average of its shots' scores.

Valid detection rate: w.

Concept scores are not normalized against each other: it is not possible to compare them, or to define a threshold that provides a boolean result (whether the concept is present or not present). Nevertheless, in order to have an insight on their performance, we manually created a valid detection score by examining the top 100 images for each concept and counting the number of true positive. The percentage of true positives found will be designated by *valid detection rate* w in the remaining of this paper.

Mapping text-based visual cues to visual concepts.

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 [15]. Keywords are extracted from the "visual cues" using the Alchemy API[4], and then each of those keywords is mapped with concepts for which a detector is available. This was done by computing a semantic distance between the keyword and the names of the concepts, based on Wordnet synsets [10]. Hence, each keyword was aligned to several concepts with a confidence score: this score gives a clue on the proximity between the keyword and the concept.

Table 1: Concepts mapped to the visual query from example 1: "Castle", with their associated confidence score β

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

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.

Computing visual scores regarding each query.

For each query q, we compute the visual score v_i^q associated to every scene i as the following:

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

$$v_i^q = \sum_{c \in C'^q} w_c \times v_i^c, \quad (2)$$

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 baselines: i) using only text-based search and ii) using text-based search with all available visual concepts C (e.g. the 151 visual concepts).

3.3.4 *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 [5, 1]. Here, we chose a simple weighting fusion function as follows:

$$s_i = t_i^\alpha + v_i^{1-\alpha} \quad (3)$$

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.

4. EXPERIMENTS

4.1 The dataset

We conducted our work on the dataset offered by the MediaEval 2013 Search and Hyperlinking task. MediaEval [3] is a benchmark initiative for multimedia evaluation including different tasks. The cited task tackles the issue of information seeking in a video dataset. The scenario proposed is that of a user looking for a known video segment, and who could be interested in watching related content proposed by the system. The 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 what we described in 3.1. 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.

4.2 Visual scores

To produce the visual scores we used the approach presented in [16], using a sub-set of ten different low-level descriptors calculated on key-frames. Each detector was used to train a linear SVM on 151 semantic concepts of TRECVid 2012 SIN task, these results 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 key-frame of our dataset. A simple late fusion approach was applied on the ten scores for each key-frame and results in one score for each concept per key-frame. These scores are then normalized by the min-max function. We have no information about the quality of the models trained on TRECVid 2012, since only the scores on the key-frames were provided to us. Thus, we have computed manually the performance of the models on the first ranked 100 key-frames for each visual concept, which have the maximum predicted scores for each concept. We have used these scores as the valid detection rate of each concept, denoted as w.

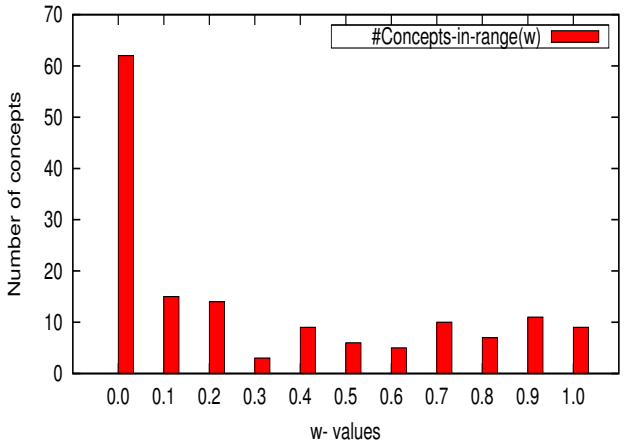


Figure 2: The predictor confidence scores of the visual concepts (w), for simplicity we show scores grouped in ten ranges.

Figure 2 shows, the histogram of w values. As this histogram shows, there are many concepts whose confidence score is equal to zero: $w = 0$. This means that these concepts will be ignored when calculating the visual scores according to function 2.

In this paper, we also compare the performance of the system using these confidence detector scores w and the case when having the same confidence for each detector, i.e. when $w = 1$.

4.3 Query mapping

Table 2 reports the minimum (Min), maximum (Max) and mean ($Mean$) number of concepts per query with different thresholds θ on the mapping confidence β. It also shows the number of queries that have at least one concept at each confidence level of θ ($\#Q(\#c'^q > 0)$). It is clear that when θ increases, the number of associated concepts decreases (see the Max and Mean values), and when $\theta > 0.7$ very few concepts will be included for each query. Furthermore, there are only 21 out of the 50 queries that have at least one concept with a strong confidence score (i.e. β) for the mapping (see $\#Q(\#c'^q > 0)$ with $\theta = 0.9$).

Table 2: Number of concepts associated to queries .

THR (θ)	Min	Max	Mean	$\#Q(\#c'^q > 0)$
0.0	5	45	20	50
0.1	5	45	19	50
0.2	5	41	18	50
0.3	2	37	15	50
0.4	0	25	11	49
0.5	0	19	7	49
0.6	0	19	5	48
0.7	0	12	3	44
0.8	0	6	1	29
0.9	0	2	1	21

4.4 Optimizing the α parameter of the fusion function

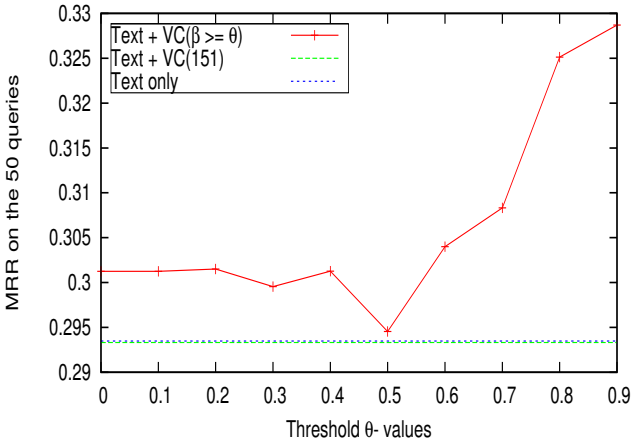
MediaEval does not provide a relevant development set for the search task. However, we chose to tune the α parameter (equation 3) using the aforementioned initial results (the text-based and the concept-based scores) with different subsets of 20 queries. We have randomly chosen ten different subsets, each includes 20 queries out of the 50. As mentioned before, the α parameter controls the range, in which we expect the visual content to improve the text-based search. The optimal value for this parameter 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 the tuning is to get the values of α that enable to obtain the best performance of our system.

Table 3 reports the optimal values of α for each threshold θ using the (manually computed) visual predictor confidences $w = Score(c)$ and the case when all concept confidences are the same $w = 1$. These values were chosen after applying the majority vote on the ten selected subsets of different 20 queries each. As we can see, the values of α for $\theta < 0.5$ are close to 0.9 in both cases, which means the effectiveness of the visual scores is very small comparing to the text-based system. Furthermore, for $0.5 \geq \theta < 0.7$, the α values are different between both cases, they are between 0.5 and 0.7. When $\theta \geq 0.7$, the values are stable and the influence of the visual scores is coherent.

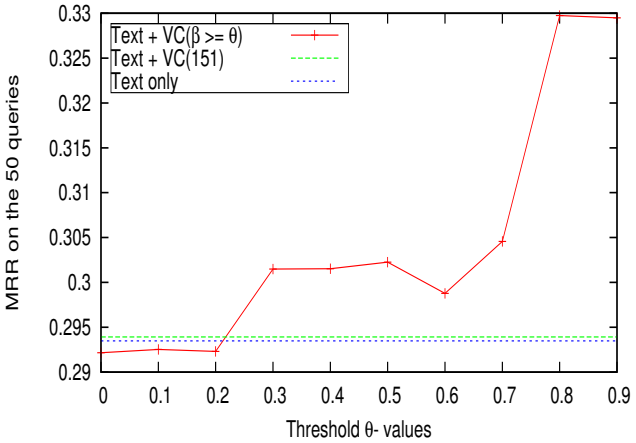
4.5 Evaluation on all 50 queries

The goal, of this experiment, is to study the influence of the visual concept mapping to text-based queries, that was done based on WordNet. 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). We have set the values of the α parameter as obtained by cross-validation (see Table 3), with the two confidence scoring ($w = Score(c)$ and $w = 1$).

Figure 3 shows the system performance (with MRR measure) when combining the visual content (selected using threshold θ) with the text-based search approach. The performance is shown with the two studied cases: when having a concepts validation rates $w = Score(c)$ (in 3(a)) and when $w = 1$ (in 3(b)). When $\theta = 0$, all mapped concepts (using the WordNet-based mapping) 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



(a) Concepts validation rates $w = Score(c)$



(b) Concepts validation rates $w = 1$

Figure 3: MRR values on the 50 queries with different θ-values using concepts validation rates: $w = Score(c)$ (a) and $w = 1$ (b).

concept mapping, and as it increases the number of mapped concepts decreases. Indeed, we want to study the impact of combining visual concepts with the text-based scores for query searching task. The system performance with the evaluation of θ is compared to the two aforementioned baselines: i) using the text-based scores only and ii) combining the text-based scores with the visual scores of the 151 visual concepts. As we can see in the two sub-figures, combining the visual scores of all concepts does not improve the text-based approach, while significant improvement can be achieved by combining only mapped concepts with $\theta \geq 0.3$ to each query. However, best performance is obtained when $\theta \geq 0.8$ and the gain comparing to the baseline approaches is about $11 - 12\%$ in both cases. The impact of the concept detector confidence (i.e. w) is not of that much importance, we believe that this may be due to the fact that many concepts have a valid detection rate $w = 0$. Thus, the use of $w = 1$ for each concept is a good choice for large values of θ. There is a strange bottom value with $\theta = 0.5$ using $w = Score(c)$ (the sub-figure 3(a)). We believe this is due to the noise in concept mapping, as well as the fact that many

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

	$\theta = 0.0$	$\theta = 0.1$	$\theta = 0.2$	$\theta = 0.3$	$\theta = 0.4$	$\theta = 0.5$	$\theta = 0.6$	$\theta = 0.7$	$\theta = 0.8$	$\theta = 0.9$
$w = Score(c)$	0.9	0.9	0.9	0.9	0.9	0.8	0.8	0.8	0.8	0.8
$w = 1.0$	0.9	0.9	0.9	0.9	0.9	0.5	0.5	0.7	0.7	0.7

concepts were mapped with $w = 0$ as a valid detection rate. However, when θ increases this noise is removed. The same performance was observed with both mGAP and the MASP measures, but for simplicity we report only the results with the MRR measure.

This experiment considers all the MediaEval search task queries (i.e. 50 queries), whether the visual task can be mapped to visual concepts or not. We believe that the real improvement should be computed on only the 21 queries that contain at least one mapped visual concept when $\theta \geq 0.9$ (according to table 2). In the next section we will report the performance on the subset of these 21 queries only.

4.6 Evaluation on a subset of 21 queries

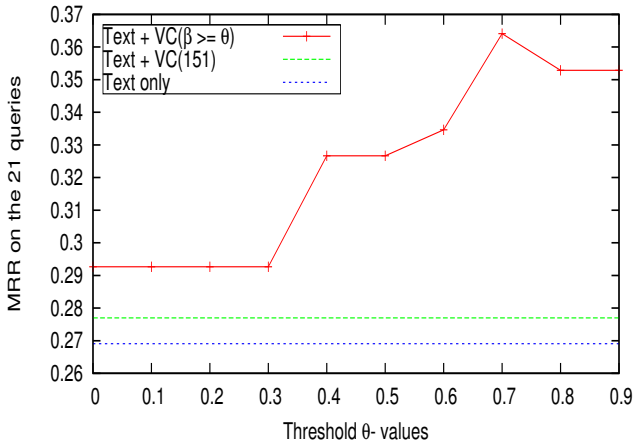
We have run the same evaluation as mentioned in the previous section but on only 21 appropriately selected queries. Each of these queries was mapped to at least one visual concept with high confidence mapping $\beta \geq 0.9$. This results on the 21 queries for which the visual information is important, and where the textual description maps to visual concept detectors with a high probability. Figure 4 shows the performance (in terms of MRR) of the 21 queries in function of the threshold θ, and again this is compared to the baselines on the same set of query.

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 $\theta \geq 0.7$ in both cases, with gain of about $32 - 33\%$ comparing to the text-based search system. This concludes that mapping text-based queries to concepts improves the performance of the search system. Furthermore, using only concepts with high confidence values $\beta \geq 0.7$ leads to better performance with gain about $32 - 33\%$.

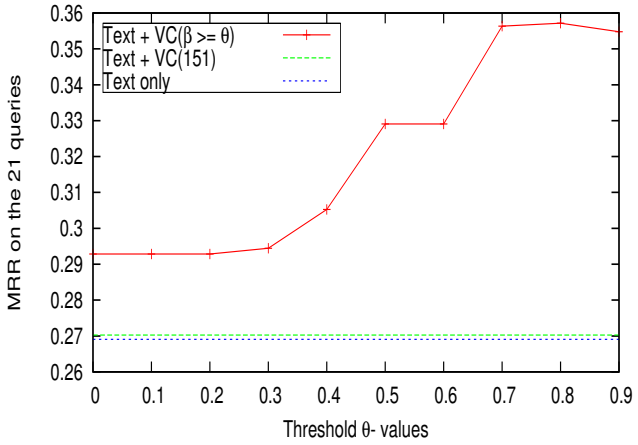
5. CONCLUSION

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. Starting from a textual query that includes some description of visual components of the searched segment, and we performed a search on a large video collection of television broadcast material by fusing text-based and visual-based scores at the scenes level in order to compute the final ranking. Indeed, we attempted to overcome the so-called problem of semantic gap by automatically mapping text from the query to semantic concepts, for which we have associated detectors.

We conducted our evaluations on the MediaEval 2013 search task. Experimental results show that carefully selecting the visual concepts related to a query improves the performance of the search system. Moreover, with an appropriate concept mapping ($\beta \geq 0.7$) a significant improvement of about $32 - 33\%$ in MRR measure of the system's performance was achieved.



(a) Concepts validation rates $w = Score(c)$



(b) Concepts validation rates $w = 1$

Figure 4: MRR values on only 21 queries that have minimum one concept with high confidence ($\beta \geq 0.9$) from WordNet, with different θ-values using concepts validation rates:$w = Score(c)$ (a) and using $w = 1$ (b)

As perspectives of this paper, we plan to study different concept mapping strategies. The work presented here is based on a simple mapping using WordNet, and further investigation in this direction are needed, in order to provide a stronger mapping. Furthermore, we plan to evaluate the effectiveness of our approach when applying the mapping on a larger number of visual concepts.

Acknowledgments

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

6. REFERENCES

[1] R. Benmokhtar and B. Huet. An ontology-based evidential framework for video indexing using high-level multimodal fusion. *Multimedia Tools and Applications*, pages 1–27, 2011.

[2] 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.

[3] 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.

[4] M. Eskevich, G. Jones, C. Wartena, M. Larson, R. Aly, T. Verschoor, and R. Ordelman. Comparing retrieval effectiveness of alternative content segmentation methods for Internet video search. In *CBMI12, the 10th International Workshop on Content-Based Multimedia Indexing*, pages 1–6, Annecy, France, June 2012.

[5] 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.

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

[7] 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.

[8] A. Hamadi, G. Quénot, and P. Mulhem. Conceptual Feedback for Semantic Multimedia Indexing. In *CBMI13, the 11th International Workshop on Content-Based Multimedia Indexing*, pages 53–58, Veszprèm, Hungary, June 2013.

[9] 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.

[10] D. Lin. An Information-Theoretic Definition of Similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 296–304, San Francisco, CA, USA, 1998.

[11] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004.

[12] M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-Scale Concept Ontology for Multimedia. *IEEE MultiMedia*, 13(3):86–91, July 2006.

[13] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, B. Shaw, W. Kraaij, A. F. Smeaton, and G. Quénot. TRECVID 2012 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics. In *Proceedings of TRECVID 2012*. NIST, USA, 2012.

[14] T. Quack, U. Mönich, L. Thiele, and B. S. Manjunath. Cortina: a system for large-scale, content-based web image retrieval. In *Proceedings of the 12th annual ACM international conference on Multimedia*, MULTIMEDIA '04, pages 508–511, New York, NY, USA, 2004. ACM.

[15] 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.

[16] 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.

[17] P. Sidiropoulos, V. Mezaris, I. Kompatsiaris, H. Meinedo, M. Bugalho, and I. Trancoso. Temporal Video Segmentation to Scenes Using High-Level Audiovisual Features. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(8):1163–1177, August 2011.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] K. van de Sande, T. Gevers, and C. Snoek. Evaluating Color Descriptors for Object and Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, September 2010.

[24] M. Yeung, B.-L. Yeo, and B. Liu. Segmentation of video by clustering and graph analysis. *Computer Vision and Image Understanding*, 71(1):94–109, July 1998.