Multimodal Reranking of Content-based Recommendations for Hyperlinking Video Snippets

Chidansh Bhatt
Idiap Research Institute
Martigny, Switzerland
cbhatt@idiap.ch

Nikolaos Pappas
Idiap and EPFL
Martigny, Switzerland
npappas@idiap.ch

Maryam Habibi
Idiap and EPFL
Martigny, Switzerland
mhabibi@idiap.ch

Andrei Popescu-Belis
Idiap Research Institute
Martigny, Switzerland
apbelis@idiap.ch

ABSTRACT

In this paper, we present an approach for topic-level search and hyperlinking of video snippets, which relies on content-based recommendation and multimodal re-ranking techniques. We identify topic-level segments using transcripts or subtitles and enrich them with other metadata. Segments are indexed in a word vector space. Given a text query or an anchor, the most similar segments are retrieved using cosine similarity scores, which are then combined with visual similarity scores, computed as the distance from the anchor's visual concept vector. This approach has performed well on the MediaEval 2013 Search and Hyperlinking task, evaluated over 1260 hours of BBC TV broadcast, in terms of overall mean average precision. Experiments showed that topic-segments based on transcripts from automatic speech recognition level systems (ASR) led to better performance than the ones based on subtitles for both search and hyperlinking. Moreover, by analyzing the effect of multimodal re-ranking on hyperlinking performance, we emphasize the merits of rich visual information available in the anchors for the hyperlinking task, and the merits of ASR for large-scale search and hyperlinking.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems

Keywords

Topic segmentation; video search; video hyperlinking.

1. INTRODUCTION

The explosive growth and the widespread accessibility of media content on the Web have led to a surge of research activities in multimedia search and hyperlinking. Users from a variety of backgrounds, such as knowledge workers from the creative industry, journalists, students, researchers, and home users can benefit from effective search and hyperlinking of content. Traditionally, search scenarios presented in the information retrieval literature aim to find information resources relevant to an information need from a given collection, and to reduce the information overload problem.

Hyperlinking a multimedia item to several others, based on similarity or relatedness, can be achieved by recommender systems, which are systems that seek to predict the preference that a user would give to an item. Although searching or linking video to additional information sources seems to be a sensible approach to satisfy the users' information needs, the perspective of users in real-life scenarios is still not fully understood. Various scenarios of use for video hyperlinking can be considered: (a) interactive non-linear access to videos, allowing users to generate narratives by following links in a video; (b) improving the entertainment value by enriching one medium with another one; and (c) exploring additional information sources while accessing content in a linear fashion [2]. Also, the users' perspective on what the links should entail may vary according to different scenarios.

In this paper, we present a solution for video search and hyperlinking which answers the user-centric requirements defined in the MediaEval 2013 Search and Hyperlinking (S & H) task [8]. The system we have proposed targets scenario (c) from above, where a user first searches for a known segment in a video collection ("search" sub-task); then, occasionally, finds that information insufficient and wishes to watch other related segments ("hyperlinking" sub-task). In other words, the search sub-task requires finding a determined segment of a TV show based on a query that was built with a "known item" in mind. The hyperlinking sub-task requires finding items from the collection that are related to "anchors", which are segments within known items, possibly using their larger context. The organizers of the task at MediaEval 2013 have provided 1260 hours of broadcast TV shows from BBC as test material, along with 50 test queries for searching, and 98 test anchors for hyperlinking.

The key issue – the relevance of the search/linking results with respect to the search/linking query – still remains a challenging and open research problem. Due to the success of text search, most popular video search engines such as Google or Yahoo! build upon text search techniques by using the non-visual information (such as surrounding text and user-provided tags) associated with visual content. On the one hand, the literature suggests that this approach to multimedia search or hyperlinking cannot always achieve satisfying results as it entirely ignores the visual content [28]. But on the other hand, it has been found that the visual in-

ICMR '14, April 1–4, 2014, Glasgow, United Kingdom
Copyright 2014 ACM 978-1-4503-2782-4/14/04 ...$15.00.
http://dx.doi.org/10.1145/2578726.2578752.

formation obtained by state-of-the-art processing techniques degrades the search performance on these tasks [9]. Also, the difficulties of automatically segmenting videos into shorter retrieval units are not mentioned clearly.

The unified approach to both search and hyperlinking proposed in this paper is based on techniques inspired from content-based recommender systems for multimedia recordings, which provide the most similar audio-visual segments to a given text query or to another segment, based on words. We consider an information-filtering rather than an information retrieval approach, targeting a scenario in which users prefer to provide more information in their search queries in exchange for a potentially smaller number of search iterations until they find the desired result. The same approach is used for hyperlinking, but in addition we use the visual concepts detected in the anchor segment and in the indexed ones in order to re-rank answers based on visual similarity. In other words, while textual descriptions are convenient as queries for known-item search (rather than asking users for visual query examples), anchors with richer multi-modal content can advantageously be used for hyperlinking.

The paper is organized as follows. In Section 2, we briefly review the literature on search, hyperlinking, content-based recommendation, and multimodal re-ranking. In Section 3, we present the components of the proposed system: topic segmentation, segment search and hyperlinking. In Section 4, we introduce the dataset, and then present and discuss evaluation results for the search and hyperlinking subtasks at MediaEval 2013, as well as additional evaluations done after the campaign.

2. RELATED WORK

An integrated system for the MediaEval 2013 Search and Hyperlinking task requires the combination of methods for multimedia search and for the automated creation of hyperlinks between items. In previous work, the search tasks attempted at the TRECVid workshops envisioned a scenario where the user poses a multimodal query and the system returns the most relevant video shots from a collection of videos [20]. Multimedia information retrieval [16] or searching spoken document archives [15] are well-known research topics, with a considerable literature and evaluation campaigns. However, the goal of hyperlinking multimedia segments has only emerged as an important use case more recently, and the MediaEval benchmarking evaluation took the initiative to clearly formulate such a standardized task. In an early idea put forward by Google researchers [14], a "query-free" system was designed for enriching television news with articles from the Web, using queries derived from the closed-captioning text.

The techniques proposed in this paper are inspired from previous work on content-based recommender systems [17]. Such systems use similarities between items computed from descriptors of their content. We will consider vector space models to define such similarities, and more specifically a model with tf-idf coefficients [26], but "semantic" spaces using some form of dimensionality reduction such as LDA, LSI, or Random Projections could also be used, as in experiments on content-based recommendation of multimedia on which we reported elsewhere ([21], Section IV).

Longer video programs, such as those provided for MediaEval 2013 [8], need to be divided into shorter video segments to provide the result links. This segmentation can be done at the shot level based on the visual channel, on sentences or speech segments from ASR transcripts, on temporal pauses, on lexical cohesion features, or can simply use fixed-length segments [9]. While most of the segmentation approaches are pre-determined for generating video segments over a collection that do not change for each query, it is possible also to apply on-the-fly segmentation based on each query [23]. A survey on such visual-based and audio-based shot segmentation is provided in [7].

The re-ranking of multimedia search results has received increasing attention in recent years. This is defined as the reordering of visual documents based on the information contained in the initial search result set, or based on a knowledge base, in order to improve search performance [29]. Among four paradigms for research on visual search re-ranking (self re-ranking, example-based re-ranking, crowd re-ranking, and interactive re-ranking), we chose Linear Multimodal Fusion (LMF), an example-based re-ranking approach. LMF [19, 12] is the most straightforward and easy-to-implement re-ranking method, merging several ranked lists by linearly combining the relevance scores for each document. Moreover, the query-independent approaches might have limited effectiveness because the optimal combination strategies usually vary considerably for different query topics. We explored the results with various weights to identify the influence of each modality on the scores.

3. METHODOLOGY

3.1 System Overview

The proposed system makes use of several components, represented as rectangles with sharp corners in Figure 1. We generate the data units, namely topic-based segments, from the human-made subtitles provided by the BBC or from the ASR transcripts. We experimented with transcripts from LIMSI [11] and from LIUM [27], both kindly provided to the MediaEval participants. The topic segmentation was performed over the words using the TextTiling algorithm implemented in NLTK.

For search, we compute word-based similarity (from transcript and metadata) between queries and all segments in the collection, using a vector space model and tf-idf weighting. Similarly, for hyperlinking, we first rank all segments based on their similarity with the anchor. In addition, we use the visual concept detection provided to the participants by the organizers: key frames from Technicolor [18] and concepts detected by Visor [6]. We thus generate a score matrix and then the list of nearest neighbors. Scores from text and visual similarity are fused to re-rank the final linking results.

In the following sections, we provide details about each of these components.

3.2 Topic Segmentation

Topic segmentation was performed over subtitles and transcripts using TextTiling [13] as implemented in the NLTK toolkit [5] (available at http://nltk.org/). Topic shifts are determined based on the analysis of lexical co-occurrence patterns, which are computed from 20-word pseudo-sentences, to ensure uniform length. Then, similarity scores are assigned at sentence gaps using block comparison. The peak differences between the scores are marked as boundaries, which we fit to the closest speech segment break. We selected TextTiling for its robustness and simplicity, although

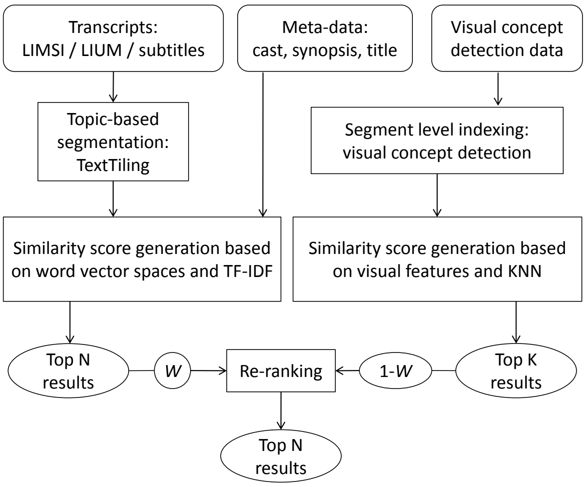

Figure 1: Overview of the proposed system for search and hyperlinking. Rectangles with sharp corners represent components of the system, while rounded boxes represent data.

more advanced techniques such as TopicTiling [24] (same core algorithm but with LDA topic modeling) are also available and could be tested in the future.

Table 1 shows the total number of segments, the average segment size (in seconds) and the standard deviation STD (in seconds) for each of the three alternative transcript types of the BBC TV shows available for segmentation: subtitles vs. ASR from LIMSI [11] vs. ASR from LIUM [27]. The longer size of the LIUM-based segments and the larger variability of subtitle-based segments should be noted. We also found some mismatches between the durations in metadata files and the timing found in the subtitles and the LIMSI transcripts (for 488 respectively 956 videos) and discarded the corresponding segments from further computation and evaluation.

Data	Number of Segments	Average size	STD	Mismatches
Subtitles	114,448	53	287	488
LIMSI	111,666	53	68	956
LIUM	84,783	68	64	738

Table 1: Topic segmentation statistics (size and STD are in seconds).

3.3 Segment Search

Segment search was performed by indexing the text segments in a word vector space with tf-idf weight [26], representing each textual query as well as the words from the "visual cues" provided with them, into the same space. We retrieved the most similar segments to the query using cosine similarity in the word vector space.

In other words, the tf-idf weights w_{ij} for a given segment i were (classically) computed as $w_{ij} = tf_{ij} \cdot idf_j$, where tf_j is the term frequency of word j in document d_i and idf_j is the inverse document frequency of word j. The similarity between two segments $\vec{s}_i$ and $\vec{s}_j$ was then computed by the cosine similarity between them as follows:

$$sim_{cos}(\vec{s}_i, \vec{s}_j) = \frac{\vec{s}_i \cdot \vec{s}_j}{||\vec{s}_i||_2 \times ||\vec{s}_j||_2} \quad (1)$$

Text	Features	# words	Total rank
LIMSI	1-gram	10k	500
LIMSI	1-gram	20k	453
LIMSI	1-gram	50k	417
LIMSI	1-gram	10k	606
LIMSI	1-gram + 2-gram	20k	573
LIMSI	1-gram + 2-gram	50k	540
LIMSI	1-gram + 2-gram + 3-gram	10k	606
LIMSI	1-gram + 2-gram + 3-gram	20k	573
LIMSI	1-gram + 2-gram + 3-gram	50k	517

Table 2: Combination of features and number of words for the LIMSI transcript ordered by *increasing performance* for each n-gram combination on the development set. The total rank score is the sum of the ranks of the correct answers for all the queries.

We have generated four sets of results for the set of test queries provided for the MediaEval 2013 search task. Of course, more configuration can be tested, but at MediaEval the total number of runs per participant was limited at 5.

(a) using the LIMSI ASR transcript;

(b) using the LIUM ASR transcript;

(c) using the LIUM ASR transcript plus the metadata associated in the dataset to each recording (cast, synopsis, series, and episode name) – the words from the metadata were appended to each segment;

(d) using human-made the subtitles.

For all the above texts we followed standard pre-processing to create the word vector representation, namely conversion to lower-case, tokenization, and stop word removal, using the NLTK library. Then, we determined which type of tf-idf features performed better for a given segmentation of the text on the development set. For example, we examined whether unigrams affect the performance compared to bigrams and how to select the vocabulary size such that we obtain appropriate representations.

We thus tested several parameters on the small (with 4 queries) development set provided before the MediaEval campaign, with the LIMSI transcript: the order of n-grams (1, 2, or 3) and the size of the vocabulary (10k, 20k, 30k, 40k, 50k words). In Table 2 we list the combinations of features that have been evaluated for the LIMSI ASR transcript on the development set, ordered by performance. The performance measure that we used was not the metric finally adopted for MediaEval (see Section 4.2 below), but we simply added the ranks of the correct answers for all the queries, thus the lower the better.

Since the development set is too small, we cannot select our method only best on the lowest scores (because they might mean overfitting the queries) but we need also to select a combination that will most likely generalize better in the test set. Thus, we selected the combination that captures

Text	Features	# words	Total rank
Subtitles	1-gram + 2-gram + 3-gram	50k	2008
LIUM + Meta	1-gram + 2-gram + 3-gram	50k	1077
LIUM	1-gram + 2-gram + 3-gram	50k	1043
LIMSI	1-gram + 2-gram + 3-gram	50k	517

Table 3: Ranking of four different types of text by *increasing performance* on the development set. The total rank score is the sum of the ranks of the correct answers for all the queries.

the most relations between words and has the best scores among the different vocabularies i.e. 1-gram + 2-gram + 3-gram for a vocabulary of 50k words, as seen in Table 2. With these features, we found on the development set that the LIMSI ASR transcript performed best, followed by LIUM, LIUM with metadata, and subtitles, as shown in Table 3. Note that we did not calibrate the size of the segments by TextTiling parameters. We submitted the results of four runs for the MediaEval search sub-task on the evaluation set.

3.4 Segment Hyperlinking

For the hyperlinking of new segments to given anchors, indexing is performed as above, though using only unigrams and a vocabulary of 20,000 words. Two scenarios were considered by the MediaEval organizers. As per the condition for the first scenario, the participants were restricted from using the broader boundaries of the initial segment that required linking. In the second scenario, the participants were allowed to take into account information about the broader boundaries of the initial segment that required linking.

Thus, in the first one (noted 'A' for anchors only), only the anchor information was given, therefore we extended the anchor text with text from segments containing/overlapping the anchor boundaries. For the second scenario (noted 'C' for anchors plus context), we considered the text within the start time and end time of the provided known-item that included the anchor, along with text from segments overlapping the known-item boundaries. We also enriched the subtitle or ASR texts using the textual metadata (title, series, episode) and webdata (cast, synopsis). The segments and anchors were indexed into a vector space with tf-idf weights, and the top N most similar segments were found by cosine similarity.

Then, we re-ranked results based on visual feature similarity, using the visual concept detection scores per keyframe, provided by the MediaEval organizers. Keyframes were first aligned to topic-based segments using shot information [18], with an average of 5 keyframes per segment. Similarly, this was performed for the anchors average of 8 frames and anchors plus contexts average of 55 frames. For each segment, we generated a visual feature vector using the concepts with the highest scores from the keyframes of the segment. We also generated the visual feature vector for each anchor and anchor plus contexts.

Using a K-nearest neighbors method, we ranked all segments by decreasing visual similarity to each anchor. Using the Scikit-learn toolkit in Python, with a ball tree datastructure and Euclidean distance, we generated the K nearest neighbors (segments) for each anchor. Typically, given that we intended to provide N = 1,000 links per anchor, we chose $K \gg N$, around 100,000. Then, we re-ranked text-based results using the visual scores of these segments, respectively with weight W for the text-based ranks and $1-W$ for the visual ranks. For the MediaEval workshop submission, we chose $W = 0.8$ when using text from subtitles, as we assumed that they are almost entirely accurate, and W = 0.6 when using text from ASR transcripts, assumed to contain more noise. Finally, we ignored segments shorter than 10 seconds and chunked larger segments into 2-minute segments, following guidelines received from the organizers of the task.

We submitted three runs to MediaEval for human evaluation of relevance: two with the subtitle words (scenarios A and C) and one with the LIMSI ASR transcript (scenario C). To extend our investigation after the MediaEval evaluations, and to take advantage of human ratings kindly made available by the organizers, we furthermore experimented with $W \in \{0, 0.2, 0.4, 0.5, 0.6, 0.8, 1.0\}$ to investigate the actual effect of different weights on hyperlinking performance.

4. EXPERIMENTAL RESULTS

As already stated, the Idiap system for MediaEval 2013 [3] has been evaluated on the search and hyperlinking sub-tasks. The dataset and evaluation metrics are those described in the MediaEval Search and Hyperlinking task guidelines and outlined in [8][1]. We provide below our results and a detailed discussion on achieved performance in the light of different features, approaches and other observed results, as well as our additional testing after the campaign.

4.1 Dataset

The dataset for both subtasks was a collection of 2,323 videos, totaling 1,260 hours of broadcast, kindly provided by the BBC. The average length of a video was roughly 30 minutes (varying from less than 10 minutes to 2 hours and more) and most videos were in English. The BBC kindly provided human generated textual meta-data and manual transcripts (subtitles) for most of the videos. The output of two automatic speech recognition (ASR) systems (LIMSI-CNRS / Vocapia and LIUM cited above in Section 3.1) was provided, along with visual analysis results: 1,200,000 shots / keyframes extracted by Technicolor and 1,000 visual concept probabilities for the top 10,000 key-frames from the on-the-fly video detector Visor. We utilized all the provided data except the INRIA-generated 10 similar face information for each of provided 557,324 faces in the key-frames,because we found that this was indicating face-detection information (location of a face in the frame) but did not contain the face recognition information (labeling of faces).

For the definition of realistic queries and anchors, the MediaEval S&H organizers conducted a study with 30 users which resulted in 50 known-items and the corresponding queries used for the search evaluation task. Subsequently,

[1] The guidelines are available online at http://multimediaeval.org/mediaeval2013/hyper2013/.

Submission	MRR	mGAP	MASP
Subtitles	0.064	0.044	0.044
LIUM + Meta	0.085	0.054	0.053
LIUM	0.090	0.058	0.057
LIMSI	**0.110**	**0.060**	**0.060**

Table 4: Results of our system (Idiap 2013) for the search task measured by MRR, mGAP and MASP.

users were asked to mark so-called anchors, or segments, within the known-item for which they would like to see links. This resulted in 98 anchors given to the participants in the task, among which a random set of 30 were used for actual evaluation of the hyperlinking task, performed as described below.

4.2 Evaluation Metrics

For the search sub-task, three metrics were used to evaluate the submissions of the participants: mean reciprocal rank (MRR), mean generalized average precision (mGAP) [22] and mean average segment precision (MASP) [10]. The "jump-in" point of the relevant content was considered over a time window of 10, 30 or 60 seconds.

To evaluate the hyperlinking subtask, the organizers resorted to crowdsourcing via Amazon's Mechanical Turk platform. They used a pooling method to group the videos from the top 10 ranks of the submitted runs of each of the participants (but no more than five runs per participant). This resulted in 9,195 anchor-target pairs, which represented 7,637 different pairs for crowdsourcing assessment. These assessments provided the ground truth used to calculate precision at fixed rank cutoffs and mean average precision (MAP) for all the participants' runs. As retrieval systems can return segments of arbitrary start and length, average precision calculation required adjustment to varying segmentation boundaries as explained in detail in [1]. Moreover, a new evaluation measure considered overlap relevance, binned relevance and tolerance to irrelevance for precision calculations, and was additionally applied after the campaign.

4.3 Search Results and Discussion

The results of our system named Idiap 2013 in the campaign for the search task are shown in Table 4. The four variants show the same ranking as on the development set. The LIMSI ASR transcript outperforms the LIUM one, which is not helped by metadata, a fact that is likely due to the low frequency of metadata words. Surprisingly, subtitles yield the lowest scores. Analyzing results per query, in 12 out of 50 test queries our best run gets the known item in the top 10 answers. These queries vary between runs, so they are not necessarily "simple" queries (assuming some queries out of 50 are easier to answer than others). One exception is query number 18 from the test set ("What does a ball look like when it hits the wall during Squash") which was answered correctly in all our runs, likely due to the number and specificity of words it contains – e.g. when compared to query number 49, "What foods are good to cook with?", which is more ambiguous and has a large array of potential answers from shows related cooking. At the opposite end of the scores, for 14 other queries, the known item is not found at all by our system among its top 1000 results.

In comparison to other systems' scores, our system ranked towards the lower third: the best MRR scores reached 0.40 and about half of the submitted runs were above 0.20. Our rather low scores (also on mGAP and MASP) could be due to the short average size of our segments, which were not calibrated to match the average size of known items. The most successful (and popular) strategy for this task was the fixed-length segmentation, though how this was calibrated by the participants must still be clarified.

4.4 Hyperlinking Results and Discussion

Three different sets of evaluation results for the Idiap 2013 system on the MediaEval 2013 hyperlinking sub-task are available:

I. Our initial run submission [3] to the hyperlinking subtask was scored using MAP after the deadline, separately from the other submissions, due to a time conversion problem in our results, undetected when validating the submission, but corrected afterwards.

II. The same runs as submitted in (I) were evaluated again with pooled top overlapping results across all the participants' submissions in the hyperlinking sub-task and scored with modified evaluation metrics like the overlapping, the tolerance based, and the binned MAPs.

III. We ran 18 additional experiments with different parameters to further investigate the results, and scored them with the overlapping, binned and tolerance based MAPs.

Tables 5 and 6 show the results of evaluation set (I) and (III), while the official results of evaluation set (II) will be discussed in the future in a synthesis paper by the task organizers. However, we summarize them as follows in terms of MAP scores:

- MAP with overlap relevance: our submissions ranked first with values of 0.52, 0.48 and 0.43; the next submission scored one run at 0.32, while all other runs of all other participants were below 0.3 (and two thirds were below 0.1).
- MAP with binned relevance: two systems reached 0.25, including ours, with our other two runs being 3rd and 4th. Three quarters of the systems were below 0.1.
- MAP with tolerance to irrelevance: our runs were ranked 2nd to 4th, with 0.12–0.14 scores, while the best run scored 0.16. Scores then decreased uniformly across runs up to 0.01.

Submission	P_5	P_10	P_20	MAP
L_V_M_O_T6V4_C	**0.620**	**0.583**	**0.413**	**1.00**
S_V_M_O_T8V2_C	0.400	0.443	0.370	0.832
S_V_M_O_T8V2_A	0.400	0.433	0.340	0.782

Table 6: Version (I) of our system is results for hyperlinking: precision at top 5, 10 and 20, and MAP.

In the official results of MediaEval S&H 2013, our three submissions to the hyperlinking task – two runs in the C scenario and one run in the A scenario – were thus always ranked among the first four top positions based on MAP values. Particularly, our two runs for scenario C (with context) had higher performance compared to our run for scenario A

LIMSI transcript with meta-data and visual features in the C scenario (I_V_M_O_C)

Metric	T0V10	T2V8	T4V6	T5V5	T6V4	T8V2	T10V0	Average
map	*0.2450*	0.4978	0.5468	0.5532	0.5209	**0.5602**	0.5521	0.4965
P_5	*0.3467*	0.6467	0.6733	**0.6933**	0.6667	0.6600	0.6467	0.6190
P_10	*0.3233*	0.5767	0.6067	0.6167	0.6333	0.6300	**0.6600**	0.5781
P_20	*0.2567*	0.4850	0.5267	**0.5550**	0.5400	0.5500	0.5267	0.4914

Subtitles with meta-data and visual features in the C scenario (S_V_M_O_C)

Metric	T0V10	T2V8	T4V6	T5V5	T6V4	T8V2	T10V0	Average
map	*0.2565*	0.4566	0.5021	0.5119	**0.5200**	0.4870	0.5151	0.4641
P_5	*0.3933*	0.5067	0.5400	0.5400	**0.5467**	0.4267	0.5133	0.4952
P_10	*0.2967*	0.5333	0.5300	0.5200	**0.5367**	0.5100	0.5333	0.4942
P_20	*0.2533*	0.5067	0.5200	0.5117	**0.5117**	0.4833	0.4983	0.4692

Subtitles with meta-data and visual features in the A scenario (S_V_M_O_A)

Metric	T0V10	T2V8	T4V6	T5V5	T6V4	T8V2	T10V0	Average
map	*0.1141*	0.3179	0.3914	0.4112	0.4272	0.4311	**0.4809**	0.3676
P_5	*0.0867*	0.3467	0.3800	0.4000	0.4067	0.4400	**0.4467**	0.3581
P_10	*0.0900*	0.3500	0.3867	0.4133	0.4433	0.4833	**0.5300**	0.3852
P_20	*0.0667*	0.3567	0.3950	0.4100	0.4217	0.4367	**0.4933**	0.3685

Table 5: Result set (III) of Idiap 2013 system for hyperlinking task: precision at top 5, 10 and 20, and MAPs (with consideration of overlap relevance, binned relevance and tolerance to irrelevance). The highest lowest values of precision and MAP are indicated respectively in bold and italics, for each weight W of textual similarity (from 0% to 100%) in the multimodal re-ranking procedure (noted T(W)V(1-W) without the trailing zeroes).

(without context). Similarly, some of the submissions from other participants showed a higher performance measured in terms of MAP over the top 20 positions for scenario C than for scenario A, which was expected given that scenario C offers more information than A.

The transcript type did not influence fundamentally the results, with the three sources being ranked very heterogeneously – the main factor being rather the underlying segmentation and hyperlinking methods. In any case, human-made subtitles did not appear to confer any specific advantage to any system. Rather, using the LIMSI ASR transcript always outperformed using subtitles, for our system and for other participants as well. This might be due to the query words having on average a higher word recognition rate for LIMSI ASR transcript than with subtitles, or the lower out-of-vocabulary words in LIMSI ASR transcript compared to subtitles which has an impact on indexing.

4.5 Analysis of Hyperlinking Results

Our system utilizes different textual and visual features. Each of the runs is coded with a combination of letters indicating the information used by the system, following the MediaEval S&H instructions, as follows: 'I' for LIMSI ASR based topic segments; 'S' for subtitle based topic segments; 'M' when using the metadata provided by the BBC (title, series, episode); 'O' when using other metadata such as cast and synopsis; 'V' when using visual concept detection scores for key-frames. In our system, we indicate by 'T(W)' the weight (in %) of the textual features for similarity, while 'V(1-W)' is the complementary weight of visual features. The scenarios for hyperlinking are coded 'C' when using contextual information (i.e. textual and visual) and 'A' when not using it. For example, I_V_M_O_T6V4_C indicates that the system utilized the LIMSI transcript for topic segmentation, along with metadata and other data, to derive the textual similarity scores of segments and, their visual concept similarity scores, which are re-ranked with W = 60% weight for textual similarity and 40% weight for visual similarity, in scenario C.

Based on these results, we will illustrate the utility of the proposed multimodal re-ranking method for the hyperlinking sub-task from the following perspectives:

- Effect of different evaluation methods.
- Importance of contextual information.
- Different weights for the textual vs. visual features.

The result set (I) in Table 6 have the highest MAP compared to other two result sets. This indicates that if user evaluations can be done for each of the individual runs with larger number of retrieved items, it may provide more insight than evaluating limited number of retrieved items from common runs across all the participants. On the other hand, such exclusive run evaluations are of course expensive; moreover, consideration of only the top few results across different runs can bring in more diversity to the results to be evaluated. However, the overall value of MAP is reduced by nearly 50% when removing the overlapping result segments, and is further reduced when the tolerance based threshold is used. Thus, there is still a large scope for future improvement with the proposed methodology. Still, the actual utility of such constraints on evaluation methods and metrics for hyperlinking task must still be confirmed.

When using similar features and run parameters, a higher MAP value was found when context (textual and visual features surrounding the known-item) was used for scenario C, indicating that this might actually add useful information, especially with our strategy of extending context boundaries to the closest segments. The majority of the participants have similarly observed that contextual information has helped achieve higher MAP. In our runs, we also observed that the size of the query anchor without contextual information for linking in scenario A was spanning across 8 to 11 visual shots and with contextual information in scenario C it spanned across 50 to 55 visual shots on average. Thus, the richer textual and visual information help achieve better MAP value for scenario C.

One of the most important observations from our result set (III) in Table 5 for combination of weights from 0% to 100% for the multimodal re-ranking is that, visual feature based segment similarity scores can be given weights between 0% and 50% in order to achieve a higher MAP. It does not seem useful to give more than 50% weight to the visual features, as all the highest MAPs are observed on T5V5 to T10V0. It is an interesting pattern that the subtitle-based runs achieve higher MAP results when higher weight is given to the textual features (e.g., T6V4 for S_V_M_O_C and T10V0 for S_V_M_O_A). The quality of text is likely more accurate with subtitles and thus giving higher weight to text is better than giving it to less accurate visual features. On the other hand, the LIMSI ASR transcript shows some instances with better results when textual and visual features are given equal weights (e.g., T5V5 achieved highest MAP of 0.6933 among all the results).

Our best system reaches MAP of 0.80 and 0.50 respectively on anchors 31 and 39, while the MRR of the corresponding search queries (item_23 for 31, item_25 for 39) is close to zero. This is an indication that the visual features may be helpful. Because, we did not utilized visual features for the search task and it might have caused such low MRR score for some of the queries compared to their higher MAP score during the hyperlinking task.

5. CONCLUSION AND PERSPECTIVES

In this paper, we proposed a novel search and hyperlinking system, based on our experience with content-based recommender systems for multimedia. We proposed a multimodal re-ranking technique, and used topic-based segmentation over transcripts or subtitles. Obtained results are leading us towards consideration of the contextual information and the visual features for the generation of hyperlinks from a multimodal perspective. Due to improved results for hyperlinking, our confidence in visual features for indexing has increased. This should motivate multimedia researchers to rely on visual features with a balanced weight given to textual features in certain scenarios.

An important issue that has appeared during the analysis of MediaEval 2013 results concerns the need for diversity in the result set, in addition to overall relevance to a query or an anchor segment. In other words, a set with more diverse segments will appear as more informative and will bring more novelty than a set with equally relevant but very similar results. “Diverse ranking” is a well studied topic in information retrieval [25], using either implicit or explicit techniques, which can be considered also in the present scenario. The implicit methods attempt to demote similar results to reduce overall redundancy [31], which is directly applicable here as well. The explicit methods attempt to model the possible interpretations of a query or an anchor to maximize the coverage of their aspects with selected documents [30], which would require in the present scenario more work on query analysis.

While accurate segmentation and hyperlinking are important for the back-end of a multimedia search and retrieval system, they should ultimately be used within a front-end component, i.e. a user interface allowing people to make use of these results. Within such a system, performance (or quality) is not only a matter of back-end accuracy, but results also from a suitable presentation of the results. Therefore, in future work, both aspects of system quality should be considered. For instance, we will explore the refinement and evaluation of a “navigation graph” [4] which displays the strongest links between segments of multimedia recordings of lectures, along with keywords and keyframes for each segment. Such a visualization appears to be a promising and cost-effective approach to navigating multimedia repositories, but the respective contributions to user satisfaction of hyperlinking accuracy and display/navigation strategies remain to be assessed.

6. ACKNOWLEDGMENTS

This work was supported by the Swiss National Science Foundation (AROLES project n. 51NF40-144627) and by the European Union (inEvent project FP7-ICT n. 287872). We would like to thank Maria Eskevich and Robin Aly for their valuable help with the task.

7. REFERENCES

[1] R. Aly, M. Eskevich, R. Ordelman, and G. J. F. Jones. Adapting binary information retrieval evaluation metrics for segment-based retrieval tasks. Technical report, ArXiv e-prints, 2013.

[2] R. Aly, R. Ordelman, M. Eskevich, G. J. F. Jones, and S. Chen. Linking inside a video collection – what and how to measure? In *Proceedings of the 22nd International Conference on World Wide Web Companion*, pages 457–460, 2013.

[3] C. Bhatt, N. Pappas, M. Habibi, and A. Popescu-Belis. Idiap at MediaEval 2013: Search and hyperlinking task. In M. A. Larson, X. Anguera, T. Reuter, G. J. F. Jones, B. Ionescu, M. Schedl, T. Piatrik, C. Hauff, and M. Soleymani, editors, *MediaEval*, volume 1043 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2013.

[4] C. Bhatt, A. Popescu-Belis, M. Habibi, S. Ingram, F. McInnes, S. Masneri, N. Pappas, and O. Schreer. Multi-factor segmentation for topic visualization and recommendation: the MUST-VIS system. In *Proceedings of ACM Multimedia 2013, Grand Challenge Solutions*, pages 37–42, Barcelona, 2013.

[5] S. Bird. NLTK: the Natural Language Toolkit. In *COLING/ACL Interactive Presentations*, Sydney, 2006.

[6] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *British Machine Vision Conference*, pages 76.1–76.12, 2011.

[7] M. Del Fabro and L. Böszörmenyi. State-of-the-art and future challenges in video scene detection: a survey. *Multimedia Systems*, pages 1–28, 2013.

[8] M. Eskevich, R. Aly, R. Ordelman, S. Chen, and G. J. F. Jones. The Search and Hyperlinking task at MediaEval 2013. In *MediaEval 2013*, Barcelona, 2013.

[9] M. Eskevich, G. J. F. Jones, R. Aly, R. J. Ordelman, S. Chen, D. Nadeem, C. Guinaudeau, G. Gravier, P. Sébillot, T. de Nies, P. Debevere, R. Van de Walle, P. Galuscakova, P. Pecina, and M. Larson. Multimedia information seeking through search and hyperlinking.

In *Proceedings of the 3rd ACM International Conference on Multimedia Retrieval*, pages 287–294, 2013.

[10] M. Eskevich, W. Magdy, and G. J. F. Jones. New metrics for meaningful evaluation of informally structured speech retrieval. In *Advances in Information Retrieval*, volume 7224 of *Lecture Notes in Computer Science*, pages 170–181. Springer Berlin Heidelberg, 2012.

[11] J.-L. Gauvain, L. Lamel, and G. Adda. The LIMSI Broadcast News Transcription System. *Speech Communication*, 37(1-2):89–108, 2002.

[12] A. G. Hauptmann, M. G. Christel, and R. Yan. Video retrieval based on semantic concepts. *Proceedings of the IEEE*, 96(4):602–622, 2008.

[13] M. A. Hearst. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64, 1997.

[14] M. Henziker, B.-W. Chang, B. Milch, and S. Brin. Query-free news search. *World Wide Web: Internet and Web Information Systems*, 8:101–126, 2005.

[15] M. Larson, R. Ordelman, F. Metze, W. Kraaij, and F. de Jong. *Proceedings of the 2010 International Workshop on Searching Spontaneous Conversational Speech (SSCS) at ACM Multimedia*. ACM, Florence, 2010.

[16] M. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, 2(1):1–19, 2006.

[17] P. Lops, M. de Gemmis, and G. Semeraro. Content-based recommender systems: State of the art and trends. In F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, editors, *Recommender Systems Handbook*, pages 73–105. Springer-Verlag, 2011.

[18] A. Massoudi, F. Lefebvre, C.-H. Demarty, L. Oisel, and B. Chupeau. A video fingerprint based on visual digest and local fingerprints. In *ICIP*, Atlanta, GA, 2006.

[19] T. Mei, X.-s. Hua, W. Lai, L. Yang, Z.-j. Zha, Y. Liu, Z. Gu, G.-j. Qi, M. Wang, J. Tang, X. Yuan, Z. Lu, and J. Liu. Msra-ustc-sjtu at TRECVid 2007: High-level feature extraction and search. In *TREC Video Retrieval Evaluation Online Proceedings*, 2007.

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

[21] N. Pappas and A. Popescu-Belis. Combining content with user preferences for TED lecture recommendation. In *11th Int. Workshop on Content Based Multimedia Indexing (CBMI)*, Veszprém, 2013.

[22] P. Pecina, P. Hoffmannova, G. J. Jones, Y. Zhang, and D. W. Oard. Overview of the CLEF-2007 cross-language speech retrieval track. In *Advances in Multilingual and Multimodal Information Retrieval*, volume 5152 of *Lecture Notes in Computer Science*, pages 674–686. Springer Berlin Heidelberg, 2008.

[23] J. Preston, J. Hare, S. Samangooei, J. Davies, N. Jain, D. Dupplaw, and P. H. Lewis. A unified, modular and multimodal approach to search and hyperlinking video. In *MediaEval 2013 / Search and Hyperlinking of Television Content*, October 2013.

[24] M. Riedl and C. Biemann. TopicTiling: a text segmentation algorithm based on LDA. In *Proceedings of the ACL 2012 Student Research Workshop*, pages 37–42. Association for Computational Linguistics, 2012.

[25] S. E. Robertson. The probability ranking principle in IR. In K. Sparck Jones and P. Willett, editors, *Readings in information retrieval*, pages 281–286. Morgan Kaufmann Publishers Inc., 1997.

[26] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management Journal*, 24(5):513–523, 1988.

[27] H. Schwenk, P. Lambert, L. Barrault, C. Servan, H. Afli, S. Abdul-Rauf, and K. Shah. LIUM's SMT machine translation systems for WMT 2011. In *6th Workshop on Statistical Machine Translation*, pages 464–469, Edinburgh, 2011.

[28] C. G. M. Snoek and M. Worring. Concept-based video retrieval. *Foundations and Trends in Information Retrieval*, 4(2):215–322, 2009.

[29] S. L. Tao Mei, Yong Rui and Q. Tian. Multimedia search reranking: A literature survey. In *ACM Computing Surveys*, 2014.

[30] S. Vargas, P. Castells, and D. Vallet. Explicit relevance models in intent-oriented information retrieval diversification. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 75–84. ACM, 2012.

[31] J. Wang and J. Zhu. Portfolio theory of information retrieval. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 115–122. ACM, 2009.