POLITECNICO DI TORINO
Repository ISTITUZIONALE

Effective video hyperlinking by means of enriched feature sets and monomodal query combinations

Original
Effective video hyperlinking by means of enriched feature sets and monomodal query combinations / Kavoosifar, MOHAMMAD REZA; Apiletti, Daniele; Baralis, ELENA MARIA; Garza, Paolo; Huet, Benoit. - In: INTERNATIONAL JOURNAL OF MULTIMEDIA INFORMATION RETRIEVAL. - ISSN 2192-6611. - ELETTRONICO. - (2019). [10.1007/s13735-019-00173-y]

Availability:
This version is available at: 11583/2736714 since: 2020-02-10T09:51:12Z

Publisher:
Springer

Published
DOI:10.1007/s13735-019-00173-y

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright
Springer postprint/Author's Accepted Manuscript

This version of the article has been accepted for publication, after peer review (when applicable) and is subject to Springer Nature's AM terms of use, but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: http://dx.doi.org/10.1007/s13735-019-00173-y

(Article begins on next page)

12 October 2024

Noname manuscript No.
(will be inserted by the editor)

Effective video hyperlinking by means of enriched feature sets and monomodal query combinations

Mohammad Reza Kavoosifar · Daniele Apiletti · Elena Baralis · Paolo Garza · Benoit Huet

Received: date / Accepted: date

Abstract Video content has been increasing at an unprecedented rate in recent years, bringing the need for improved tools providing efficient access to specific contents of interest. Within the management of video content, hyperlinking aims at determining related video segments from a collection with respect to an input video anchor.

This paper describes the system we designed to address feature selection for the video hyperlinking challenge, as defined by TRECVID, one of the top worldwide venues for multimedia benchmarking. The proposed solution is based on different combinations of textual and visual features, enriched to capture the various facets of the videos: automatically generated transcripts (ASR), visual concepts, video metadata, Named-Entity Recognition, and concept-mapping techniques.

The different combinations of monomodal queries are experimentally evaluated, and the impact of both parameters and single features are discussed to identify their contributions. The best-performing approach at the TRECVID 2017 video hyperlinking challenge was the Ensemble Feature Selection (EFS), which includes three different monomodal queries based on enriched feature sets.

Keywords Video Retrieval · TRECVID · Multimedia Indexing · Feature Selection

1 Introduction

The constant growth in both amount and variability of digital multimedia content being stored requires the development of techniques that not only identify files containing relevant content, but also bring the user as close as possible to the beginning of the relevant passage within this file to maximize the efficiency of information access.

Politecnico di Torino - Dipartimento di Automatica e Informatica - Torino, Italy
E-mail: {mohammadreza.kavoosifar, daniele.apiletti, elena.baralis, paolo.garza} @polito.it
EURECOM - Sophia Antipolis, France
E-mail: benoit.huet@eurecom.fr

Video hyperlinking deals with retrieval of video segments from a video collection. The retrieved segments should be topically related to given query video segments. The main objective of the task is to explore methods which enable users to easily browse the video collection using hyperlinks provided for the segments of their interests.

In this paper, we describe the framework used by the Eurecom-Polito team [20] to address the Hyperlinking task inside a video collection at TRECVID 2017 [2]. We have proposed a system that exploits different combinations of monomodal queries. Each query is based on textual features, enriched with concepts and entities aimed at maximizing the relevance of the selected video segments. The exploited features are: (i) automatic speech recognition transcripts [17, 22], (ii) visual concepts, (iii) entities extracted by Named-Entity Recognition techniques, and (iv) a concept mapping technique, which is based on WordNet [14].

The rest of the paper is organized as follows. Section 2 presents related works. Section 3 introduces the proposed system and its main phases. Section 4 provides details into the query formulation phase. Section 5 describes the different combinations of features exploited to retrieve relevant video segments. Section 6 presents and discusses the experimental results obtained by the proposed combinations on the TRECVID 2017 dataset. Finally, Section 7 draws conclusions and discusses future developments of the proposed work.

2 Related work

The automatic generation of hyperlinks within video collections has recently become a major subject, specifically in some evaluation benchmarks such as MediaEval and TRECVID [11, 12]. The key idea is to create hyperlinks between video segments within a collection, enriching a set of anchors that represent interesting entry points in the collection itself. Links can be seen as recommendations for potential viewers, whose intent is not known at the time of linking. The goal of the links is thus to help viewers gain insights on a potentially massive collection of videos so as to find information of interest, following a search and browse paradigm. To this aim, several techniques have been proposed.

Besides the unimodal approaches, such as [16], which relies on textual features only, multi-modal techniques taking into account different feature sets have emerged. In [38], Soleymani et al. have proposed a multi-modal system designed to analyze users behavior and interaction with browsed visual content for different image search intents, whereas the approach proposed in our paper exploits combinations of many different features, both textual and visual.

Additional paradigms propose models predominantly based on one specific modality (e.g., image search) and try to improve them using information from other modalities (e.g., captions) [28, 30]. Similarly, [25, 29] propose a text-to-video mapping. On the other hand, [18] described a system for content-based video retrieval from large surveillance video archives, using behavior, actions and appearance of objects. Recent high-performing approaches in video browsing revolve around retrieval of simplified sketches (e.g., by using simple color signatures [6]) and displaying the collection in a more informative way (e.g., using a graph-based keyframe arrangement for browsing [4]). A more in-depth sketch analysis where

deep semantic classifiers are employed for sketch auto-completion has been demonstrated also in an earlier work [41]. A vertical application of hyperlink techniques is presented in [5], where an effective signature-based approach has been proposed to link endoscopic images with video segments.

A new indexing and retrieval system is presented in [42]. It detects multiple object events or crowd events (e.g., group walking, group splitting, etc.). However, the generic video hyperlinking use case requires not only the detection of group items, but also single items or objects which are appearing inside the videos.

Some other approaches are also developed in Multimodal Video Retrieval. The IMOTION system [33] represents a multimodal content-based video search and browsing application offering a rich set of query modes based on a feature-fusion approach. The VERGE interactive search engine [27] is capable of browsing and searching into video content by providing integrated content-based analysis and retrieval modules, such as video shot segmentation, concept detection, clustering, and visual-similarity and object-based search. In terms of using features, the approach proposed in the current paper exploits a different set of features, for instance by including also video metadata, and by avoiding the need to perform video processing tasks since it relies on textual provided features.

Leveraging different information sources is a task investigated by [40]. They include video and text for efficient video browsing, however, the search and hyperlinking task [11] is to seek for meaningful videos with respect to a text query. Advances have been reported in the area of cross-modal systems by IRISA team [10] and VIREO teams [7]. Cross-modal systems are based on two (or more) modalities that are known to share a common set of categories.

The IRISA group exploited an enriched version of their 2016 algorithm, a crossmodal *Bidirectional Deep Neural Networks (BiDNN)* Joint Learning [46], which ranked first in TRECVID 2016. In their 2016 algorithm [45], training is performed cross-modally and in both directions: one modality is presented as an input and the other as the expected output, and vice-versa at the same time (i.e., the second one is presented as input and the first one as expected output). This is equivalent to using two separate deep neural networks and tying them (sharing specific weight variables) to make them symmetrical. Finally, for the phase of video hyperlinking, segments are compared. For each video segment, the two modalities are considered: embedded automatic transcripts with embedded CNN (a very deep Convolutional Neural Network [36]) representation) and a multimodal embedding is created with a bidirectional deep neural network. Then, the two multi-modal embeddings are compared with a cosine distance to obtain a similarity measure. However, for TRECVID 2017, the IRISA group, contrary to their 2016 algorithm, decided to put more emphasis on the choice of visual descriptors. Additionally, the use of metadata was explored in one of the runs.

The VIREO group introduced a deep model called *Semantic Representation Network (SRN)* which evaluates the relatedness between visual and text data. The structure of SRN contains different layers. At first, it consists of two networks, which share weights with each other, for inputs of anchors and targets. Then it encodes both target and anchor into the same feature space. After that, the holographic

layer would evaluate the relatedness between anchor and target by exploiting circular correlation [43], which measures vector correlation in the frequency domain using FFT (Fast Fourier Transform). Finally, the softmax layers output the probabilities of similarity and dissimilarity between anchors and targets. For the phase of video hyperlinking and for their 4 submitted runs at TRECVID 2017, they considered 2 algorithms. For Run-1 (Visual baseline), they exploited SRN and cosine similarity. Then for Run-3 (Multimodal baseline), they combined visual Run-1 and the text features extracted from ASR (Automatic Speech Recognition). For the other 2 runs (Run-2 and Run-4), they formulated the problem as an optimization algorithm (considering k-nearest neighbors) and adopted LID-first algorithm [9] for re-ranking of baseline results. The goal of this algorithm is to promote the ranks of targets with "lower data risk", specifically, in lower local dimensions, being hubs of data, and sufficiently diverse from neighboring regions.

Even if such proposals are all very promising, our approach gained higher MAiSP (Mean Average interpolated Segment Precision) [31] in the TRECVID 2017 workshop, thanks to the proposed combinations of multi-modal features.

Finally, in [44] additional studies on cross-modal systems are presented, however they work well only in terms of text-to-image retrieval, while our approach considers both image-to-text and text-to-image aspects.

3 System overview

Videos from digital archives and collections can be interconnected by their topic, events presented, activities depicted, people shown, and many other aspects. The video hyperlinking (LNK) task at TRECVID envisages a scenario where users are willing to find further information on some aspects of the video segment they are watching, hence they expect to be provided hyperlinks to related video content within a given archive or collection.

The video hyperlinking (LNK) task at TRECVID 2017 aims to foster progress in systems for effectively accessing video content. The task input is a query consisting of an anchor video segment. The task goal is to produce a ranked list of relevant segments with respect to the querying anchor.

To address such task, in this paper we propose and evaluate a system based on different combinations of both textual and visual features. All the data and metadata provided by the task organizers are exploited in our approach, with the only exception of visual concepts, for which we used those extracted by the Caffe framework with the BVLC GoogleNet model [39] trained to classify images into 1000 different ImageNet categories.

The proposed approach also considers extra features to identify the most relevant terms and concepts in each query. To this aim, we exploit the Stanford Named Entity Recognizer (NER) [15] to find entities, and a concept mapping technique based on WordNet [14].

Overall, the proposed system is based on the following features:

- Automatic speech recognition transcripts (LIMSI) [17, 22].
- Visual concepts, provided by the ImageNet GoogleNet model.

Fig. 1: System stages

- Metadata of the videos (specifically, title, description, and tags).
- Results of named-entities recognition and concept mapping.

The system exploits a three-step approach, with each step associated to a computation stage, as presented in Figure 1:

1. Data segmentation (Section 3.1).
2. Indexing (Section 3.2).
3. Query formulation and retrieval (Section 3.3).

3.1 Data segmentation

Since the hyperlinking task result consists of video segments, the first step splits full-length videos into short segments. To this aim, a fixed segmentation of 120 seconds has been used. Fixed segmentation has been preferred over shot segmentation since our experimental experience from TRECVID 2016 [19] showed that user-generated amateur videos are more suitable to be processed with the former approach. The 120-second period is the result providing better coverage and more choice than lower-length segmentations. Furthermore, the 120-second length is the upper bound for an anchor in the hyperlinking task (the minimum length is 10 seconds).

3.2 Indexing

Apache Solr[1] [13] has been used to index the textual and visual features associated with each segment. Multiple indexes have been created for the video segments, each based on one of the following features: (i) the LIMSI transcripts of the segments, (ii) the visual concepts of the segments, and (iii) the metadata of the full videos.

The specific indexing structure implemented by Solr is known as inverted index. An inverted index stores, for each term, the list of documents where the term is present. This makes term-based queries very efficient [21], and it is exploited by the proposed approach.

All the textual data associated with the segments have been preprocessed to remove irrelevant words and punctuation. Specifically, we used 665 different English stop words[2] [8], narrowing down the word list of each segment to its core concepts.

The transcripts exploited by our approach are provided by the LIMSI tool, as in our experiments on the training anchors, on average the LIMSI [22] transcripts allow to achieve better results than the LIUM [17] ones.

[1] http://lucene.apache.org/solr
[2] https://www.ranks.nl/stopwords

3.3 Query formulation and segment retrieval

This stage aims at generating an optimal query text to be used for the segment retrieval on Solr indexes. Ideally, the best query text would completely describe the video anchor in terms of contents and context, besides providing the user intentions and preferences, to allow specific and personalized results.

The proposed approach is designed to build an enriched query text from the available features. To this aim, the video query segment (anchor) is converted into a textual query string by including all the textual information associated with the anchor (i.e., LIMSI transcripts and visual concepts), the metadata of the full video containing the anchor (i.e., title and tags), and additional text obtained by Named Entity Recognition (NER) and a concept mapping technique.

Finally, the resulting enriched query is used to query the Apache Solr indexes, hence identifying related segments, ranked by relevance.

The core of this paper addresses the evaluation of different strategies for creating queries based on the content of the video anchors. Such specific tasks are described in Sections 4 and 5.

4 Query formulation

The key idea of the proposed system consists of different combinations of mono-modal queries into an overall multi-modal approach. The specific mono-modal queries are described in the current Section, whereas their combinations are presented in Section 5.

The characteristics of the four monomodal queries are the followings:

1. **LIMSI-based query + Named-Entity Recognition.**
   For each anchor, a textual query is built by considering the words appearing in the LIMSI transcript of the anchor. Then, Named-Entity Recognition (NER) is applied on the anchor LIMSI transcripts to extract relevant names of entities and give them higher relevance in the query. NER labels sequences of words in a text which are related to the names of entities, for instance people and company names, or gene and protein names. The basic idea is that the segments containing the same entities as the anchor are potentially more relevant, hence a higher weight is assigned to those words in the query, as well as groups of 2, 3, and 4 adjacent words, e.g., "United States of America".
   The resulting query is executed on the LIMSI transcript index.
   For example, if the LIMSI text is: *"Handmade portraits: Staceyrebecca"*, the query would be: *"Handmade portraits" (W1.0) OR "Staceyrebecca" (W1.6)* since "Staceyrebecca" is a know entity and it is assigned a higher weight (1.6 instead of 1 in our case, the parameter value of query boost weight is discussed in Section 6.2).
2. **Visual-concept-based query + concept mapping technique.**
   For each video anchor, a textual query is built by considering the "names" of visual concepts appearing in the anchor. The visual concepts with a score greater than 0.3, as provided by the GoogleNet model, are selected (the parameter value of visual concept filter is discussed in Section 6.2).
   Furthermore, a concept mapping technique based on WordNet is applied to find the most relevant concepts inside the query. The mapping is performed by using

the words appearing in the full video metadata and the list of visual concepts of the segment. To maximize the word-list enrichment for concepts and metadata, we applied WordNet using both the synonyms and hypernyms of the words. Furthermore, also groups of 2, 3 and 4 adjacent words are considered.
The concept mapping technique aims at increasing the relevance of the visual concepts of the considered anchor that are related to the content of the whole video. For this reason, each visual concept name of the anchor is compared with the words appearing in the metadata of the video containing the anchor. If the visual concept, or its synonym (or hypernym) based on WordNet, appears in the metadata of the video, then the visual concept is assigned a higher weight in the query. The resulting query is executed on the visual concept index.
For example, if metadata text is: *"Top 100 golf tips for kids"*, and visual concepts are: *"digital clock, golf ball"*, the resulting query would be: *"digital clock" (W1.0) OR "golf ball" (W1.6)*, since *"golf"* is matching.

3. Metadata-based query for segment selection.
Metadata can be used to select either segments or videos. Metadata are associated to the full video, i.e., all segments of a video share the same metadata. A textual query built from a segment (anchor) metadata will be the same for all segments of the same video.
If the query is executed on a metadata index, only full videos can be selected, with all their corresponding segments. Instead, to select specific segments, metadata queries are executed on the LIMSI transcript index, since transcripts are specific for each segment. Named-Entity Recognition (NER) is applied to extract relevant entities and give them higher relevance in the query, by following the same procedure described for queries #1 and #2.
For example, if metadata is: *"United Kingdom weekly Talk Show"*, the query on LIMSI transcripts would be: *"United Kingdom" (W1.6) OR "weekly" (W1) OR "Talk Show" (W1.6).*

4. Metadata-based query for video selection.
This query is the same as the previous one, but it is executed on the metadata index, hence returning videos and not segments. For this reason, the results of such query cannot be used directly to propose the resulting segments, since all the segments of the related videos would be selected. However, this query helps in filtering a pre-selection of videos among which related segments are highly likely to be found (see Section 5.2).

5 Query combinations

To address the video hyperlinking task, we propose and evaluate four strategies based on different combinations of queries. Building on past experience [19], we decided to combine multiple mono-modal queries (described in Section 4) into a globally multi-modal system, resulting into the following combinations:

1. Ensemble Feature Selection (EFS) (Section 5.1)
2. Metadata-based approach (Section 5.2)
3. Pipeline approach (Section 5.3)

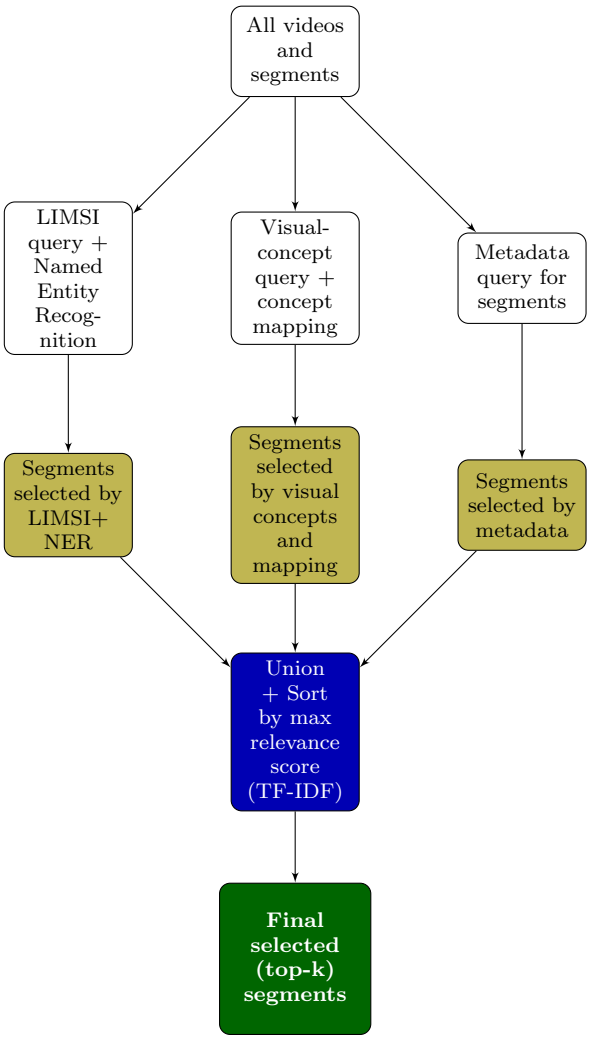

Fig. 2: Ensemble Feature Selection (EFS)

4. LIMSI-NER approach (Section 5.4)

LIMSI-NER is not an actual combination, but a simple mono-modal approach. However, it is considered since it is a core part embedded in the other proposed combinations: its separate evaluation is a noteworthy addition for the experimental comparison to identify its specific contribution to the overall results.

For each of the four combinations, an experimental run has been submitted to TRECVID, and its results are presented in Section 6.

5.1 Ensemble Feature Selection (EFS)

The EFS combination exploits the three monomodal queries in parallel. It aims at dynamically selecting the overall best segments among all the segments returned by each monomodal query by means of a two-step approach, as depicted in Figure 2.

In the first step, each monomodal query is executed separately, returning its own set of resulting segments. The queries exploited in EFS are: (i) the LIMSI-based segment-transcript query + Named-Entity Recognition, (ii) the visual-concept-based segment query + concept mapping technique, and (iii) the metadata-based query for segment selection. In the second step, the three resulting subsets of segments are merged and ranked in terms of relevance score. We used TF-IDF provided by Solr [37] as a metric for the relevance score. TF-IDF (Term Frequency - Inverse Document Frequency), is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus [23]. If the same segment is returned more than once from different queries, the system keeps only the copy of the segment with the highest relevance score value. The output of the second step is the final result of the EFS approach, where a segment is considered relevant depending on its maximum relevance score among the three queries.

5.2 Metadata-based approach

The approach based on metadata performs a first step exploiting the video metadata index to pre-select correlated videos (see Figure 3). This filter restricts the subsequent segment selection to such videos only. The idea is to discard possible matches with segments having some relevant feature match (e.g., transcript) but belonging to uncorrelated videos, under the hypothesis that those are probably false positive matches. In the second step, relevant segments are retrieved in parallel by both LIMSI and visual-concept queries. Finally, a union and sort is applied similarly to the EFS approach.

5.3 Pipeline approach

The pipeline approach is focused on the exploitation of a selected subset of data which are deemed to be the most informative among the available features, specifically machine-generated data and segment-level data. To this aim, only two queries are included: LIMSI and visual concepts. Since the purpose is to maximize the selection of highly relevant segments, only those being relevant for both queries are considered. Hence, a segment is considered relevant if and only if it is (i) selected by both queries separately, and also (ii) independently ranked high in each single query result.

The pipeline approach implements two separate flows, each consisting of two steps (see Figure 4). One flow first selects the top-k segments according to LIMSI and, among them, it then executes the visual-concept query. On the other flow, the same steps are performed switching the order of the two queries: first the top-k visual concepts, then the LIMSI. Finally, the last step works on the two subsets of segments returned by each flow: they are merged and ranked in terms of relevance score. Please note that the sequence *LIMSI + visual concepts* is not

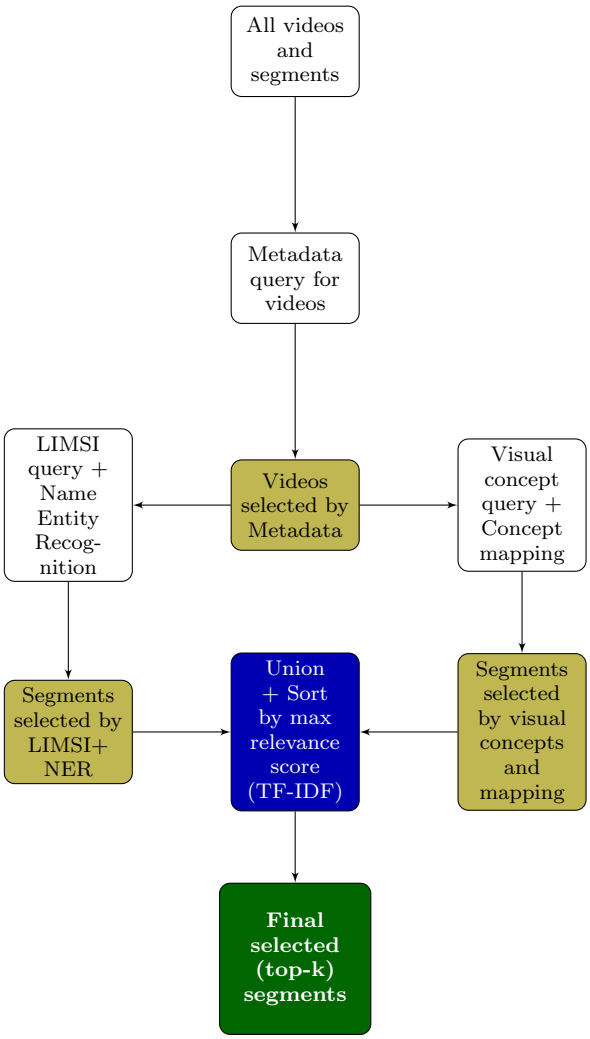

Fig. 3: Metadata based approach

equal to *visual concepts + LIMSI* because only the top-k segments are selected after the first steps.

5.4 LIMSI-NER approach

This approach exploits only the LIMSI transcript query with the Named-Entity Recognition (NER) technique (Figure 5). The aim of this approach is to compare the experimental results of the multi-modal combinations with a straightforward mono-modal approach. The LIMSI-based query has been selected since it reported the best results on the training anchors, being better than both the visual concepts alone and the LIUM-based transcripts.

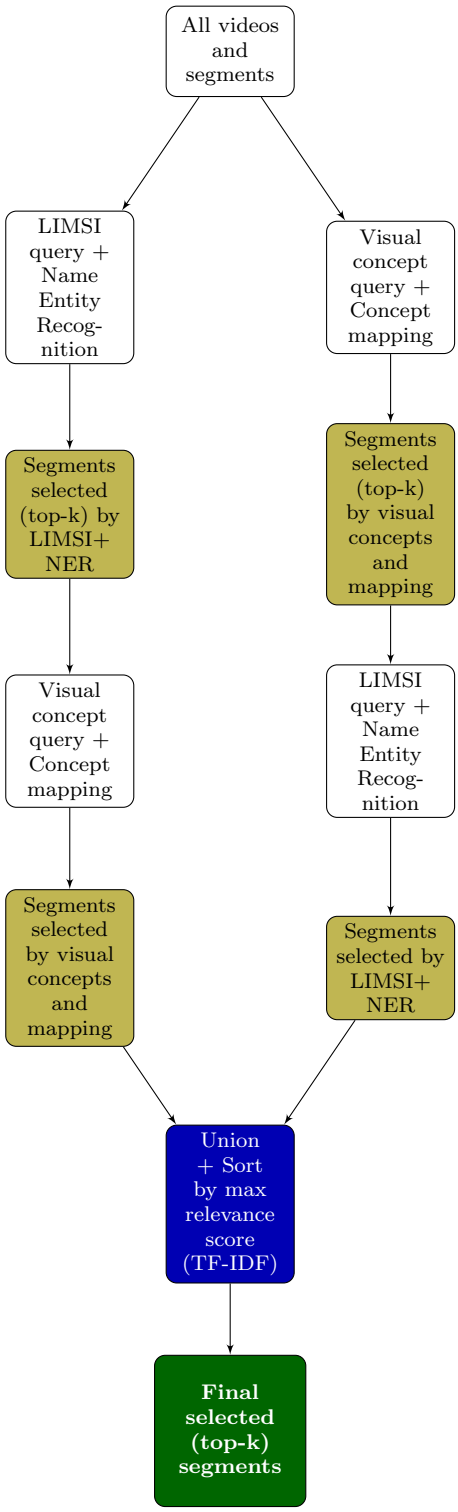

Fig. 4: Pipeline approach

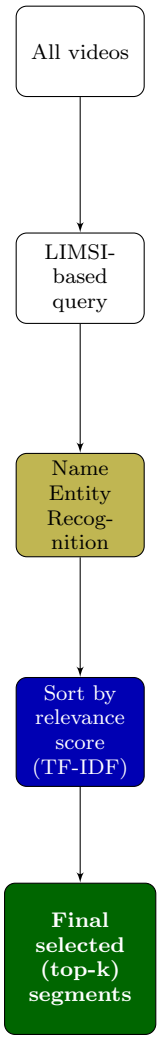

Fig. 5: LIMSI-NER approach

6 Experimental evaluation

The video dataset used for the TRECVID 2017 competition has been provided by blip.tv, and it is identified by the name "Blip10000" [35]. It consists of 14,838 videos, for a total of 3,288 hours. The mean length of videos is around 13 minutes.

Videos are characterized by metadata (we considered title, short program descriptions, and tags), Automatic Speech Recognition (ASR), transcripts (LIUM and LIMSI), visual concepts, shots, and keyframes.

The videos present a variety of topics from computer science tutorials and sightseeing guides to homemade song covers. They are provided in many languages but a vast majority of them are in English, while the anchor video fragments were exclusively in English.

Metric	EFS	Pipeline	LIMSI-NER	Metadata
P@5	**0.840**	0.808	0.725	0.704
P@10	**0.808**	0.748	0.667	0.556
MAP	**0.164**	0.114	0.093	0.082
MAiSP	**0.253**	0.185	0.155	0.132

Table 1: Results of the different approaches submitted to TRECVID according to each evaluation metric.

The training set provided by TRECVID contains 90 query anchors and their corresponding set of ground-truth related segments. The test set consists of 25 different query anchors.

The four proposed approaches have been submitted to the TRECVID 2017 video hyperlinking benchmark and their results are presented in Section 6.1, whereas the impact of the set of parameter values are discussed in Section 6.2.

6.1 Experimental results

Results have been evaluated according to the following metrics:

- Precision at rank 5 (P@5), i.e., the number of true positives in the top 5 selected segments.
- Precision at rank 10 (P@10).
- Mean Average Precision (MAP), which considers true positives all segments overlapping with a segment that was considered relevant in the ground truth [1].
- An adapted MAP called Mean Average interpolated Segment Precision (MAiSP) [31]

Table 1 reports the results provided by TRECVID 2017 for each of our approaches.

EFS (Ensemble Feature Selection) consistently yields the best results according to all metrics. We recall that it exploits all the available features (LIMSI transcripts, visual concepts, and metadata) by executing three mono-modal queries, one for each feature, and then ranking all the resulting segments by descending relevance score. This approach allows each feature/query to contribute at its best to the overall results, hence reaching the highest score in the TRECVID competition among all participants in terms of MAiSP (see Figure 7).

The rest of the approaches are ranked uniformly by all metrics: the pipeline approach is the second best, then LIMSI-NER, and finally the metadata-based one.

If we consider the LIMSI-only approach as a baseline, adding the visual concepts is a relevant addition: indeed, pipelining the two queries based on LIMSI and visual concepts yields a significant improvement over the LIMSI alone, in particular for P@5 and P@10. To further improve the results, besides the change in structure, EFS adds the metadata-based feature. Even if the metadata-based approach alone is the worst performing technique among the four, when exploited in the EFS, it contributes to a better resulting segment selection.

This behavior leads us to investigate the contribution of each query to the overall results of EFS. Such analysis is provided in Figure 6: for the 25 test anchors, the number of segments selected by each query are reported, hence providing the

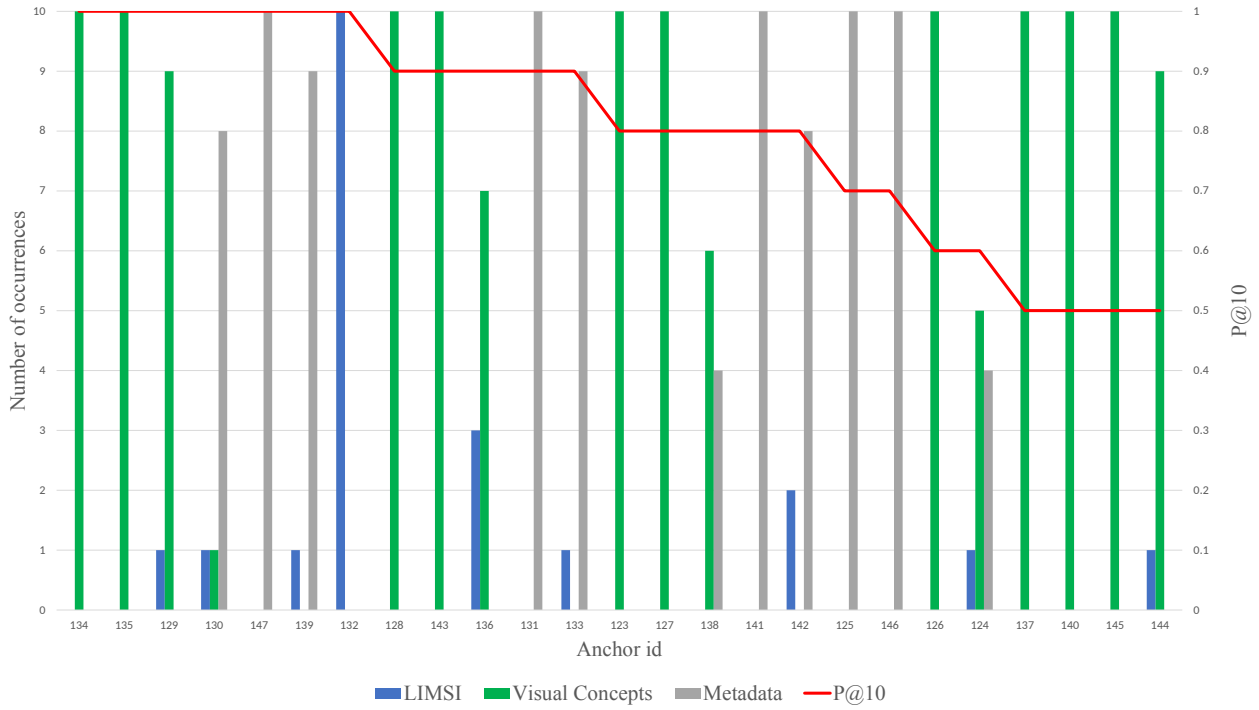

Fig. 6: Composition of EFS results: for each test anchor, the number of relevant segments provided by each query (LIMSI, visual concepts, and video metadata) within the top 10 resulting segments are reported, together with the total number of actually relevant segments (P@10).

composition of the results (top 10 segments) as contributions of each monomodal feature query. To visually assess if the selected segments actually contribute to the overall result, the red line indicates the overall P@10 score, which is also used to sort horizontally the 25 anchors in decreasing order of precision. On the left part of Figure 6, anchors leading to the highest P@10 score are reported: for the first 7 anchors, all the top 10 selected segments are relevant. Such segments are provided by completely different mixes of queries: visual concepts alone (90-100%) contribute for the first 3 anchors, metadata alone (80-100%) contribute for other 3 anchors, and LIMSI transcripts alone contribute to all 10 segments for the 7th anchor. The same pattern occurs in the remaining 18 anchors with lower P@10 score: 9 anchors lead to segments selected exclusively by visual concepts, 6 anchors lead to segments selected exclusively by metadata, and 3 anchors lead to segments selected by different queries.

Results suggest that depending on the anchor, completely different queries can lead to the correct relevant segment selection, hence it is crucial to consider ensembles of feature, i.e. combinations of monomodal queries. Actually, multi-feature approaches (EFS and pipeline) lead to better results according to all metrics (Table 1). In particular, the advantage of the EFS approach is that it lets the best

EFS queries	P@10	Number of segments
Metadata	0.891	92
LIMSI	0.762	21
Visual concepts	0.752	137

Table 2: Average P@10 of each query contributing to the EFS, over all the test anchors.

performing query among the three features (visual concepts, metadata, LIMSI) drive the results.

Overall, visual concept seems to be the most useful feature, followed by video metadata. The visual concept contribution is coherent with the improvement shown by the pipeline over the LIMSI (Table 1), as the pipeline adds visual concept queries.

Video metadata contribution is more contrasting. The metadata-based approach achieved the lowest result according to all metrics (Table 1). This outcome was unexpected as (i) metadata are among the most useful features in the P@10 analysis of EFS (Figure 6), and (ii) performance on the training anchors was higher than pipeline and LIMSI approaches. However, a fundamental difference is present between the EFS inclusion of the video metadata and the metadata-based approach: the latter exploits metadata to pre-filter videos. Hence, video-metadata exploitation as one of the possible segment queries, like in EFS, yields good results in a significant portion of the test set (Figure 6), whereas the video-wide pre-filtering of the metadata-based approach leads to poor final performance.

A deeper analysis of EFS results actually shows that video metadata are the best performing queries in terms of average P@10 contribution to the EFS results over all test anchors, as reported in Table 2. Among the segments selected by the video metadata query, 89.1% are relevant, whereas only 75.2% of those selected by visual concepts are relevant.

Regarding the TRECVID 2017 video hyperlinking task, MAiSP [31] is the most important performance measure, as it considers the start and end of segments, and evaluates the whole segment prediction. Figure 7 reports the results, in terms of MAiSP, of the different competing approaches. Three teams with 12 different approaches participated in the competition. EFS (Ensemble Feature Selection) ranked first with a MAiSP higher than 0.25. A significant difference has been reported with respect to the second-best approach, the pipeline one, who reached a MAiSP score of less than 0.20. Three of the approaches proposed in this paper ranked in the first three positions: EFS, pipeline, and LIMSI-NER, respectively, whereas the metadata approach ranked 5th. Differences in MAiSP are higher between the first and the second best, then from third ranked (with a MAiSP of 0.15) differences are lower: the interval 0.15-0.10 of MAiSP includes most of competing approaches, with the exception of the best two and the last.

In terms of precision at rank 5 and 10 (P@5 and P@10), EFS and pipeline approaches reached values higher than 0.8, together with other approaches proposed by the VIREO team.

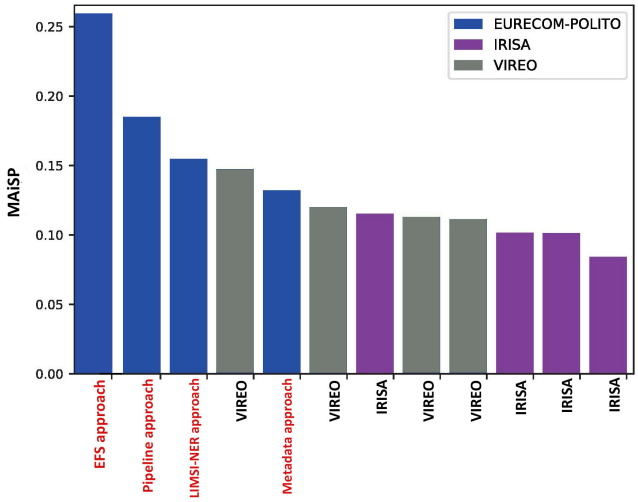

Fig. 7: Results of all the approaches submitted to TRECVID 2017 in terms of MAiSP

6.2 Analysis of parameters and training set

This section presents the training work on the ground truth and addresses the impact of the different parameter values on the performance of the proposed approaches.

The ground truth provided by TRECVID organizers consists of the top-10 related segments for each of the 90 training anchors. However, approximately 60% of the training segments are not annotated with ground truth, so when one of these segments is selected, we cannot state whether such result has to be accepted or rejected.

Figure 8 presents the results of each approach over all the 90 training anchors, indicating how many of the resulting segments are accepted, rejected or not yet evaluated, based on the ground truth. Accepted segments represent the P@10 on the training set. The full set of metrics on the training set, for each approach, are reported in Table 3.

Results on the training set (Table 3) and on the test set (Table 1) lead to the same top-performing algorithm, i.e., EFS. However they rank the other approaches differently: metadata, pipeline, and LIMSI in the training set, and pipeline, LIMSI and metadata in the test set. The most noteworthy difference is the metadata approach, which is the second-best on the training set and becomes the worst on the test set. We consider that such data-dependent changes in results are due to the small number of samples in both datasets, so few specific samples can influence the overall ranking of the approaches.

The default parameter-value configuration considered for all approaches is as follows.

- K-filter: **1000**
- Stemming algorithm: **SnowballPorter**
- Filter threshold of visual concepts: **0.3**

Measures	**EFS**	Metadata	Pipeline	LIMSI-NER
P@10	**0.289**	0.227	0.221	0.212
MAP	**0.096**	0.077	0.071	0.065
MAiSP	**0.084**	0.062	0.059	0.054

Table 3: Pre-evaluation results based on ground-truth

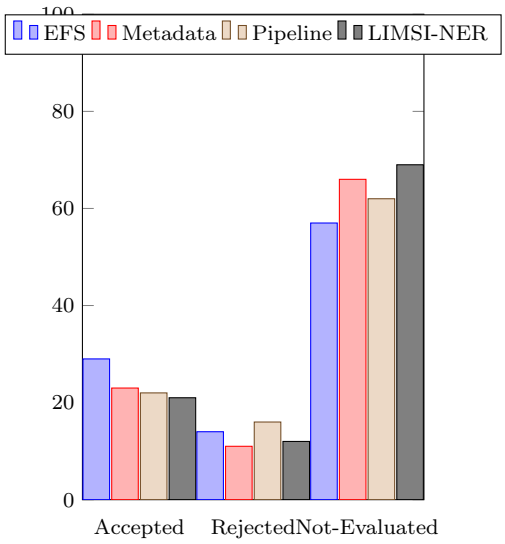

Fig. 8: % of pre-evaluation segment tags based on ground-truth for training anchors

Algorithm	EFS	Metadata	Pipeline	LIMSI-NER
SnowballPorter	**0.289**	**0.227**	**0.221**	**0.212**
PorterStem	0.278	0.215	0.205	0.198
Hunspell	0.224	0.187	0.178	0.153
KStem	0.219	0.181	0.173	0.145

Table 4: P@10 results for stemming algorithms in Solr

- Query boost weight: **1.6**
- NER classifier: **Multi Classifier**
- WordNet similarity algorithm: **Lin**
- Lin algorithm threshold: **0.7**

K-filter indicates the top-k number of segments to be kept in the final step of each approach: it is fixed to 1000 because in TRECVID each participant/approach was allowed to submit up to 1000 segments for each run.

The analysis of the other parameters is described in following.

1. Stemming algorithms in Solr

To be able to search the text efficiently and effectively, Solr splits the text into tokens during both indexing and query execution. Those tokens can also be pre- and post-filtered for additional flexibility. This allows for case-insensitive

Threshold	EFS	Metadata	Pipeline
0.2	0.256	0.219	0.213
0.3	**0.289**	**0.227**	**0.221**
0.5	0.243	0.211	0.207
0.7	0.231	0.204	0.198

Table 5: P@10 results for filter threshold of visual concepts

Boost value	EFS	Metadata	Pipeline	LIMSI-NER
1.2	0.268	0.211	0.202	0.197
1.3	0.268	0.211	0.202	0.197
1.4	0.273	0.215	0.208	0.202
1.5	0.281	0.221	0.215	0.208
1.6	**0.289**	**0.227**	**0.221**	**0.212**
1.7	0.283	0.223	0.217	0.209
1.8	0.280	0.219	0.214	0.206

Table 6: P@10 results for query boost value

Classifier	EFS	Metadata	Pipeline	LIMSI-NER
No Classifier	0.197	0.164	0.152	0.136
Single Classifier	0.271	0.210	0.207	0.193
Multi Classifier	**0.289**	**0.227**	**0.221**	**0.212**

Table 7: P@10 results for NER classifiers

Algorithm	EFS	Metadata	Pipeline	LIMSI-NER
LESK	0.279	0.217	0.209	0.198
Lin	**0.289**	**0.227**	**0.221**	**0.212**
Wu-Palmer	0.281	0.220	0.208	0.202

Table 8: P@10 results for WordNet similarity algorithms

search, misspelled product names, synonyms, etc. [37]. For our approaches, we analyzed four stemming token filters:

1. PorterStem transforms the token stream by applying the Porter stemming algorithm.
2. SnowballPorter stems words using a Snowball-generated stemmer.
3. Hunspell is a TokenFilterFactory that creates instances of HunspellStemFilter.
4. KStem is a high-performance kstem filter for English.

Table 4 reports the experimental results on the training set for the 4 proposed approaches. The SnowballPorter is always the best stemmer in terms of precision at rank 10.

2. Threshold of visual concept recognition

To maximize the effectiveness of visual concepts, a properly thresholding of the visual-concept recognition score is required. Based on the analysis reported in Table 5, high threshold values such as 0.5 and 0.7 remove useful concepts for some anchors, thus reducing the final precision. On the contrary, low filter

Threshold	EFS	Metadata	Pipeline	LIMSI-NER
0.6	0.281	0.218	0.214	0.203
0.7	**0.289**	**0.227**	**0.221**	**0.212**
0.8	0.275	0.215	0.210	0.198

Table 9: P@10 results for Lin algorithm threshold

threshold values such as 0.2 let irrelevant concepts to be included. The best result is obtained with a filter threshold value of 0.3 for all the three approaches exploiting visual concepts (LIMSI-NER approach does not exploit visual concepts).

3. Query boost weight

The query boost parameter is used to determine the higher weight of query words selected by the concept-mapping technique and the named-entity recognition, as allowed by the Solr query engine [26]. All weight values from 1.2 to 1.8 with a 0.1 step have been analyzed and P@10 results are reported in Table 6). The query boost weight 1.6 yields the best results for all approaches.

4. NER classifier

Stanford NER is also known as CRFClassifier. It provides a general implementation of (arbitrary order) linear chain Conditional Random Field (CRF) sequence models. There are two kinds of CRFs provided by Stanford Named Entity Recognizer: Single CRF NER Classifier and Multiple CRFs NER Classifier. These two classifiers have been compared, besides the No Classifier option; results are reported in Table 7. The Multiple CRFs NER Classifier obtained the highest P@10 for all approaches.

5. WordNet similarity for concept mapping

WordNet has been exploited to determine a quantitative similarity to related words. We considered four different algorithms for this analysis:

1. The Wu-Palmer (Wu & Palmer) [47] calculates relatedness by considering the depths of the two synsets in the WordNet taxonomies, along with the depth of the LCS (Least Common Subsumer).
2. Resnik [32] similarity score denotes how similar two word senses are, based on the Information Content (IC) of the Least Common Subsumer (most specific ancestor node).
3. Lin [24] adapts Resnik's method and defines the similarity of two concepts as the ratio between the amount of information needed to state the commonality between them and the information needed to fully describe them.
4. LESK [3] metric measures the overlap between the glosses of the two concepts and also concepts directly related via relations such as hypernyms and meronyms.

Based on the results reported in Table 8), the Lin algorithm yields the best performance in terms of precision at rank 10 for all approaches.

The Lin similarity algorithm requires an inner parameter: a threshold used to filter the selected mapped concepts. Although a previous analysis was already performed on this threshold value [34], we report in Table 9 the results on our

specific dataset for a short range of values (0.6, 0.7, 0.8). The current results confirm the previous study, indicating that a 0.7 threshold yields the highest precision at rank 10 for all approaches.

7 Conclusions and future work

The paper addressed the video hyperlinking problem by proposing enriched query formulations and their combinations. Features considered in the proposed approaches are textual and include ASR transcripts, visual concepts and video metadata, enriched with Named-Entity Recognition and a concept-mapping techniques. Experiments addressed the parameter impact of the different components involved in the query enrichment process and results from the TRECVID submission of the proposed approaches. In particular, the Ensemble Feature Selection (EFS) approach reached higher performance than all other competitors at TRECVID for the specific video hyperlinking task.

A discussion on the contributions of the different components has been provided, showing how the same features (e.g., video metadata), used at different stages of the selection process, can lead to contradictory results. Detailed analysis of such contributions reported that each monomodal query is specifically useful for a subset of the test anchors. Hence, approaches (i) considering ensembles of different monomodal queries and (ii) able to let the best specific query emerge for each test anchor, yielded the best overall results consistently across different metrics.

Future work will address the inclusion of additional features, such as OCR (Optical Character Recognition) results, and new combinations of the different modalities to better capture their specific contributions.

References

[1] Aly R, Eskevich M, Ordelman R, Jones GJ (2013) Adapting binary information retrieval evaluation metrics for segment-based retrieval tasks. arXiv preprint arXiv:13121913

[2] Awad G, Butt A, Fiscus J, Joy D, Delgado A, Michel M, Smeaton AF, Graham Y, Kraaij W, Quénot G, Eskevich M, Ordelman R, Jones GJF, Huet B (2017) TRECVID 2017: Evaluating ad-hoc and instance video search, events detection, video captioning and hyperlinking. In: Proceedings of TRECVID 2017, NIST, USA

[3] Banerjee S, Pedersen T (2002) An adapted lesk algorithm for word sense disambiguation using wordnet. In: International Conference on Intelligent Text Processing and Computational Linguistics, Springer, pp 136–145

[4] Barthel KU, Hezel N, Mackowiak R (2016) Navigating a graph of scenes for exploring large video collections. In: International Conference on Multimedia Modeling, Springer, pp 418–423

[5] Beecks C, Schoeffmann K, Lux M, Uysal MS, Seidl T (2015) Endoscopic video retrieval: A signature-based approach for linking endoscopic images with video segments. In: Multimedia (ISM), 2015 IEEE International Symposium on, IEEE, pp 33–38

[6] Blažek A, Lokoč J, Matzner F, Skopal T (2015) Enhanced signature-based video browser. In: International Conference on Multimedia Modeling, Springer, pp 243–248
[7] Bois R, Vukotić V, Simon AR, Sicre R, Raymond C, Sébillot P, Gravier G (2017) Exploiting multimodality in video hyperlinking to improve target diversity. In: International Conference on Multimedia Modeling, Springer, pp 185–197
[8] Bradford RB, Pozniak J (2016) A systematic approach to design of a text categorizer. In: Systems, Man, and Cybernetics (SMC), 2016 IEEE International Conference on, IEEE, pp 509–514
[9] Cheng ZQ, Zhang H, Wu X, Ngo CW (2017) On the selection of anchors and targets for video hyperlinking. In: Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, ACM, pp 287–293
[10] Demirdelen M, Budnik M, Sargent G, Bois R, Gravier G (2017) IRISA at TRECVID 2017: Beyond crossmodal and multimodal models for video hyperlinking. In: Working Notes of the TRECVID 2017 Workshop
[11] Eskevich M, Jones GJ, Aly R, Ordelman RJ, Chen S, Nadeem D, Guinaudeau C, Gravier G, Sébillot P, De Nies T, et al (2013) Multimedia information seeking through search and hyperlinking. In: Proceedings of the 3rd ACM conference on International conference on multimedia retrieval, ACM, pp 287–294
[12] Eskevich M, Aly R, Racca D, Ordelman R, Chen S, Jones GJ (2014) The search and hyperlinking task at mediaeval 2014
[13] Eskevich M, Bui QM, Le HA, Huet B (2015) Exploring video hyperlinking in broadcast media. In: Proceedings of the Third Edition Workshop on Speech, Language & Audio in Multimedia, ACM, pp 35–38
[14] Fellbaum C (1998) WordNet: An Electronic Lexical Database. Bradford Books
[15] Finkel JR, Grenager T, Manning C (2005) Incorporating non-local information into information extraction systems by gibbs sampling. In: Proceedings of the 43rd annual meeting on association for computational linguistics, Association for Computational Linguistics, pp 363–370
[16] Galuščáková P, Saleh S, Pecina P (2016) Shamus: Ufal search and hyperlinking multimedia system. In: European Conference on Information Retrieval, Springer, pp 853–856
[17] Gauvain JL (2010) The quaero program: Multilingual and multimedia technologies. In: International Workshop on Spoken Language Translation (IWSLT)
[18] Hoogs A, Perera AA, Collins R, Basharat A, Fieldhouse K, Atkins C, Sherrill L, Boeckel B, Blue R, Woehlke M, et al (2015) An end-to-end system for content-based video retrieval using behavior, actions, and appearance with interactive query refinement. In: Advanced Video and Signal Based Surveillance (AVSS), 2015 12th IEEE International Conference on, IEEE, pp 1–6
[19] Huet B, Baralis E, Garza P, Kavoosifar MR (2016) Eurecom-Polito at TRECVID 2016: Hyperlinking task. In: Working Notes of the TRECVID 2016 Workshop
[20] Huet B, Baralis E, Garza P, Kavoosifar MR (2017) Eurecom-Polito at TRECVID 2017: Hyperlinking task. In: Working Notes of the TRECVID 2017 Workshop
[21] Kumar J (2015) Apache Solr Search Patterns. Packt Publishing Ltd

[22] Lamel L (2012) Multilingual speech processing activities in quaero: Application to multimedia search in unstructured data. In: Baltic HLT, pp 1–8
[23] Leskovec J, Rajaraman A, Ullman JD (2014) Mining of massive datasets. Cambridge university press
[24] Lin D, et al (1998) An information-theoretic definition of similarity. In: Icml, Citeseer, vol 98, pp 296–304
[25] Liu X, Troncy R, Huet B (2011) Finding media illustrating events. In: Proceedings of the 1st ACM International Conference on Multimedia Retrieval, ACM, p 58
[26] McCandless M, Hatcher E, Gospodnetic O (2010) Lucene in action: covers Apache Lucene 3.0. Manning Publications Co.
[27] Moumtzidou A, Mironidis T, Apostolidis E, Markatopoulou F, Ioannidou A, Gialampoukidis I, Avgerinakis K, Vrochidis S, Mezaris V, Kompatsiaris I, et al (2016) Verge: a multimodal interactive search engine for video browsing and retrieval. In: International Conference on Multimedia Modeling, Springer, pp 394–399
[28] Nakagawa A, Kutics A, Tanaka K, Nakajima M (2003) Combining words and object-based visual features in image retrieval. In: Image Analysis and Processing, 2003. Proceedings. 12th International Conference on, IEEE, pp 354–359
[29] Okuoka T, Takahashi T, Deguchi D, Ide I, Murase H (2009) Labeling news topic threads with wikipedia entries. In: Multimedia, 2009. ISM'09. 11th IEEE International Symposium on, IEEE, pp 501–504
[30] Quattoni A, Collins M, Darrell T (2007) Learning visual representations using images with captions. In: Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, IEEE, pp 1–8
[31] Racca DN, Jones GJ (2015) Evaluating search and hyperlinking: An example of the design, test, refine cycle for metric development. In: MediaEval
[32] Resnik P (1995) Using information content to evaluate semantic similarity in a taxonomy. arXiv preprint cmp-lg/9511007
[33] Rossetto L, Giangreco I, Tănase C, Schuldt H (2017) Multimodal video retrieval with the 2017 imotion system. In: Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, ACM, pp 457–460
[34] Safadi B, Sahuguet M, Huet B (2014) When textual and visual information join forces for multimedia retrieval. In: Proceedings of International Conference on Multimedia Retrieval, ACM, p 265
[35] Schmiedeke S, Xu P, Ferrané I, Eskevich M, Kofler C, Larson MA, Estève Y, Lamel L, Jones GJ, Sikora T (2013) Blip10000: A social video dataset containing spug content for tagging and retrieval. In: Proceedings of the 4th ACM Multimedia Systems Conference, ACM, pp 96–101
[36] Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556
[37] Smiley D, Pugh E, Parisa K, Mitchell M (2015) Apache Solr enterprise search server. Packt Publishing Ltd
[38] Soleymani M, Riegler M, Halvorsen P (2018) Multimodal analysis of user behavior and browsed content under different image search intents. International Journal of Multimedia Information Retrieval 7(1):29–41
[39] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: Proceed-

ings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1–9

[40] Tan S, Ngo CW, Tan HK, Pang L (2011) Cross media hyperlinking for search topic browsing. In: Proceedings of the 19th ACM international conference on Multimedia, ACM, pp 243–252

[41] Tanase C, Giangreco I, Rossetto L, Schuldt H, Seddati O, Dupont S, Altiok OC, Sezgin M (2016) Semantic sketch-based video retrieval with autocompletion. In: Companion Publication of the 21st International Conference on Intelligent User Interfaces, ACM, pp 97–101

[42] Tani MYK, Ghomari A, Lablack A, Bilasco IM (2017) Ovis: ontology video surveillance indexing and retrieval system. International Journal of Multimedia Information Retrieval 6(4):295–316

[43] Tay Y, Phan MC, Tuan LA, Hui SC (2017) Learning to rank question answer pairs with holographic dual lstm architecture. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, pp 695–704

[44] Verma Y, Jha A, Jawahar C (2018) Cross-specificity: modelling data semantics for cross-modal matching and retrieval. International Journal of Multimedia Information Retrieval 7(2):139–146

[45] Vukotić V, Raymond C, Gravier G (2016) Bidirectional joint representation learning with symmetrical deep neural networks for multimodal and crossmodal applications. In: Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, ACM, pp 343–346

[46] Vukotić V, Raymond C, Gravier G (2018) A crossmodal approach to multimodal fusion in video hyperlinking. IEEE MultiMedia 25(2):11–23

[47] Wu Z, Palmer M (1994) Verbs semantics and lexical selection. In: Proceedings of the 32nd annual meeting on Association for Computational Linguistics, Association for Computational Linguistics, pp 133–138