

IRISA at MediaEval 2015: Search and Anchoring in Video Archives Task

Anca-Roxana Simon, Guillaume Gravier, Pascale Sébillot

▶ **To cite this version:**

Anca-Roxana Simon, Guillaume Gravier, Pascale Sébillot. IRISA at MediaEval 2015: Search and Anchoring in Video Archives Task. Working Notes Proceedings of the MediaEval Workshop, 2015, Wurzen, Germany. hal-01196176

HAL Id: hal-01196176
https://hal.science/hal-01196176

Submitted on 9 Sep 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

IRISA at MediaEval 2015: Search and Anchoring in Video Archives Task

Anca-Roxana Şimon*, Guillaume Gravier**,
Pascale Sébillot***
IRISA & INRIA Rennes
Univ. Rennes 1*, CNRS**, INSA***
35042 Rennes Cedex, France
firstname.lastname@irisa.fr

ABSTRACT

This paper presents our approach and results in the Search and Anchoring in Video Archives task at MediaEval, 2015. The *Search* part aims at returning a ranked list of video segments that are relevant to a textual user query. The *Anchoring* part focuses on identifying video segments that would encourage further exploration within the archive. A two step approach is implemented for both sub-tasks. The first step is common to both. This step consists in generating a list of potential anchor segments and response-query segments relying on a hierarchical topical structuring technique; In the second step, for each query, the best 20 segments are selected according to content based comparisons, while for the anchor detection sub-task, the segments are ranked based on a cohesion measure. The use of the hierarchical topical structure helps to propose segments of variable length at different levels of details with precise jump-in points for them. More, the algorithm deriving the structure relies on the burstiness phenomenon in word occurrences which gives an advantage over the classical bag-of-words model.

1. INTRODUCTION

This paper presents the participation of IRISA at the MediaEval 2015 Search and Anchoring in Video Archives task [2]. The task is composed of two sub-tasks. The first one is a search scenario in a video collection. Starting from a two-field query, where one field refers to spoken content and the other refers to the visual content, the goal is to retrieve video segments that contain the requested information (audio or visual). The second sub-task consists in automatically selecting video segments, called *anchors*, for a list of videos in the collection, for further hyperlinking within the archive. The solutions should help the users to find relevant information in the archive and also to improve the browsing and navigation experience by providing anchor points that can lead to further discoveries in the archive.

The system that we propose consists of a two step process, with the first one being in common for both sub-tasks. The first step consists of extracting video segments that will be used in a second step to represent anchor segments or segments that respond to the queries issued by users. This first step is usually done using fixed-length segmentation [3, 8] or linear topic segmentation strategies [1]. We believe it is important to extract segments with precise jump-in points and at various levels of details. This means creating anchors that cover a more general topic or different points of view on some topic. While for the search part the results retrieved could offer a general perspective or a more focused one. Differently from what it is generally done, we choose to represent each video as a *hierarchy of topically focused segments* using the algorithm proposed in [10].

The advantage of using the algorithm proposed in [10] is that it helps to identify the salient information in the videos. Moreover, having a hierarchical representation, the segments we provide as results can be at different granularity, i.e., more specific or more general, offering different levels of details. The algorithm is build upon the *burstiness* phenomenon in word occurrences. In practice words tend to appear in bursts, i.e., if a word appears once it is more likely to appear again, instead of independently [6]. Several studies for statistical laws in language have proposed burst detection models that analyze the distributional pattern of words [9, 7]. We believe that such an approach brings more focus to what is extracted from the videos, and not only to the content-based comparisons and analysis part.

2. SYSTEM OVERVIEW

The aim of our approach is first to find precise jump-in points to the salient segments in the videos, at various levels of details. These segments are obtained by applying the algorithm proposed in [10] (denote it HTFF), which outputs a hierarchy of topically focused fragments for each video. HTFF relies on text-like data. Therefore, we exploit spoken data obtained from automatic transcripts and manual subtitles [4] and visual concepts detected for each video[11]. More details about the data can be found in [2]. After obtaining the topically focused fragments we perform content analysis and comparisons to propose the top segments for the two sub-tasks.

Subsections 2.1–2.3 detail the following: 2.1, the generation of potential anchor and query-response segments; 2.2, the selection of the top 20 segments for each query; 2.3, the ranking of the anchor segments;

2.1 Hierarchy of topically focused segments

Each video in the test collection is partitioned into a hierarchy of topically focused fragments with the automatic segmentation algorithm HTFF, which is domain independent, needs no a priori information and has proven to offer a good representation of the information contained in the video. It

Copyright is held by the author/owner(s).
MediaEval 2015 Workshop, Sept. 14-15, 2015, Wurzen, Germany

can be applied on any text-like data. For the search sub-task we are provided also with the visual query, i.e., visual concept words. Thus, we can apply HTFF also on the textual representation of a video formed by the visual concepts detected for each keyframe in that video. For the search sub-task we use also the LIMSI transcripts, while for the anchor detection sub-task we rely on the LIMSI transcripts and the manual transcripts. For applying the algorithm, data in the transcripts are first lemmatized and only nouns, non modal verbs and adjectives are kept.

The core of HTFF is Kleinberg's algorithm [5] used to identify word bursts, together with the intervals where they occurred. A burst interval corresponds to a period where the word occurs with increased frequency with respect to the normal behavior. Kleinberg's algorithm outputs a hierarchy of burst intervals for each word, taking one word at a time (for more details see [5]). The HTFF algorithm generates a hierarchy of salient topics using an agglomerative clustering of burst intervals found with [5]. The result is a set of nested topically focused fragments which are hierarchically organized. Next, we describe how the best segments are proposed for each sub-task.

2.2 Search sub-task

A cosine similarity measure is computed between each query and the content of the segments previously retrieved. This measure is computed with segments from all levels in the hierarchy and the ones for which higher similarity is obtained compared to the others will be ranked higher. In this setting, short, focused and highly similar segments are favored. This procedure is done both for textual and visual query independently.

2.3 Anchor selection sub-task

After having the list of salient segments for each of the video for which anchors need to be extracted, we compute a cohesion measure to rank these fragments. The measure is a probabilistic one where lexical cohesion for a segment S_i is computed using a Laplace law as in [12], i.e.,

$$C(S_i) = log \prod_{j=1}^{n_i} \frac{f_i(w_j^i) + 1}{n_i + k} \quad ,$$

where n_i is the number of word occurrences in S_i, $f_i(w_j^i)$ is the number of occurrences of the word w_j^i in segment S_i and k is the number of words in $\mathcal{V}$. The quantity $C(S_i)$ increases when words are repeated and decreases consistently when they are different. Using HTFF for anchor detection does not ensure any number of anchor segments to be found for a video. Therefore, some videos might have more or less anchors proposed than others. This is realistic, since the number of anchors that can be found in a video depends on the information contained.

3. RESULTS

For the search sub-task, 30 test set queries were defined. The top 10 results for each query were evaluated for each method, using crowd-sourcing technologies. The official evaluation results for the search sub-task are reported in Table 1. LIMSI denotes the system using LIMSI automatic transcripts and textual query, while Visual denotes the system relying on visual concepts and visual query. The best results are obtained with the LIMSI system. Analyzing the list of all the segments proposed by participants, it can be observed that with our approach the segments proposed are shorted in duration. Figure 1 illustrates the duration of the segments that were judged relevant or not with both our systems (LIMSI and Visual) compared to those proposed by other participants. The segments we proposed are on average less than half the size of the segments proposed by other participants. This was detrimental to our approach. Some of the short segments, proposed with our methods, are judged not relevant. While long segments which cover these short segments are judged relevant. However, many of our short segment do not overlap with longer segments proposed by others, so in the end they remain judged as not relevant.

	P_5	P_10	P_20
LIMSI	0.34	0.31	0.19
Visual	0.12	0.11	0.06

Table 1: Precision values obtained for all proposed methods for the search sub-task.

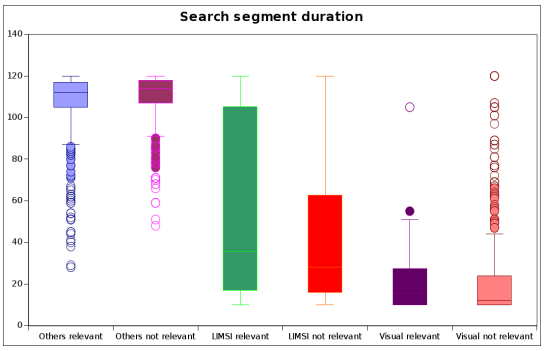


Figure 1: Boxplots showing segment duration variation proposed by others participants and our systems (i.e., LIMSI and Visual)

	Precision	Recall	MRR
LIMSI	0.557	0.435	0.773
Manual	0.469	0.38	0.735

Table 2: Precision, recall and MRR values obtained for all proposed methods for the anchor detection sub-task.

For the anchor sub-task, a list of 33 videos was defined, for which anchors had to be proposed. The top-25 ranks for each video and each method were judged by crowd-sourcing using Amazon Mechanical Turk workers who gave their opinion on these segments taken from the context of the videos. Precision, recall and Mean Reciprocal Rank (MRR) measures have been used. The results obtained for both our systems LIMSI (using LIMSI automatic transcripts) and Manual (using subtitles) are reported in Table 2. The best results were obtained when relying on automatic transcripts.

4. CONCLUSION

The results obtained on both sub-tasks show that while for anchor detection short segments are a good idea, for the search sub-task, assessors seem to need more context to find a segment relevant. For future work on the search sub-task we consider selecting larger segments from a higher level in the hierarchy (i.e., coarse grain). Additionally, combining visual and textual burst could improve the results. For the anchor detection task, different ways to rank the segments could be considered, favoring segments which contain named entities or visual bursts.

5. REFERENCES

[1] C. Bhatt, N. Pappas, M. Habibi, and A. Popescu-Belis. IDIAP at MediaEval 2013: Search and hyperlinking task. In *Working Notes Proc. of the MediaEval Workshop*, 2013.

[2] M. Eskevich, R. Aly, R. Ordelman, D. N. Racca, S. Chen, and G. J. F. Jones. SAVA at MediaEval 2015: Search and anchoring in video archives. In *Working notes of the MediaEval 2015 Workshop*, Wurzen, Germany, 2015.

[3] P. Galuščáková and P. Pecina. CUNI at MediaEval 2014 search and hyperlinking task: Search task experiments. In *Working Notes Proc. of the MediaEval Workshop*, 2014.

[4] J.-L. Gauvain, L. Lamel, and G. Adda. The LIMSI broadcast news transcription system. *Speech Communication*, 37(1-2):89–108, 2002.

[5] J. Kleinberg. Bursty and hierarchical structure in streams. In *8th ACM SIGKDD International Conf. on Knowledge Discovery and Data Mining*, pages 91–101, 2002.

[6] R. E. Madsen, D. Kauchak, and C. Elkan. Modeling word burstiness using the dirichlet distribution. In *Proceedings of the 22Nd International Conference on Machine Learning*, ICML '05, pages 545–552, New York, NY, USA, 2005. ACM.

[7] R. E. Madsen, D. Kauchak, and C. Elkan. Modeling word burstiness using the dirichlet distribution. In *Proceedings of the 22Nd International Conference on Machine Learning*, ICML '05, pages 545–552, New York, NY, USA, 2005. ACM.

[8] D. N. Racca, M. Eskevich, and G. J. F. Jones. Dcu search runs at MediaEval 2014 search and hyperlinking. In *Working notes of the MediaEval 2014 Workshop*, Barcelona, SPAIN, 2014.

[9] A. Sarkar, P. H. Garthwaite, and A. De Roeck. A bayesian mixture model for term re-occurrence and burstiness. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, CONLL '05, pages 48–55. Association for Computational Linguistics, 2005.

[10] A.-R. Simon, P. Sébillot, and G. Gravier. Hierarchical topic structuring: from dense segmentation to topically focused fragments via burst analysis. In *Recent Advances in NLP*, Hissar, Bulgaria, 2015.

[11] T. Tommasi, R. B. N. Aly, K. McGuinness, K. Chatfield, and et al. Beyond metadata: searching your archive based on its audio-visual content. In *International Broadcasting Convention*, 2014.

[12] M. Utiyama and H. Isahara. A statistical model for domain-independent text segmentation. In *Association for Computational Linguistics*, Toulouse, France, 2001.