Detecting Hot Spots in Web Videos

José Luis Redondo García[1], Mariella Sabatino[1],
Pasquale Lisena[1], Raphaël Troncy[1]

EURECOM, Sophia Antipolis, France,
{redondo, mariella.sabatino, pasquale.lisena, raphael.troncy}@eurecom.fr

Abstract. This paper presents a system that detects and enables the exploration of relevant fragments (called Hot Spots) inside educational online videos. Our approach combines visual analysis techniques and background knowledge from the web of data in order to quickly get an overview about the video content and therefore promote media consumption at the fragment level. First, we perform a chapter segmentation by combining visual features and semantic units (paragraphs) available in transcripts. Second, we semantically annotate those segments via Named Entity Extraction and topic detection. We then identify consecutive segments talking about similar topics and entities that we merge into bigger and semantic independent media units. Finally, we rank those segments and filter out the lowest scored candidates, in order to propose a summary that illustrates the Hot Spots in a dedicated media player. An online demo is available at http://linkedtv.eurecom.fr/mediafragmentplayer.

Keywords: Semantic Video Annotation, Media Fragments, Summarization

1 Introduction

Nowadays, people consume all kind of audiovisual content on a daily basis. From breaking news to satiric videos, personal recordings or cooking tutorials, we are constantly feed by video content to watch. A common practice by viewers consists in fast browsing through the video, using sometimes the key frames provided by the video sharing platform, with the risk of missing the essence of the video. This phenomena is even more obvious when it comes to educational web content. A study made over media entertainment streaming services reveals that the majority of partial content views (52.55%) are ended by the user within the first 10 minutes, and about 37% of these sessions do not last past the first five minutes [5]. In practice, it is difficult and time consuming to manually gather video insights that give the viewers a fair understanding about what the video is talking about. Our research tackles this problem by proposing a set of automatically annotated media fragments called Hot Spots, which intend to highlight the main concepts and topics discussed in a video. We also propose a dedicated exploring interface that eases the consumption and sharing of those hot spots.

The challenge of video segmentation has been addressed by numerous previous research. Some of them rely exclusively on low-level visual features such as color histograms or visual concept detection clustering operations [4]. Other approaches rely on text, leveraging the video transcripts and sometimes manual annotations and comments attached to the video [6] while the combination of both text and visual features is explored in [1]. Our approach combines also both visual and textual features with the added value of leveraging structured knowledge available in the web of data.

2 Generating and Exploring Hot Spots in Web Videos

This demo implements a multimodal algorithm for detecting and annotating the key fragments of a video in order to propose a quick overview about what are the main topics being discussed. We conduct an experiment over a corpora of 1681 TED talks [1], a global set of conferences owned by the private non-profit Sapling Foundation under the slogan: "Ideas Worth Spreading"

2.1 Media Fragments Generation

First, we perform *shot* segmentation for each video using the algorithm described in [3]. Shots are the smallest unit in a video, capturing visual changes between frames but not necessary reflecting changes of topic being discussed in the video. Therefore, we introduce the notion of *chapters* corresponding to wider chunks illustrating particular topics. In order to obtain such fragments, we use specific marks embedded in the available video transcripts for all TED talks that indicate the start of new paragraphs. In a last step, those fragments are combined with visual shots. Hence, we adjust the boundaries of each chapter using both paragraph and shot boundaries.

2.2 Media Fragments Annotation

We rely on the subtitles available for the 1681 TED talks for annotating the media fragments which have been generated. More precisely, we detect topics and named entities. For the former, we have used the dedicated TextRazor topic detection method[2], while for the latter, we used the NERD framework [2]. Both entities and topics come with a relevance score which we use to give a weight to this particular semantic unit within the context of the video story. Topics and named entities are attached to a chapter.

2.3 Hot Spots Generation

Once all chapters are delimited and annotated, we iteratively cluster them, in particular, when temporally close segments are similar enough in terms of topics named entities. More precisely, we compute a similarity function between

[1] `http://www.ted.com/`
[2] `https://www.textrazor.com/documentation`

consecutive pairs of segments S_1 and S_2 until no new merges are possible. This comparison leverages on the annotations attached to each segment by analyzing the number of coincidences between topics $T = max_3 \left\{ \sum_{topic_i} Rel_i \right\}$ and entities $E = max_{5W's} \left\{ \sum_{entity_i} Rel_i \right\}$, where Rel_i is the TexRazor's relevance:

$$d\left(S_1, S_2\right) = w_{topic} \cdot \left(\frac{|T_1 \bigcap T_2|}{max\left\{|T_1|, |T_2|\right\}} \right) + w_{entity} \cdot \left(\frac{|E_1 \bigcap E_2|}{max\left\{|E_1|, |E_2|\right\}} \right) \quad (1)$$

After this clustering process, the video is decomposed into less but longer chapters. However, there are still too many candidates to be proposed as Hot Spots. Therefore, we filter out those fragments which contain potentially less interesting topics. We define a function for measuring the interestingness of a video segment, which directly depends on the relevance and frequency of the annotations and which is inversely proportional to its length. In our current approach, the Hot Spots are those fragments whose relative relevance falls under the first quarter of the final score distribution.

In a last step, for each Hot Spot, we also generate a summarization to be shown in a dedicated media player where we highlight the main topics T and entities E which have been discovered.

2.4 Exploring Hot Spots within TED Talks

The Hot Spots and their summaries are visualized in a user friendly Media Fragment URI compliant media player. The procedure to get the Hot Spots for a particular Ted talk is the following: the user enters a valid TED Talk URL to get a landing page (Figure 1a). When the results are available, the hot spots are highlighted on the timeline together with the label of the most relevant chapter annotation (Figure 1b). This label can be extended to a broader set of entities and topics (Figure 1c). Finally, the user can always share those hot spots segments using media fragment URIs (Figure 1d).

3 Discussion

We have presented a demo for automatically discovering Hot Spots in online and educational videos. We leverage on visual analysis and background knowledge available in the web of data for detecting what fragments illustrate the best the main topics discussed in the video. Those Hot Spots allow the viewer to quickly decide if a video is worth watching and will provide incentive for consuming videos at the fragment level. In addition, Hot Spots can be explored in a dedicated media fragment player which also display the attached semantic annotations.

We plan to carry out an exhaustive evaluation of our approach involving real users feedback, in order to optimize the results of our Hot Spot detection algorithm and to improve the usability and efficiency of the developed interface.

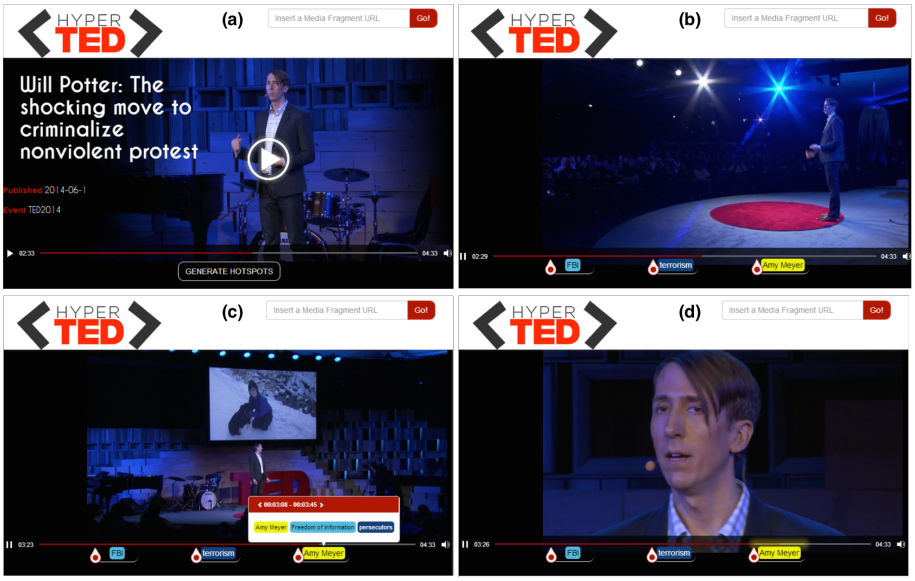

Fig. 1: Visualizing the Hot Spots of a TED Talk (available at `http://linkedtv.eurecom.fr/mediafragmentplayer/video/bbd70fff-e828-4db5-80d0-1a4c9aea430e`)

We also plan to further exploit the segmentation results and their corresponding annotations for establishing links between fragments belonging to different videos in order to generate true hyperlinks within a closed collection such as TED talks and make results available following Linked Data principles.

References

1. S.-F. Chang, R. Manmatha, and T.-S. Chua. Combining text and audio-visual features in video indexing. In *In IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05)*, 2005.
2. G. Rizzo and R. Troncy. NERD: A Framework for Unifying Named Entity Recognition and Disambiguation Extraction Tools. In *13th Conference of the European Chapter for Computational Linguistics (EACL'12)*, Avignon, France, 2012.
3. P. Sidiropoulos, V. Mezaris, I. Kompatsiaris, H. Meinedo, M. Bugalho, and I. Trancoso. Temporal video segmentation to scenes using high-level audiovisual features. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(8):1163–1177, 2011.
4. C. G. Snoek and M. Worring. Multimodal video indexing: A review of the state-of-the-art. *Multimedia tools and applications*, 25(1):5–35, 2005.
5. H. Yu, D. Zheng, B. Y. Zhao, and W. Zheng. Understanding user behavior in large-scale video-on-demand systems. In *In 1st ACM SIGOPS/EuroSys European Conference on Computer Systems*, pages 333–344, 2006.
6. Z.-J. Zha, M. Wang, J. Shen, and T.-S. Chua. Text mining in multimedia. In *Mining Text Data*, pages 361–384. Springer, 2012.