Interactive exploration of journalistic video footage through multimodal semantic matching

Ibrahimi, Sarah; Chen, Shuo; Arya, Devanshu; Câmara, Arthur; Chen, Yunlu; Crijns, Tanja; Van Der Goes, Maurits; Mensink, Thomas; Van Miltenburg, Emiel; More Authors

DOI
10.1145/3343031.3350597

Publication date
2019

Document Version
Final published version

Published in
MM 2019 - Proceedings of the 27th ACM International Conference on Multimedia

Interactive Exploration of Journalistic Video Footage through Multimodal Semantic Matching

Sarah Ibrahimi[1], Shuo Chen[1], Devanshu Arya[1], Arthur Câmara[2], Yunlu Chen[1], Tanja Crijns[3], Maurits van der Goes[3], Thomas Mensink[4], Emiel van Miltenburg[5], Daan Odijk[3], William Thong[1], Jiaojiao Zhao[1], Pascal Mettes[1]
[1]University of Amsterdam, [2]Delft University of Technology, [3]RTL Nederland, [4]Google Research, [5]Tilburg University

ABSTRACT

This demo presents a system for journalists to explore video footage for broadcasts. Daily news broadcasts contain multiple news items that consist of many video shots and searching for relevant footage is a labor intensive task. Without the need for annotated video shots, our system extracts semantics from footage and automatically matches these semantics to query terms from the journalist. The journalist can then indicate which aspects of the query term need to be emphasized, e.g. the title or its thematic meaning. The goal of this system is to support the journalists in their search process by encouraging interaction and exploration with the system.

CCS CONCEPTS

• **Information systems** → *Search interfaces*; • **Human-centered computing** → *Systems and tools for interaction design.*

KEYWORDS

video; multimodal; exploration; semantics; matching

1 INTRODUCTION

Journalists produce hours of video content every day. A typical news broadcast has several news items, consisting of both raw footage from reporters and existing footage e.g. from news wires or old news items. All the raw footage and edited news items are archived for future reuse. Nowadays, documentation departments manually annotate this footage. Editors, or even a dedicated search team, can search such archives for relevant videos related to news events. Because this process is time consuming, we aim to automate the annotation process and optimize the search for journalistic footage.

We present a comprehensive system that offers a way to search through videos for relevant footage without the need to manually annotate any of these videos. The framework consists of two main components: multimodal semantic extraction and interaction, see Fig. 1. We extract semantics from videos and match them with composite query terms. Next, the framework provides the journalist an interactive way to explore retrieved footage, either by updating queries or by altering control sliders for relevant concepts. Several multimodal video retrieval systems have previously been proposed [2, 4, 9]. In these systems the interactive component is either missing [2] or in the form of relevance feedback [4] to improve future model results. In [9], the user is able to refine results to indicate the importance of different features. Contrarily, our system places exploration as its central component. Users interact with the system by deciding on the importance of both features and their semantics. This is vital for journalistic footage, as search can be direct or semantic. Consider the word 'fire', which can be used for a real fire, but also for the expression 'being under fire'. In journalistic footage, it is *a priori* unknown which type is desirable. We emphasize exploration to handle query diversity.

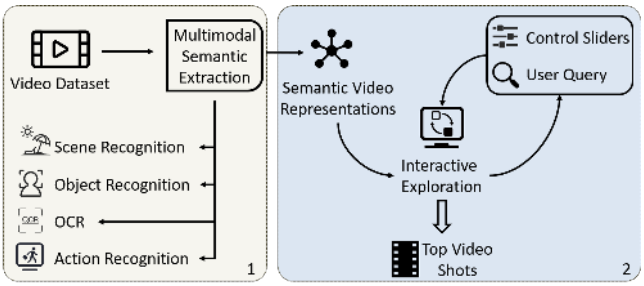

Figure 1: Overview of the framework. Part 1 focuses on Multimodal Semantic Extraction. Part 2 lets the user interact with the system to retrieve results.

2 FRAMEWORK

First, we extract semantics from video footage, followed by a phase of interaction between the user and the system. This consists of semantic query matching and exploration of the retrieved footage.

2.1 Multimodal Semantic Extraction

For multimodal semantic extraction, we focus on four types of recognition to get an overview of the most important aspects of the video. These are scenes, objects, optical characters, and actions. These are extracted for every tenth frame of each video, except for actions where a spatio-temporal approach is used.

For *Scene Recognition*, we employ a ResNet18[3] architecture pre-trained on the Places365 dataset [12]. Each frame obtains a classification as an indoor or outdoor scene, a probabilities for 365 scenes (e.g. baseball field, hotel room), and a probabilities for 102 scene attributes (e.g. water, mountains). These probabilities are averaged for each shot to obtain shot-level predictions. Shots are detected based on frame differencing. *Object Recognition* is done by employing a ResNet101 architecture pre-trained on 1,000 ImageNet

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10.
https://doi.org/10.1145/3343031.3350597

A B C

Figure 2: Illustration of the search process in journalistic footage. A: A user enters the query 'vuur '(fire) and the first matching results are presented. The system focuses on query terms in news titles, with top results about people being 'onder vuur '(under fire). B: The importance of news titles is reduced and the scene semantic item 'vulkaan '(vulcano) is emphasized, with drastically different results, highlighting the importance of exploration. C: Separate view of the most relevant video, with the shots and the most important multimodal semantics related to the video.

categories [10]. This results in a 1,000-dimensional probability distribution and converted in a probability on shot level akin to scenes. *Optical Character Recognition* is mainly used to extract the subtitles in the videos. We employ the open-source OCR engine Tesseract [11]. After cleaning the results by confidence score thresholding, fuzzy matching is performed to a large dictionary, which is a combination of the words in FastText [6] and OpenTaal 2.10. Results are concatenated on shot level. The spatio-temporal nature of videos is investigated by *Action Recognition*. For this task, the I3D network architecture [1] pre-trained on the Kinetics-400 dataset [7] is used. This results in a score for 400 actions in each video shot (e.g. motorcycling, riding a bike) with at least 64 frames.

2.2 Interaction

2.2.1 Semantic Query Matching. Given the multimodal semantic extraction for videos, we match the video semantics to query terms. Two types of matching can be distinguished: textual and semantic. Textual matching relies on OCR recognition and semantic matching on the scenes, object, and action recognition. For both, the Elasticsearch search engine is used to index and search the video files, resulting in a textual and semantical search functionality. BM25 is used as the Elasticsearch matching algorithm [5]. The BM25 search matches every textual field in the video file, attributing it a score. The score for each field (OCR, title, and description) is raised by a factor on the user-defined score. The final score is given by the sum of the individual scores. For the semantic search, the framework relies on the gensim [8] implementation of FastText word embeddings. By retrieving the word embeddings from the Dutch textual representation of each concept from the video models, a visually semantic search from the query embedding is performed and the top-k concepts for each field are retrieved. This search focuses on visual semantics. The cosine distance between the query and textual concept embeddings is used as a score function.

2.2.2 Exploration. The ranked match between user queries and video shots is presented in the interaction component of the system. This consists of a front-end environment that lets the user interact with the results. Fig. 2 presents several views of the framework. We aim for high recall, since journalists should not miss relevant available footage in their search. Apart from changing the query term, a set of sliders is presented to interact easily with the results. Sliders exist for top-k results for scene categories, scene attributes, objects, text related components, and actions. This slider-based approach enables the user to specify the importance of components of top results. For example, the action in the video might be more important than the scene, which cannot easily be specified by a query. Changing a slider immediately affects the results presented below the sliders, which results in fast and easy browsing through large scale video datasets. Hence this interactive component is necessary to re-rank the results to retrieve the most suitable footage.

3 CONCLUSION

This system presents a way for journalists to search in videos that are not annotated for reference materials. User queries are matched with video and textual semantics and it presents top-k results in an interactive visualization. Users explore the visualization by emphasizing semantics from different sources. This facilitates journalists in their search for suitable shots. The system enables an efficient exploration for videos and alleviates the need for manual labeling.

ACKNOWLEDGMENTS

This research was part of the ICT with Industry Workshop 2019, organised by ICT Research Platform Netherlands and NWO.

REFERENCES

[1] João Carreira and Andrew Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *CVPR*.

[2] Maria Eskevich, Huynh Nguyen, Mathilde Sahuguet, and Benoit Huet. 2015. Hyper Video Browser: Search and Hyperlinking in Broadcast Media. In *ACM Multimedia*.
[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*.
[4] Lu Jiang, Shoou-I Yu, Deyu Meng, Teruko Mitamura, and Alexander G. Hauptmann. 2015. Bridging the Ultimate Semantic Gap: A Semantic Search Engine for Internet Videos. In *ICMR*.
[5] K. Sparck Jones, S. Walker, and S.E. Robertson. 2000. A probabilistic model of information retrieval: development and comparative experiments: Part 1. *Information Processing Management* (2000).
[6] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. FastText.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651* (2016).
[7] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The Kinetics Human Action Video Dataset. *arXiv preprint arXiv:1705.06950* (2017).
[8] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*.
[9] Luca Rossetto, Ivan Giangreco, Claudiu Tănase, and Heiko Schuldt. 2017. Multimodal Video Retrieval with the 2017 IMOTION System. In *ICMR*.
[10] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. 2015. ImageNet Large Scale Visual Recognition Challenge. *IJCV* (2015).
[11] Ray Smith. 2007. An Overview of the Tesseract OCR Engine. In *Proc. Ninth Int. Conference on Document Analysis and Recognition (ICDAR)*. 629–633.
[12] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million Image Database for Scene Recognition. *TPAMI* (2017).