Video2Sentence and Vice Versa

Habibian, A.; Snoek, C.G.M.

DOI
10.1145/2502081.2502249
Publication date
2013
Document Version
Final published version
Published in
MM '13

Video2Sentence and Vice Versa

Amirhossein Habibian
ISLA, University of Amsterdam
Science Park 904
1098 XH, Amsterdam
The Netherlands
a.habibian@uva.nl

Cees G. M. Snoek
ISLA, University of Amsterdam
Science Park 904
1098 XH, Amsterdam
The Netherlands
cgmsnoek@uva.nl

ABSTRACT

In this technical demonstration, we showcase a multimedia search engine that retrieves a video from a sentence, or a sentence from a video. The key novelty is our machine translation capability that exploits a cross-media representation for both the visual and textual modality using concept vocabularies. We will demonstrate the translations using arbitrary web videos and sentences related to everyday events. What is more, we will provide an automatically generated explanation, in terms of concept detectors, on *why* a particular video or sentence has been retrieved as the most likely translation.

Categories and Subject Descriptors

I.2.10 [**Vision and Scene Understanding**]: Video analysis

General Terms

Algorithms, Experimentation

Keywords

Multimedia translation, event detection, concept vocabulary

1. INTRODUCTION

All of a sudden, video stories are everywhere. For sharing daily routines on YouTube, for private recollection of holidays, or for new professional markets. Anyone who must work with large amounts of video data is overwhelmed by its volume and the lack of tools to search and interpret the material easily. We aim to ease both retrieval and interpretation by translating a video to a sentence, similar to [1], and a sentence to video.

We consider semantic vocabularies for representing both video and human-provided sentences in a cross-media setting. Representing text and images by their semantic concepts has been shown to be beneficial for captioned-image retrieval [5]. At the same time, representing video by semantic concepts has been shown to be advantageous for video event recognition [8, 4, 3, 2]. In [4], for example, Merler *et al.* arrive at a robust high-level representation of video events using 280 concept detectors, which outperforms a low-level audiovisual representation. Our novelty is to demonstrate the capabilities of semantic representations for translating arbitrary web video into their lingual description, and vice versa. See Figures 1 and 2.

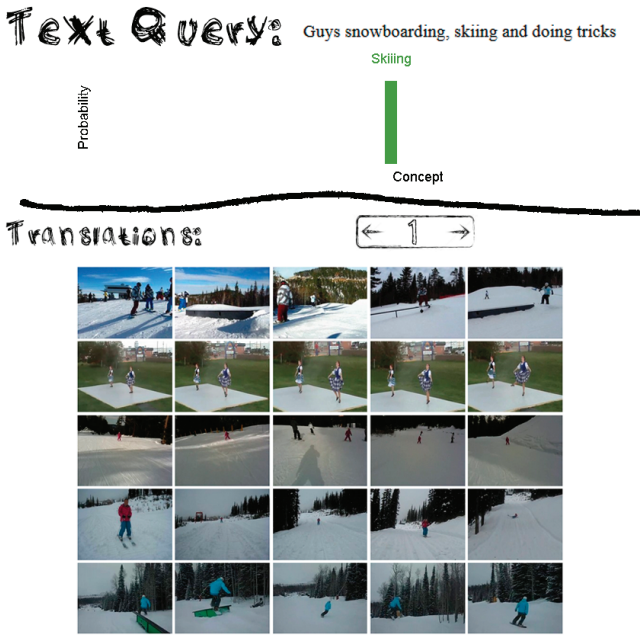


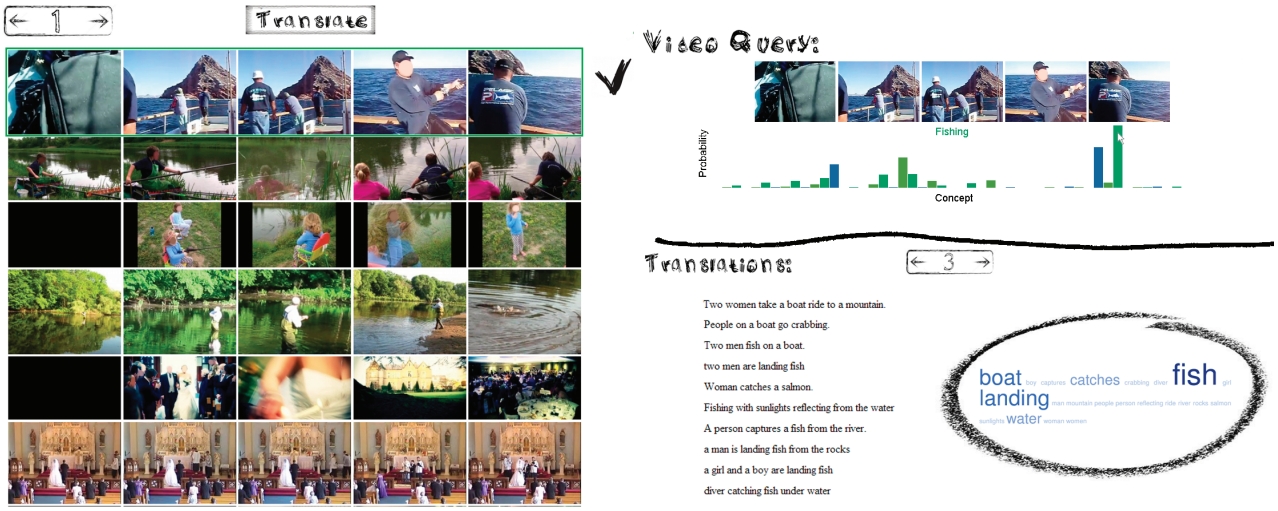
Figure 1: *Sentence2Video:* We automatically project an input sentence into a cross-media concept vocabulary, which is also used for representing video. The joint vocabulary allows us to translate a sentence to a video, and vice versa.

2. CONCEPT VOCABULARY

We demonstrate capabilities for translating arbitrary web video into a textual sentence and the other way around.

Data set We rely on the web video corpus from the TRECVID 2012 Multimedia Event Detection task [6]. For each web video in this corpus a textual description is provided that summarizes the event happening in the video by

MM'13, October 21–25, 2013, Barcelona, Spain.
ACM 978-1-4503-2404-5/13/10.
http://dx.doi.org/10.1145/2502081.2502249.



(a) *User selects video to translate.* (b) *System returns translations.*

Figure 2: *Video2Sentence:* Illustrative example for user interaction in our multimedia search engine. In (a) a user selects a candidate video for translation. In (b) the retrieved sentence translations are provided to the user together with the concept vocabulary of the query video. The tag cloud aggregates the sentences into a more precise translation.

highlighting its dominant concepts [7]. In addition, we rely on the publicly available 1,346 concept detector scores for this dataset, which we provided in [2].

Multimedia Translation We map the sentences and video into a unified semantic space and then find their cross-media relations within this space. To map the texts and videos into the semantic space, we apply two sets of concept detectors which are trained so as to detect the semantic concepts on both videos and texts. Hence the textual detectors correspond to the labels of the visual detectors. The cross-media retrieval metric simply minimizes the distance between joint-vocabulary probability vectors [5]. We will show how a vocabulary of concept detectors can be exploited for effective translation at a meaningful level.

3. DEMONSTRATION

During our demonstration we will focus on two multimedia translation use cases:

1. ***Video2Sentence:*** where we retrieve the sentences that best describe the visual content of the query video.

2. ***Sentence2Video:*** where we retrieve the videos that best illustrate the semantic content of the query sentence.

In addition, we will exhibit novel applications of this capability that aggregates the translated information for recounting. Taken together, the search engine provides a means to collect video examples illustrating a sentence and it provides a textual explanation of what happens in a video.

4. ACKNOWLEDGMENTS

This research is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20067. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government. This research is also supported by the STW STORY project and the Dutch national program COMMIT.

5. REFERENCES

[1] P. Das, R. K. Srihari, and J. J. Corso. Translating related words to videos and back through latent topics. In *WSDM*, 2013.

[2] A. Habibian, K. E. A. van de Sande, and C. G. M. Snoek. Recommendations for video event recognition using concept vocabularies. In *ICMR*, 2013.

[3] M. Mazloom, E. Gavves, K. E. A. van de Sande, and C. G. M. Snoek. Searching informative concept banks for video event detection. In *ICMR*, 2013.

[4] M. Merler, B. Huang, L. Xie, G. Hua, and A. Natsev. Semantic model vectors for complex video event recognition. *IEEE TMM*, 2012.

[5] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *ACM MM*, 2010.

[6] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In *ACM MIR*, 2006.

[7] S. Strassel, A. Morris, J. Fiscus, C. Caruso, H. Lee, P. Over, J. Fiumara, B. Shaw, B. Antonishek, and M. Michel. Creating HAVIC: Heterogeneous audio visual internet collection. In *LREC*, 2012.

[8] I. Tsampoulatidis, N. Gkalelis, A. Dimou, V. Mezaris, and I. Kompatsiaris. High-level event detection system based on discriminant visual concepts. In *ICMR*, 2011.