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

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
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] [9] [2]. In [1], 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 [3]. For each web video in this corpus a textual description is provided that summarizes the event happening in the video by
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.

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