Towards Reliable Automatic Multimodal Content Analysis

Olli-Philippe Lautenbacher
Liisa Tiittula
Maija Hirvonen
Dept. of Modern Languages
University of Helsinki
firstname.surname
@helsinki.fi

Jorma Laaksonen
Dept. of Computer Science
Aalto University
jorma.laaksonen
@aalto.fi

Mikko Kurimo
Dept. of Signal Processing
and Acoustics
Aalto University
mikko.kurimo
@aalto.fi

Abstract

This poster presents a pilot where audio description is used to enhance automatic content analysis, for a project aiming at creating a tool for easy access to large AV archives.

1 Introduction

This poster presents a pilot study for a new interdisciplinary project which aims at creating an automated, time-aligned and language-based access to large archives of audiovisual documents. The idea is to facilitate the work of researchers who wish to pinpoint particular segments of AV material without having to browse through entire data sets. The project analyses human descriptions and film-viewing patterns in order to integrate that knowledge into an automatic content analyser. The pilot was set out to compare the results of the automatic and human methods available for content description.

2 AD *vs.* AMCA

Currently verbal content description for retrieving visual data is still scarce, although different methods exist: *human-made audio description* (AD) verbalizes visual information for visually impaired people (Maszerowska & al 2014) but is a slow and costly process. *Automatic Multimodal Content Analysis* (AMCA), on the other hand, consists of computer-driven detection of visual and auditory elements from multimedia (Rohrbach & al 2015; Viitaniemi & al 2015). AMCA is cost-effective and produces consistent output, but is still insufficient for high-level semantic analysis.

Our project combines these approaches to create an automatically produced narrative, but which is more informative than a mere list of descriptive concepts.

3 The pilot and its tools

We are now tackling our first pilot, a 15-minute excerpt from a documentary (*Helsinki, forever*, Peter von Bagh, 2008), a genre which the whole project will be concentrating on.

3.1 Automatic tools

A preliminary AMCA has already been made, based on earlier filmic contents, giving lists of descriptive concepts for each picture as an output. Consider the following example:

Screenshot from *Helsinki, forever*.

For this shot of 301 frames, the AMCA provides the following occurrence numbers for concepts:

> Body_Parts (301); Man_Made_Thing (301); Outdoor (278); Legs (277); Building (254); Suits (245); Actor (184); Suburban (163); Person (141); etc.

Naturally, such concepts might seem counterintuitive for a human reading of an image, mainly because they do not inform us about the respective relevance of the various semantic elements retrieved from the picture. The AMCA concepts will thus need further filtering.

Another tool used for the visual description is automatic sentence-like caption generation per frame (Karpathy & Fei-Fei 2015), which will be combined with the abovementioned concept retriever. For the same shot, we now get for 97% of the frames:

> a man in a suit and tie standing in front of a building

For the audio, an automatic transcription of the dialogue and voice-over can be made, using

Proceedings of the 2015 Workshop on Vision and Language (VL'15), pages 6–7,
Lisbon, Portugal, 18 September 2015.

voice recognition (see Remes & al 2015). The output is a transcript that is coded on a confidence basis, informing the researcher on the degree of certainty of the recognized linguistic segments. A description of on-screen sounds, including automatic music recognition, could also enhance the validity of relevant concept retrieval.

3.2 Human input

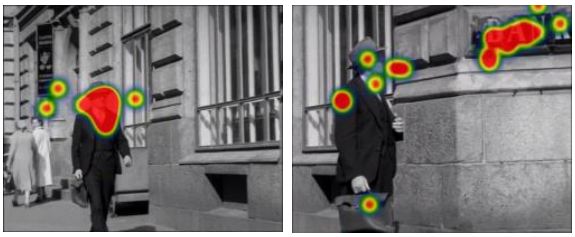
In order to improve these automatic describers, three human ADs of the excerpt were ordered from professionals. The comparison of those ADs is important for the pilot since it reveals the characteristics they share in terms of visual element selection and lexical choices (identicality of referents and words, synonymy, level of abstraction etc.). For our example shot, the ADs are (translated from Finnish):

> AD1: "A nervous looking **man** […] **stops** at the corner of the **bank** changing his *briefcase* from hand to hand and throwing glances *around him*."
>
> AD2: "A **man** […] **stops** in front of a '**Bank**' sign looking confused and *hesitating*, holding a *briefcase* with both hands."
>
> AD3: "A black suited **man stops** at the door of the **bank** and *hesitates*. He looks *around*, fingering his portfolio."

Some words are identical in all ADs (**man**; **stops**; **bank**), some concepts are almost synonymous (*briefcase* / *portfolio; hesitating* / *nervous looking*), and some expressions reveal a "point of view" (at the corner of *x* / in front of *x* / at the door of *x*; changing *y* from hand to hand / holding *y* with both hands / fingering *y*). It appears that all descriptions are similar in terms of the thematised entities and actions, but the various lexical items used in referring to them invites to re-evaluate the idea that there is only one equivalent description per image. All in all, the pilot studies the semantic variability of the descriptions by both qualitative and quantitative comparative analyses.

These human descriptions will then serve to feed the AMCA, helping to filter its concept-suggestions in terms of relevance, adequacy and degree of precision. For instance, key word lists created in a corpus analysis enable us to compare the descriptions, harmonize the content words of AD and finally merge them with the concepts suggested by the AMCA.

Furthermore, we also use eye tracking (Kruger & al, 2015) in the pilot, to identify convergence patterns in the gaze positions of average viewers watching the excerpt. This "natural viewing" gives further insight into the relevance of the visual element selection made by the AD and the AMCA. Within the selected shot, we can notice that people tend to look at the most informative parts of the image (the man's face and the "Bank" sign) especially during the first seconds of their appearance on screen:



SMI heat maps (21 viewers) on the same shot.

4 Outcomes of the pilot

This poster presentation includes a demo video of each of these tools and their respective outputs. Later on, all the collected data from the excerpt will be integrated to the AMCA to enhance its output, which can be further enriched by new human input. Such a recursive machine learning process will lead, eventually, to a reliable automatic description tool for documentary films.

References

Andrej Karpathy and Li Fei-Fei. 2015. Deep Visual-Semantic Alignments for Generating Image Descriptions. CVPR 2015.

Anna Maszerowska, Anna Matamala and Pilar Orero (eds). 2014. *Audio Description: New perspectives illustrated*. Benjamins, Amsterdam, NL / Philadelphia, USA.

Anna Rohrbach, Marcus Rohrbach, Niket Tandon and Bernt Schiele. 2015. A Dataset for Movie Description. CVPR 2015.

Jan-Louis Kruger, Agnieszka Szarkowska, Isabela Krejtz. 2015. Subtitles on the Moving Image: an Overview of Eye Tracking Studies. *Refractory – a Journal of Entertainment Media*, vol. 25.

Ulpu Remes, Ana Ramírez López, Kalle Palomäki and Mikko Kurimo. Forthcoming. Bounded conditional mean imputation with observation uncertainties and acoustic model adaptation. *IEEE Transactions on Audio, Speech and Language Processing*.

Ville Viitaniemi, Mats Sjöberg, Markus Koskela, Satoru Ishikawa and Jorma Laaksonen. 2015. Advances in Visual Concept Detection: Ten years of TRECVID. In Ella Bingham et al. (ed.): *Advances in Independent Component Analysis and Learning Machines*, 1st edition. Elsevier, Amsterdam, NL.