Analysis, Indexing and Visualization of Presentation Videos

Michele Merler
Department of Computer Science, Columbia University, New York
mmerler@cs.columbia.edu

ABSTRACT

The exponential diffusion of unstructured multimedia content on sites such as YouTube has lately fostered a rapid growth of interest in the multimedia community toward a new line of systems to recognize visual content "in the wild", that is, in less structured, unconstrained and more realistic domains. Due to the lack of structure and to the low quality of the data, algorithms and paradigms designed for professional content often cannot directly be applied to the aforementioned domains, thus presenting a new challenge.

One instance of such "wild videos" is represented by unstructured presentation videos, which are tools nowadays employed in a large variety of systems for different purposes, spanning from distance or e-learning to generation of conference proceedings, from corporate talks to student presentations.

We are interested in helping users efficiently and effectively access visual information, in particular technical information, and learn from such online multimedia sources. Our proposed system aims at summarizing, indexing, cross-referencing and browsing unstructured ("wild") presentation videos, with a focus on the quality of the produced indexes as perceived by the end users.

Categories and Subject Descriptors H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing

General Terms Experimentation, Human Factors

Keywords Presentation Video, Multi-modal Indexing

1. PROPOSED SOLUTION

We propose to index presentation videos based on four major cues: text (integrated with audio transcripts), speaker faces, graphics and mosaics.

Our system segments a video into semantically separated shots (each of which contains a different slide or blackboard image), automatically extracts the text from the content in the video (without any reference or electronic copies of the slides), builds a semantic index based on the obtained vocabulary of words, enriches such index with other visual cues such as diagrams and speaker faces, and finally represents the shot under the form of a mosaic image where the semantically relevant objects (text and graphics) are enhanced by highlighting and/or enlargement. In short, we want to situate emphasized content in enhanced context.

We extract and recognize the text of the slides directly from the presentation videos in which they appear [2]. Experiments on 8 presentation videos showed that adopting our novel binarization algorithm, Local Adaptive Otsu (LOA), before applying the open source Tesseract OCR engine helped to deal with the low quality of the detected video scene text and doubled performances, achieving 0.534 Precision and 0.744 Recall Character recognition rates.

We employ an optimal combination of 3 user preference measures (skin ratio, 3/4 pose and resolution) to maximize matching accuracy between face tracks (extracted with standard face detection and a new steady-state Kalman filter augmented tracking) and to select humanly preferred face icons as representative images for each speaker[3]. Our system achieved track matching accuracy of 94.22% on 3 presentation videos, building face summaries containing 54 out of 58 speakers from 795 detected tracks.

Finally, we will assert through user studies the performances of our system in an interactive environment, the relevance of our enhanced representation of semantic shots and the utility of the controls exposed to the user. For example a user study confirmed that humans prefer the head and shoulders, 3/4 view speaker indexes automatically produced by our system (detail in Figure 1) over frontal, face-only ones. We will use the VAST Multimedia Browser [1] as a testbed for our experiments, since it provides a powerful tool where our contributions can be easily integrated.

Our system will allow users to search for a presentation of professor X on topic Y, containing graph Z, and interact with an enhanced representation of the retrieved video shot.

2. REFERENCES