Title: Multimodal Understanding: toward joint audio/video processing for multimodal understanding process of video broadcast

Keywords: Multimedia Information Retrieval, Speech and Language Processing, Image Processing, Machine Learning

Location: Laboratoire d’Informatique Fondamentale de Marseille
Aix Marseille Université - LIF/CNRS (UMR 7279), Campus de Luminy, Marseille

Contact: Frédéric Béchet (frederic.bechet@univ-amu.fr)

Description

The growing quantity of multimedia content available through the Internet makes necessary the development of technologies that facilitate browsing and searching through this data. The scientific communities of image and audio processing have each developed numerous descriptors in order to characterize these multimedia documents, however these descriptors are all monomodal: speaker identification and speech transcription for audio processing; image classification, concept detection or scene classification for image processing.

Recently some evaluation programs such as the French Défi-Repère¹ (2011-2014) have targeted the idea of multimodal detection through tasks such as: "who is talking? Who is visible? Who is mentioned? Which names are written on screen?". The TALEP research team of the LIF lab of the Aix Marseille University is leading one of the team competing in this evaluation, and obtained the best results at the 2014 final evaluation.

The systems that have been developed during this program are all made of some monomodal components, such as speaker diarization or face detection processes, as well as some multimodal fusion modules in charge of aggregating all the different features and taking a final decision on the presence of persons in video broadcasts [Bredin12; Favre13]. Results obtained [Galibert13] have stressed on the one hand the good performance achieved by the different systems for non-ambiguous situations (the speaker is visible on screen); and on the other hand the necessity of developing better multimodal systems for complex situations.

This PhD proposal is set in this multimodal framework. The goal is to develop models that will go further than a fusion between monomodal descriptors. Based on preliminary work done in our team during the Defi-Repere on the concept of “multimodal understanding”, this research project aims at using multimodal features at each level of processing of a video document. For example, during the Defi-Repere, we have shown that visual features characterizing a TV set can help the speaker diarization and identification task. Similarly we want to add visual features into the spoken language understanding component of a video analysis system, and demonstrate that a certain level of understanding of the current video scene can be used at the language processing level.

¹ http://www.defi-repere.fr
The starting point of this research program is the 60h video corpus and the numerous annotations (both audio and visual) that have been made available through the Defi-Repere project. Only a portion of this annotation has been used during the challenge. The goal will be to use fully all these annotation levels for performing more advanced video analysis processes than person recognition. An extension on other kinds of video than video broadcasts is also considered.

Context

This PhD program is funded by the French DGA (Direction Générale de l’Armement) and the Aix Marseille University (AMU). It will be realized within the Natural Language Processing (NLP) team, called TALEP (Traitement Automatique de la Langue Ecrite et Parlée), of the Computer Science lab LIF (Laboratoire d’Informatique Fondamentale de Marseille – LIF/CNRS, UMR 7279) of AMU. One of the main topic studied in this team is the development of robust linguistic processing models applied to non-canonical text such as those produced by an Automatic Speech Recognition System or those that can be found in social media.

The TALEP team is involved in several national and international collaborative research projects linked to this PhD proposal. In particular this team is the coordinator of the consortium PERCOL, involving the University of Avignon, the University of Lille and France Telecom Orange Labs, which participated to the Défi-Reperé challenge funded by the French DGA and the ANR agency (2011-2014). The first workshop SLAM: Speech Language and Audio in Multimedia\(^2\) was organized in Marseille in 2013 by the TALEP team.

The supervisors for this PhD will be Frédéric Béchet, coordinator of the PERCOL team, and Benoit Favre. The following publication list contains some recent work from the team in areas related to this PhD proposal.

- **Multimedia processing in the Defi-Repere challenge**


- **Video segmentation**:


- **Conversation speech analysis with multi-view/multi-modal methods**:

  [Koco12] Sokol Koço, Cécile Capponi, Frédéric Béchet « Applying multiview learning algorithms to human-human conversation classification » 13th Annual


Conference of the International Speech Communication Association ISCA
Interspeech 2012, Portland, USA


Related work


[Damnati11]  Robust speaker turn role labeling of TV Broadcast News shows, Géraldine Damnati, Delphine Charlet , actes de la conference IEEE ICASSP 2011: 5684-5687


