Jan Čech
Center for Machine Perception
Department of Cybernetics, Faculty of Electrical Engineering
Czech Technical University in Prague

Bachelor's Thesis Review

Prague, January 13, 2015

Title: Audio-visual Speech Activity Detector
Author: Hana Šarbortová
Date received: January 6, 2015

The thesis presents an algorithm for detecting an active speaker in a video sequence, i.e. a person who is both heard in an audio track and visible in the camera image. The detection is achieved by recognition of synchrony between the subject's lip motion and the audio signal. The proposed method is based on the Canonical Correlation Analysis (CCA), which is a common tool for cross-modal analysis.

The thesis nicely reviews the state-of-the-art approaches, describes the CCA-based method in details together with extraction of multi-dimensional features from the lip motion detection and the audio signal. The experiments demonstrate basic properties of the proposed synchrony detection and show two applications, the audio-visual speaker detector and audio-video delay estimation as a side effect of the work. The results indicate that a reliable synchrony recognition (less than 5% error rate) is for a window of length at least 8 seconds. Assuming the subject's lips are tracked by a third-party facial landmark detector, and the speech is natural, the accuracy does not depend very much on a particular speaker, on a content of the speech or on a language.

The topic of the thesis is not trivial. The solution was not straightforward and required some research. Hana Šarbortová proved an ability to use her knowledge, understanding and to practically apply various computer vision, auditory and signal processing methods. She successfully integrated the extracted normalized features into the CCA framework and tested the proposed method experimentally on ground-truth data with promising results.

The presentation of the thesis is fluent, written in a good level of English language. In about 40 pages, the thesis provides a sufficient level of details. Besides many references to related papers, the thesis tries to be maximally self-contained, e.g. by including a brief description of both the MFCC features extracted from the audio signal and the landmark detector used in tracking the lip motion.

The candidate worked on the thesis continuously and systematically. We consulted almost every week regularly. My guidance was rather tight, but the candidate had an initiative and she came up with and implemented several own ideas. She is capable of doing a research, an engineering work, and at the same time she is careful in reporting.

My only objection could be that the progress was a bit slower than I wished originally. My original plan was to implement the CCA-based method first as a baseline to compare with a new more sophisticated algorithm that would use machine learning techniques. That was not achieved. Nevertheless considering that the candidate started working on the topic from scratch, less than one semester ago, the outcome of the bachelor's thesis is solid.

I suggest evaluating the thesis as

A – excellent.

Ing. Jan Čech, Ph.D.
Thesis Advisor