CCIT REPORT #726 April 2009

Onsets Coincidence for Cross-Modal Analysis

Zohar Barzelay and Yoav Y. Schechner

Abstract

Cross-modal analysis offers information beyond one extracted from individual modalities. Consider a non-trivial scene, that includes several moving visual objects, of which some emit sounds. The scene is sensed by a camcorder having a *single microphone*. A task for audio-visual analysis is to assess the number of independent audio-associated visual objects (AVOs), pinpoint the AVOs' spatial locations in the video and isolate each corresponding audio component. We describe an approach that helps handling this challenge. The approach does not inspect the low-level data. Rather, it acknowledges the importance of mid-level features in each modality, which are based on significant temporal changes in each modality. A probabilistic formalism identifies temporal coincidences between these features, yielding cross-modal association and visual localization. This association is further utilized in order to isolate sounds that correspond to each of the localized visual features. This is of particular benefit in harmonic sounds, as it enables subsequent isolation of each audio source. We demonstrate this approach in challenging experiments. In these experiments, multiple objects move simultaneously, creating motion distractions for one another, and produce simultaneous sounds which mix.

I. Introduction

Cross modal analysis draws a growing interest both in the computer-vision and in the signal-processing communities. Such analysis aims to deal with scenarios in which the available data is multi-modal by nature. Consequently, co-processing of different modalities is expected to synergize tasks that are traditionally faced separately. Moreover: such co-processing enables new tasks, which cannot be accomplished in a single-modal context, e.g. visually localizing an object producing sound. Indeed, audio-visual analysis [1]-[3] has seen a growing expansion of research directions, including lip-reading [4], [5], tracking [6], and spatial localization [7]-[10]. This also follows evidence of audio-visual cross-modal processing in biology [11].

Z. Barzelay and Y. Y. Schechner are with the Department of Electrical Engineering, Technion-Israel Inst. Technology, Haifa 32000, Israel (e-mail: zohar.barzelay@gmail.com; yoav@ee.technion.ac.il)