CCIT Report #620 April 2007

# Harmony in Motion

Zohar Barzelay and Yoav Y. Schechner
Department of Electrical Engineering
Technion - Israel Inst. Technology
Haifa 32000, ISRAEL
zoharb@tx.technion.ac.il, yoav@ee.technion.ac.il

## Abstract

*Cross-modal analysis offers information beyond that extracted from individual modalities. Consider a camcorder having a* single microphone *in a cocktail-party: it captures several moving visual objects which emit sounds. A task for audio-visual analysis is to identify the number of independent audio-associated visual objects (AVOs), pinpoint the AVOs' spatial locations in the video and isolate each corresponding audio component. Part of these problems were considered by prior studies, which were limited to simple cases, e.g., a single AVO or stationary sounds. We describe an approach that seeks to overcome these challenges. It acknowledges the importance of temporal features that are based on significant changes in each modality. A probabilistic formalism identifies temporal coincidences between these features, yielding cross-modal association and visual localization. This association is of particular benefit in harmonic sounds, as it enables subsequent isolation of each audio source. We demonstrate this in challenging experiments, having multiple, simultaneous highly nonstationary AVOs.*

## 1. Cross-Modal Analysis

Cross modal analysis is gaining interest in computer vision. Such analysis seeks associations between sources of input data, which have very different natures. Examples of this include registration of images acquired using sensors of different kinds [15], or association of images to text [12], such as in web pages and multimedia subtitles. It also includes audio-visual analysis [23, 25, 29], which has seen a growing expansion of research directions, including lip-reading [7, 13], tracking [24], and spatial localization [6, 9, 17, 18, 22]. This follows evidence of audio-visual cross-modal processing in biology [11].

This work deals with complex scenarios that are sometimes referred to in the literature as a *cocktail party* [9, 13, 26]: multiple sources exist simultaneously in all modalities.

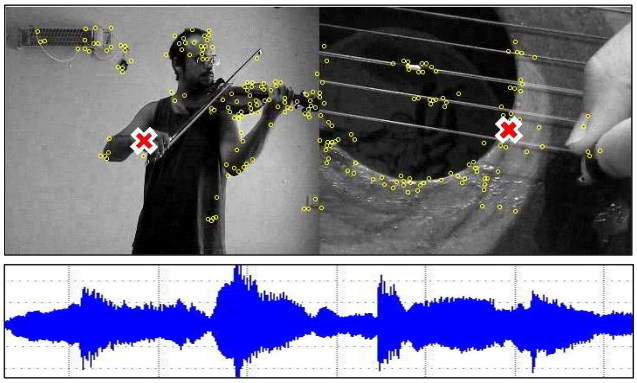

Figure 1. A frame and the audio from the `violin-guitar` movie. A camcorder and a single microphone were used. Two movies were compounded and then processed as a whole. Out of the selected and tracked visual features [Dots], two are automatically associated to the audio [Crosses]: correctly, one per source. The audio mixture is also decoupled to a guitar and a violin. See/hear this via *www.ee.technion.ac.il/∼yoav/research/harmony-in-motion.html*

This inhibits the interpretation of each source. In the domain of audio-visual analysis, a camera views multiple independent objects which move simultaneously, while some of them emanate sounds, which mix. This is depicted in Fig. 1. This paper presents a computer vision approach for dealing with this scenario. The approach has several notable results. First, it automatically *identifies the number of independent sources*. Second, it tracks in the video the multiple *spatial features*, that move in synchrony with each of the (still mixed) sound sources. This is done even in highly non stationary sequences. Third, aided by the video data, it successfully *separates the audio* sources, even though only a *single microphone* is used. This completes the isolation of each contributor in this complex audio-visual scene.

Some of the prior methods considered only parts of these tasks. Others relied on complex audio-visual hardware, such as an array of microphones that are calibrated mu-