NSF Grant Number: IIS-0205507

PIs: James Rehg, Irfan Essa, Georgia Institute of Technology (GVU Center / College of Computing)

Title: "Analysis of Complex Audio-Visual Events Using Spatially Distributed Sensors"

Research Objectives

Joint analysis of audio-visual signals to deal with
(a) arbitrary number of cameras and microphones
(b) multiple speech and nonspeech sound sources, and
(c) multiple moving people and objects,
(d) varying signal quality/fidelity.

Approach

Joint analysis at Signal and Spatial Levels

1. Representations and learning methods for signal level fusion.
2. Volumetric techniques for fusing spatially distributed audio-visual data.
3. Self-calibration of distributed microphone-camera
4. Use of visual processing method to audio and vice-verso to aid in analysis.
5. Applications of audio-visual sensing. e.g., lip and facial analysis to improve voice communications, determining focus of attention.

Broader Impact

1. Multidisciplinary research that crosses engineering and sciences in the specific fields of computer vision, acoustics, and signal processing.
2. Fusion of information from a large array of microphones and cameras allows for natural interfaces, leading to machines capable of speech reading, ubiquitous and intelligent (off-the-desktop) interfaces.
3. Supports Georgia Tech's Aware Home's research goal of building technologies to support the elderly maintain an independent and healthy lifestyle in their home.

Results

Speechreading

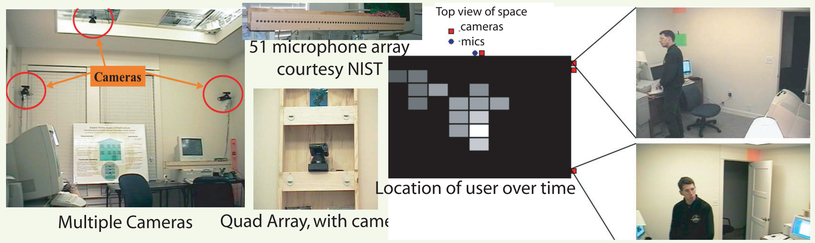
A basic limitation of visual lip motion reading is the lack of discriminating features available for measurement from video, like the Mel Frequency Cepstral Coefficients (MFCCs) used with hidden Markov Models (HMMs) for Speech Recognition. We have developed an Asymmetric Boosting Algorithm for Feature Selection, to extract such features from Visual Data.

Yin, Essa, Rehg, "Asymmetrically Boosted HMM for Speech Reading". *IEEE CVPR 2004.*

Yin, Essa, Rehg, "Boosted Audio-Visual HMM for Speech Reading". In *Proc. IEEE International Workshop on Analysis and Modeling of Faces and Gestures (AMFG) 2003 with ICCV 2003.*

Audio Visual Tracking

We have developed various different forms of phased-array microphones and multiple camera systems to robustly track a person in complex environments (home, office, etc.).



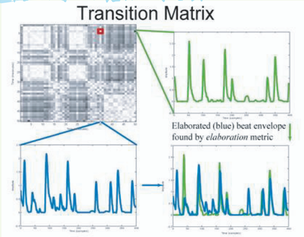
Stillman, Essa, "Towards reliable multi-modal sensing in Aware Environments", PUI 2002.(With UIST 2002).

Garg, Pavlovic, Rehg, "Boosted Learning in Dynamic Bayesian Networks for Multimodal Speaker Detection", *Proceedings of the IEEE*, 2004.

Rhythmic Similarity

We develop this approach to evaluate how close to identical two rhythms are. We proposes a similarity metric based on rhythmic elaboration that matches rhythms that share the same beats regardless of tempo or identicalness.

Parry, Essa "Rhythmic Similarity through Elaboration", ISMIR 2003.



For further information, see http://www.cc.gatech.edu/cpl/projects/aCAVE/