Multi-Sensory Perceptive Systems: Human and Machine Processing of Multi-Modal Data

John Fisher (MIT AI Lab) Ladan Shams (Caltech)
Virginia De Sa (UCSD) Malcolm Slaney (IBM Almaden)
Trevor Darrell (MIT AI Lab)

July 30, 2003

1 Description

All perception is multi-sensory perception. Situations where animals are exposed to information from a single modality exist only in experimental settings in the laboratory. For a variety of reasons, research on perception has focused on processing within one sensory modality. Consequently, the state of knowledge about multi-sensory fusion in mammals is largely at the level of phenomenology, and the underlying mechanisms and principles are poorly understood. Recently, however, there has been a surge of interest in this topic, and this field is emerging as one of fast growing areas of research in perception.

Simultaneously and with the advent of low-cost, low-power multi-media sensors there has been renewed interest in automated multi-modal data processing. Whether it be in an intelligent room environment, heterogenous sensor array or the autonomous robot, robust integrated processing of multiple modalities has the potential to solve perception problems more efficiently by leveraging complementary sensor information.

The goals of this workshop are to further the understanding of the both the cognitive mechanisms by which humans (and other animals) integrate multimodal data as well as the means by which automated systems may similarly function. It is not our contention that one should follow the other. It is our contention, that researchers in these different communities stand to gain much through interaction with each other. This workshop aims to bring these researchers together to compare methods and performance and to develop a common understanding of the underlying principles which might be used to analyze both human and machine perception of multi-modal data. Discussions and presentations will span theory, application, as well as relevant aspects of animal/machine perception.

The workshop will emphasize a moderated discussion format with short presentations prefacing each of the discussions. The presentations and discussions would be organized around the following related questions:

Automated Systems

- How should one model joint audio video properties (e.g. statistically)?
- At what level (and why) should one fuse joint audio/video measurements (e.g. at the signal, feature, or decision level)?
- How should one deal with high dimensionality?
- How are fusion methods impacted by co-located sensors (human eyes and ears) vs. distributed sensors (intelligent rooms)?
- Do the joint statistical models being used predict (or explicitly model) any of the known psychophysical phenomenon (e.g. McGurk effect, ventriloquism effect, head reflectance transfer function)?

Human (Animal) Perception

- What is a good phenomenological general characterization of the direction of crossmodal interactions? I.e., what factors determine the dominance of one modality over the others in a given situation?
- At what level crossmodal interactions occur? Early sensory, late sensory, association sensory, cognitive, all of the above?
- Why have multisensory integration? Enhancing perceptual resolution, perceptual learning, ecological validity, some other advantage?
- How can the gap between the neural data on polysensory neurons, and systems data like imaging, modeling, and psychophysical data be closed or narrowed? What kind of studies will be needed?
- What kind of theoretical framework seems most suitable for accounting for multisensory integration?

Format

Morning Session

Opening Talk

Applications using Audio-Video Fusion

- speakers will discuss current applications using joint audio/video measurements. In particular, they will be asked to highlight how and where they address issues which will be discussed later in the workshop.
- Three 15 minute talks with 5 minutes question/answer for each

Psychophysical Phenomenon (and Statistical Models)

- speakers will be asked to discuss known (or new) psychophysical phenomena associated with human (animal) perception of acoustic/visual stimuli.
- Three 15 minute talks with 5 minutes question/answer for each

Moderated Discussion (45 minutes)

Afternoon Session

Statistical Methods of Fusion

- speakers will discuss methods by which they model the joint statistics of audio and video. In particular they will be asked to identify which (if any) known psychophysical models are predicted/exhibited by their approach.
- Three 15 minute talks with 5 minutes question/answer for each

Moderated Discussion (45 minutes)

Signal-level vs. Decision-level Fusion

- speakers will be asked to discuss the merits of low-level versus higher level fusion.
- Three 15 minute talks with 5 minutes question/answer for each

Moderated Discussion (30 minutes)

Moderated Wrap-up Discussion (45 Minutes)

Presentation Time: 3.5 hours
Questions/Moderated Discussion: 4 hours
Confirmed Participants

- Andrew Blake (Microsoft Research, Cambridge, UK)
- Ross Cutler (Microsoft Research, Seattle)
- Mike Harville (HP Research, Palo Alto)
- Irfan Essa (Georgia Tech)
- Kentaro Toyama (Microsoft Research, Seattle)
- Ramani Duraiswami (University of Maryland)
- John Hershey (UCSD)
- Dom Massaro

Potential Participants

- Chelpathy Neti
- Yehia Hania

Relevant Papers Published by Organizers:

John Fisher and Trevor Darrell

- John W. Fisher III, Martin Wainwright, Erik Suddherth, and Alan Willsky, "Statistical and Information-Theoretic Methods For Self-Organization And Fusion Of Multimodal, Networked Sensors", to be published in the International Journal of High Performance Computing Applications.
- John W. Fisher III, Trevor Darrell, William T. Freeman, and Paul. Viola. Learning joint statistical models for audio-visual fusion and segregation. In Advances in Neural Information Processing Systems 13, 2000.
- Trevor Darrell, John Fisher, Paul Viola, and Freeman Bill. Audio-visual segmentation and the cocktail party effect. In Proceedings of the International Conference on Multimodal Interfaces, Oct 2000.

Malcolm Slaney

- Malcolm Slaney. ?Pattern playback in the ?90s.? Invited Keybnote talk at the 1994 Neural Information Processing Society Meeting, Denver, CO, November 1994. Published in Advances in Neural Information Processing Systems 7, edited by Gerald Tesauro, David Touretzky, and Todd Leen. MIT Press, Cambridge, MA, pp. 827-834, 1995.
- Christoph Bregler, Michele Covell, and Malcolm Slaney. "Video Rewrite: Visual speech synthesis from video.? Proceedings of the 1997 ACM SIGGRAPH, Los Angeles, pp. 353-360, 1997.
- Malcolm Slaney, Michele Covell. "FaceSync: A linear operator for measuring synchronization of video facial images and audio tracks." NIPS 2000.