

Audio-Visual Synchrony for Face Location and Segmentation

Eric J. Mislivec

Faculty Sponsor: Christopher G. Prince

Introduction

In this project a computer program was created to calculate synchrony between audio and visual data signals, based on a mutual information calculation. The result of this processing, a two dimensional mutual information map, was used to estimate the location of a speaker, whose face was then segmented on the basis of colors near the estimated location. While generally applicable to sound sources and objects moving synchronously with audio, this method is particularly interesting for faces due to the strong time-based correlation between speech sounds and facial movements.

A mutual information based approach has certain attractive benefits over other methods of sound localization and facial detection. Binaural methods (i.e., two sound source methods) of estimating sound location, including cross-correlation between channels in a stereo audio signal, have the drawback of limited precision in spatial location and the requirement of stereo audio signals (Nandy & Ben-Arie, 1996). By calculating the mutual information at each (x,y) location in a visual image, the approach discussed here can potentially provide spatial resolution corresponding to that of the input video image. Facial detection methods tend to depend on implicit knowledge of faces (e.g., supervised training methods: Li, Zhu, Zhang & Zhang, 2002; Viola & Jones, 2001), or explicit knowledge of faces (e.g., in a pilot study, we implemented a template-based face recognition method based on Sinha, 1995). However, the focus here is on

segmenting faces from the background, which could be used as training input to detection or recognition methods.

The Software: *SenseStream*

SenseStream, the program described here, was implemented in C++, and has been run on RedHat 7.3 and Mandrake 9.2 versions of the GNU/Linux operating system. *SenseStream* provides a graphical interface allowing a user to control various aspects of the processing and view a display of its results.

Input data is obtained from digital video files (MPEG-1 format) specified by the user. Visual data read from the input file is adjusted to the format specified by the user if needed. Currently the program supports grayscale, RGB and YUV data as input to the mutual information processing. Input audio data is processed to a reduced set of user selected features at a rate corresponding to the visual frame rate (typically 30Hz). These audio features currently include a root mean square energy level and the zero-crossing rate. Frames of audio and visual data are processed asynchronously, temporally aligned, and then processed for mutual information.

The algorithm used for calculating audio-visual mutual information was adapted from an approach implemented by Hershey and Movellan (2000), who used the centroid of audio-visual mutual information data to successfully estimate a speaker's location. This algorithm operates on a set of data frames from two channels of information, extending over a window of recent time. Frames from one channel, here visual image data, consist of $h \times w \times m$ -element vectors. Frames from the other channel contain a single vector of n -elements, here processed audio features. Considering a set of audio ($a(t)$) and visual ($v(x,y,t)$) vectors sampled at times $t-s+1, \dots, t$ and

spatial location (x,y) as independent samples from a joint multivariate Gaussian process $(A(t), V(x,y,t))$, the mutual information between $A(t)$ and $V(x,y,t)$ is calculated as

$$I(A(t); V(x,y,t)) = \frac{1}{2} \log_2 \frac{|\Sigma_A(t)| |\Sigma_V(x,y,t)|}{|\Sigma_{A,V}(x,y,t)|}.$$

Where $\Sigma_A(t)$ and $\Sigma_V(x,y,t)$ are $n \times n$ and $m \times m$ covariance matrices calculated from the audio and visual data, respectively, and $\Sigma_{A,V}(x,y,t)$ is a $(n+m) \times (n+m)$ joint covariance matrix between the combined audio and visual data.

This mutual information calculation is performed at each (x,y) coordinate within a visual frame, resulting in a two-dimensional mutual information map of dimension $h \times w$ corresponding to that of the input visual data. Each element within this mutual information map has been termed a *mixel* (mutual information pixel), and the map itself a *mixelgram*. The resulting mixel data is thresholded, with all values less than one standard deviation above the mean of all mixels in a frame being zeroed. The thresholded data is then smoothed with a neighbor averaging method, and a mean based M-estimator of location (Goodall, 1983) is used to calculate a weighted centroid of this smoothed data.

Interpreting this centroid as an estimate of the location of the audio source within the visual field, pixels in the input visual image near the estimated location are sampled to produce a set of minimum and maximum YUV component values. Visual segmentation is then achieved using the color segmentation mechanism of the CMVision image-processing package (Bruce, Balch & Veloso, 2000), with the sampled range of YUV values as thresholds. A mask indicating the largest contiguous region of pixels within the threshold range is obtained from the CMVision package for subsequent processing and display.

Evaluation and Report

To evaluate the function of the facial segmentation processing, the software was run on several video files of people speaking. The performance of the segmentation method was most effective in situations where the person speaking was the only significant actor (source of sound and motion). In these cases, *SenseStream* is able to perform reasonable segmentations of a speaker's face for portions of the data when the speaker is actually speaking. As there is currently no adjustment to the processing for portions of the data in which there is no speech, the facial segmentation produced unreliable results for these intervals. Also, in situations with multiple concurrent sound sources, the centroid estimate of location, and hence the facial segmentation, is largely ineffective.

Testing on various deviant datasets indicated limitations of an approach based on a Gaussian mutual information calculation. Using data in which the audio and video had been artificially placed out of sync (by delaying the audio by a second, for example) produced mixelgrams comparable to those for unaltered data. This can be explained by the fact that the mutual information calculation is based on the assumption of a Gaussian distribution of the input data, a rough approximation at best. The mutual information calculation is also sensitive to very low-level variations in the input signals over time. However, the general trend is higher mutual information, and a more ordered mixelgram, in situations where there is audio-visual synchrony.

SenseStream has also been used for purposes other than speaker localization and facial segmentation. In an M.S. thesis Vuppla (2004) evaluated the hypothesis that perceptually relevant mixelgrams indicate a strong level of audio-visual synchrony. In a recently submitted paper (Prince, Hollich, Helder, Mislivec, Reddy, Salunke, & Memon, submitted), estimates of audio-visual synchrony based on measures of whole mixelgrams were compared to infants'

responses to the same stimuli. The results of this work indicate that the relatively simple estimates of synchrony used provide a reasonable approximation of the level of synchrony perceived by infants. Similar measurements of synchrony were used as an attention selection mechanism in a sensory augmented model of infant word-learning, as reported in detail in a paper submitted to the journal *Developmental Science* (Prince & Mislivec, submitted). In this work, high levels of synchrony were used to trigger learning in a connectionist word-learning model.

Final Notes

The current state of *SenseStream* implementation would not have been practicable without the support of the UROP program. The mutual information processing of the *SenseStream* program has provided the basis for further work and collaboration with researchers at other institutions, and in other disciplines (e.g., a grant proposal, co-authored with Lakshmi Gogate, and PI Chris Prince was recently submitted). In addition to the experience gained planning and implementing the software, the ability to work with others on related projects has been invaluable. Through the course of this project, I have also become more acquainted with current work making use of cross-modal (e.g., audio-visual) information. Issues relating to audio-visual processing have long been among my areas of interest and I plan to further pursue these areas in my education and profession. The requirements of the UROP program have also encouraged me to pursue this project in professional, academic manner, giving experience in the formal process of research. Overall, the UROP program has proven to be an enjoyable and highly rewarding experience.

Works Cited

- Bruce, J., Balch, T., & Veloso, M. (2000). Fast and inexpensive color image segmentation for interactive robots. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'00)*.
- Goodall, C. (1983). M-estimators of location: An outline of the theory. In D.C. Hoagfin F. Mosteller and J.W. Tukey, (eds.), *Understanding Robust and Exploratory Data Analysis* (pp. 339-403). New York: Wiley.
- Hershey, J., & Movellan, J. (2000). Audio-vision: Using audio-visual synchrony to locate sounds. In S. A. Solla, T. K. Leen, & K. R. Muller (eds.), *Advances in Neural Information Processing Systems 12* (pp. 813-819). Cambridge, MA: MIT Press.
- Li, S. Z., Zhu, L., Zhang, Z., & Zhang, H. (2002). Learning to detect multi-view faces in real-time. *The 2nd International Conference on Development and Learning (ICDL02)*. IEEE Press.
- Nandy, D., & Ben-Arie, J. (1996). Estimating the azimuth of a sound source from the binaural spectral amplitude. *IEEE Transactions on Speech and Audio Processing*, 4, No. 1, January 1996, 45-55.
- Prince, C. G., Hollich, G. J., Helder, N. A., Mislivec, E. J., Reddy, A., Salunke, S., & Memon, N. (submitted). Taking Synchrony Seriously: Comparing Infants With A Perceptual-Level Model. Paper submitted to *The Fourth International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*, to be held at Genoa, Italy, August 25-27, 2004.
- Prince, C. G., & Mislivec, E. J. (submitted). Investigating the Sensory Grounding Hypothesis: A Sensory-Augmented Connectionist Word-Learning Model. In submission to: *Developmental Science*.
- Sinha, P. (1995). *Perceiving and Recognizing Three-dimensional Forms*. Unpublished Ph.D. Dissertation, Massachusetts Institute of Technology, Cambridge, MA.
- Viola, P., & Jones, M. J. (2001). *Robust Real-Time Object Detection*. Technical report, Cambridge Research Laboratory, Compaq Computer Corporation.
- Vuppla, K. (2004). *Evaluation of Two Synchrony Detection Implementations*. Masters Thesis, Department of Computer Science, University of Minnesota Duluth.