

Detecting Audio-Visual Synchrony

Eric J. Mislivec

Department of Computer Science

Department of Electrical and Computer Engineering

misli001@d.umn.edu

Faculty Sponsor: Christopher G. Prince

Department of Computer Science

chris@cprince.com

Abstract

In this project a computer program was created to calculate synchrony between audio and visual data signals, based on a mutual information calculation (Hershey and Movellan (2000)). 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.

Advantages of a Single Sound Signal Approach

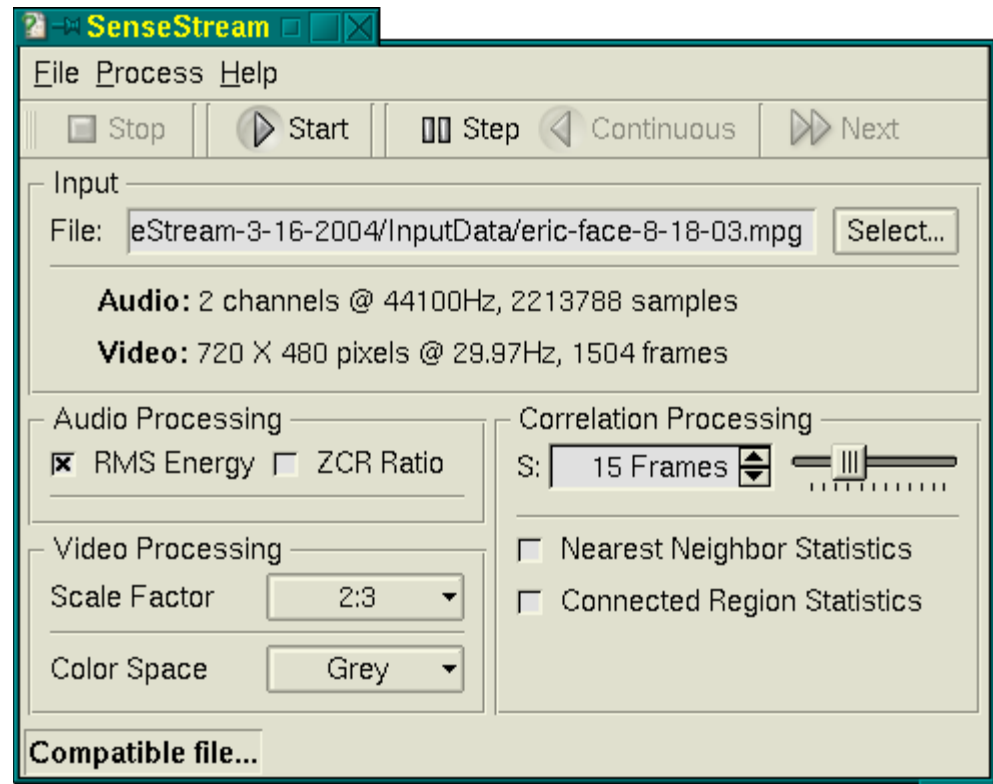
Binaural methods (i.e., two sound source methods) of estimating sound location, including cross-correlation between channels in a stereo audio signal, have the drawback of limited precision in spatial location and the requirement of stereo audio signals (Nandy & Ben-Arie, 1996). By calculating the mutual 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, without requiring stereo audio signals.

Advantages to Unsupervised Segmentation

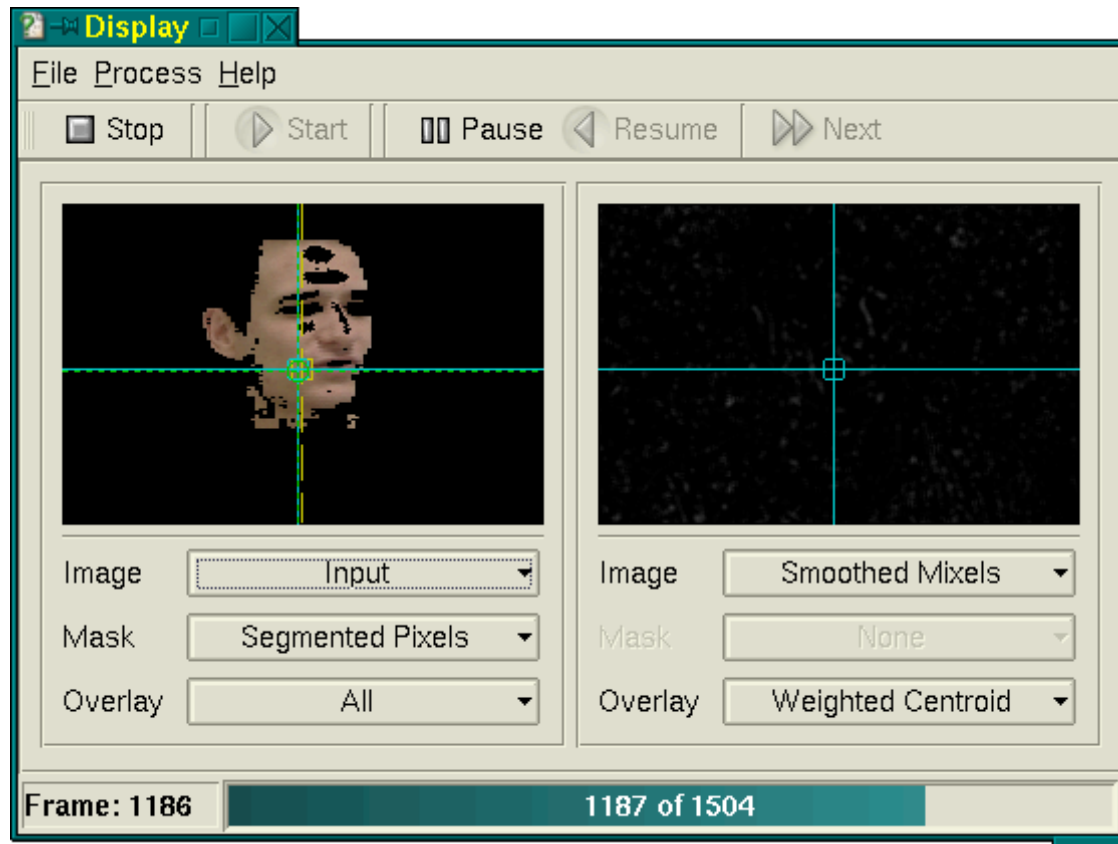
Facial detection methods tend to depend on implicit knowledge of faces (e.g., supervised training methods: Li, Zhu, Zhang & Zhang, 2002; Viola & Jones, 2001), or explicit knowledge of faces (e.g., in a pilot study, we implemented a template-based face recognition method based on Sinha, 1995). However, the focus here is on segmentation, which could be used as training input to supervised 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.



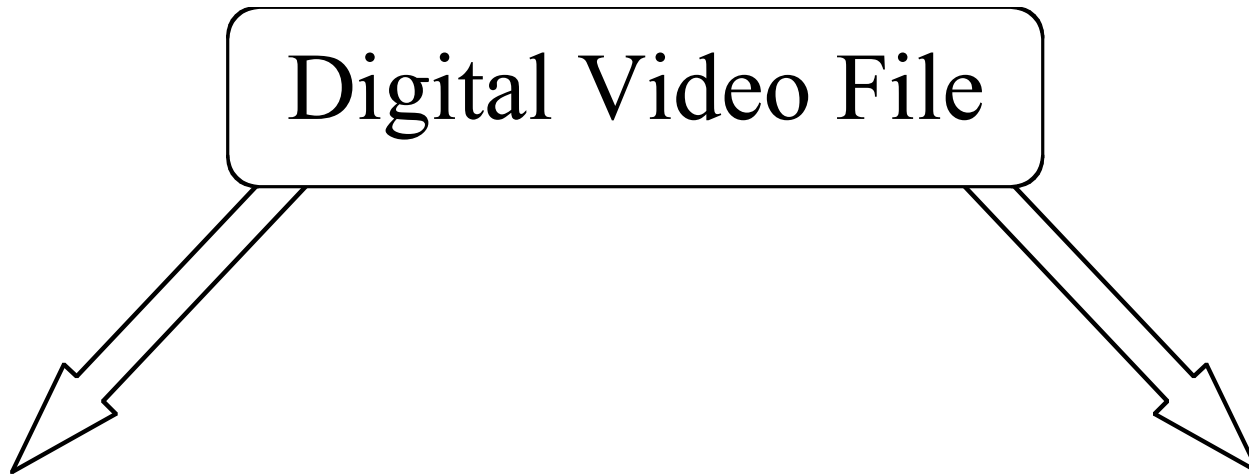
SenseStream configuration interface.



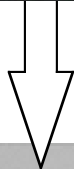
SenseStream processing display.

Input Processing

Input data is obtained from digital video files (MPEG-1 format).



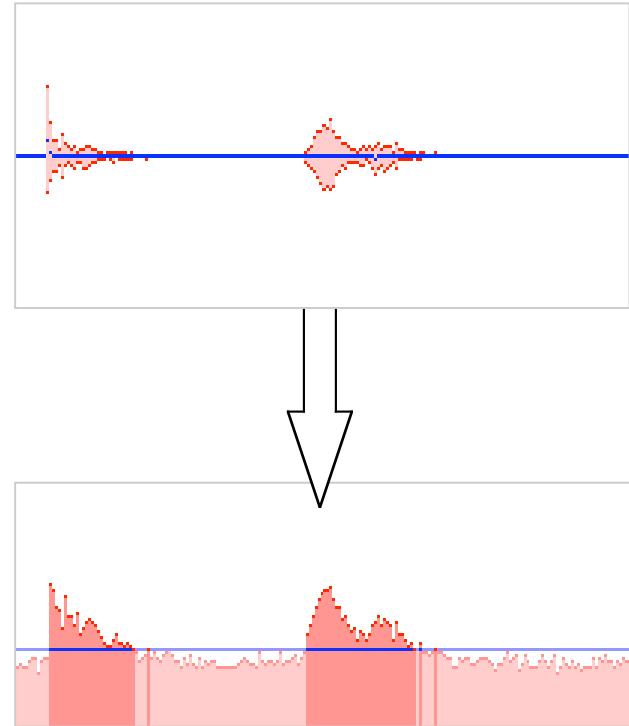
Visual Data

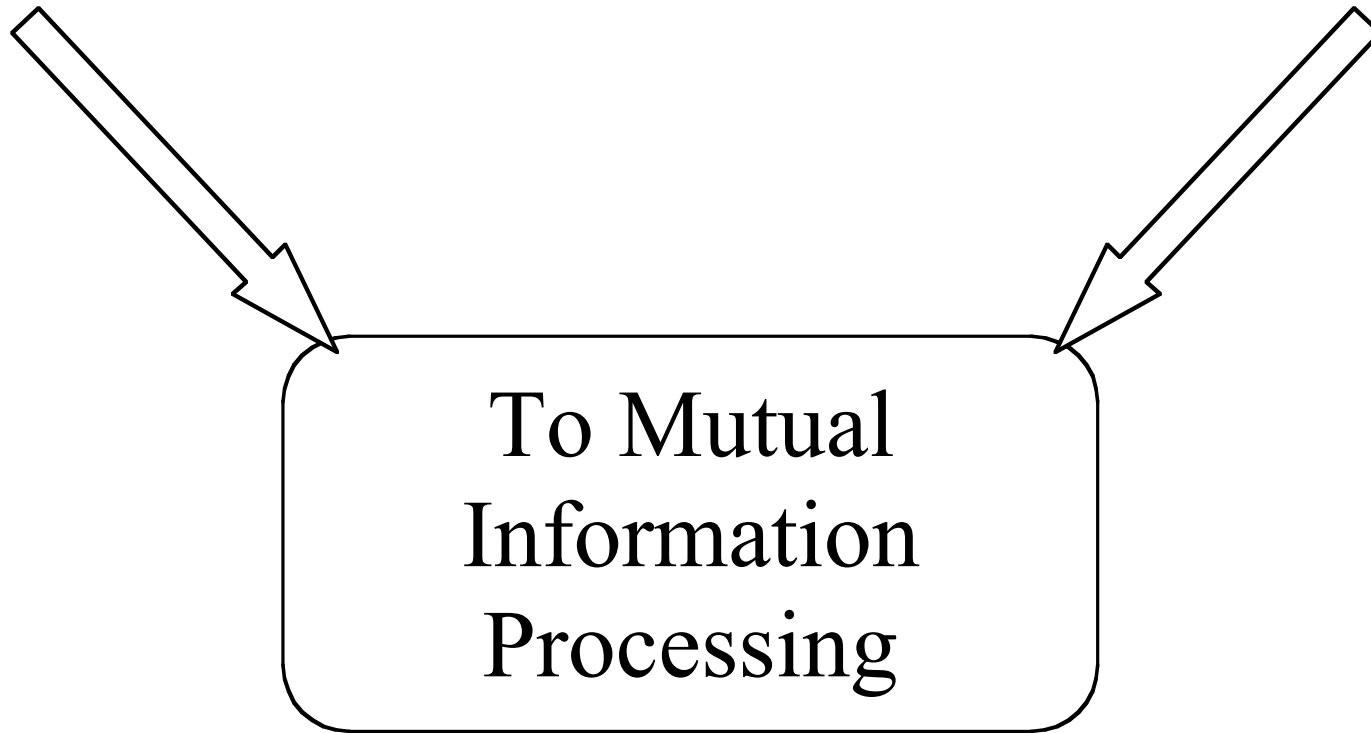


Visual data read from the input file is adjusted to grayscale, RGB or YUV format as specified by the user.

Audio Data

Input audio data is processed to a reduced set of features at a rate corresponding to the visual frame rate (typically 30Hz). These features currently include an RMS energy level and/or the zero-crossing rate.



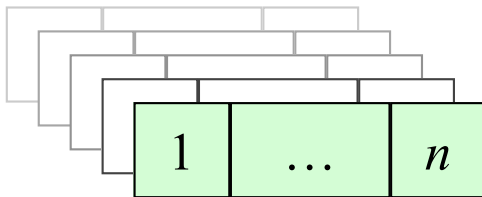


Frames of audio and visual data are processed asynchronously, temporally aligned, and then processed for mutual information.

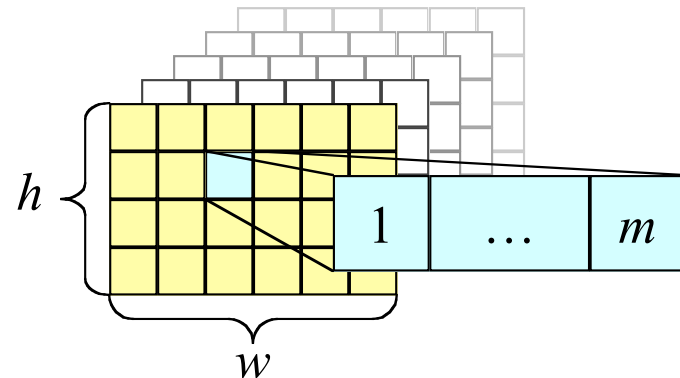
Mutual Information Calculation

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 contain a single vector of n -elements, here processed audio features.



Frames from another channel consist of $h \times w$ m -element vectors, here visual image data.



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.

	n			m		
}	1	...	n	$n+1$...	$n+m$

	1	...	n	$n+1$...	$n+m$

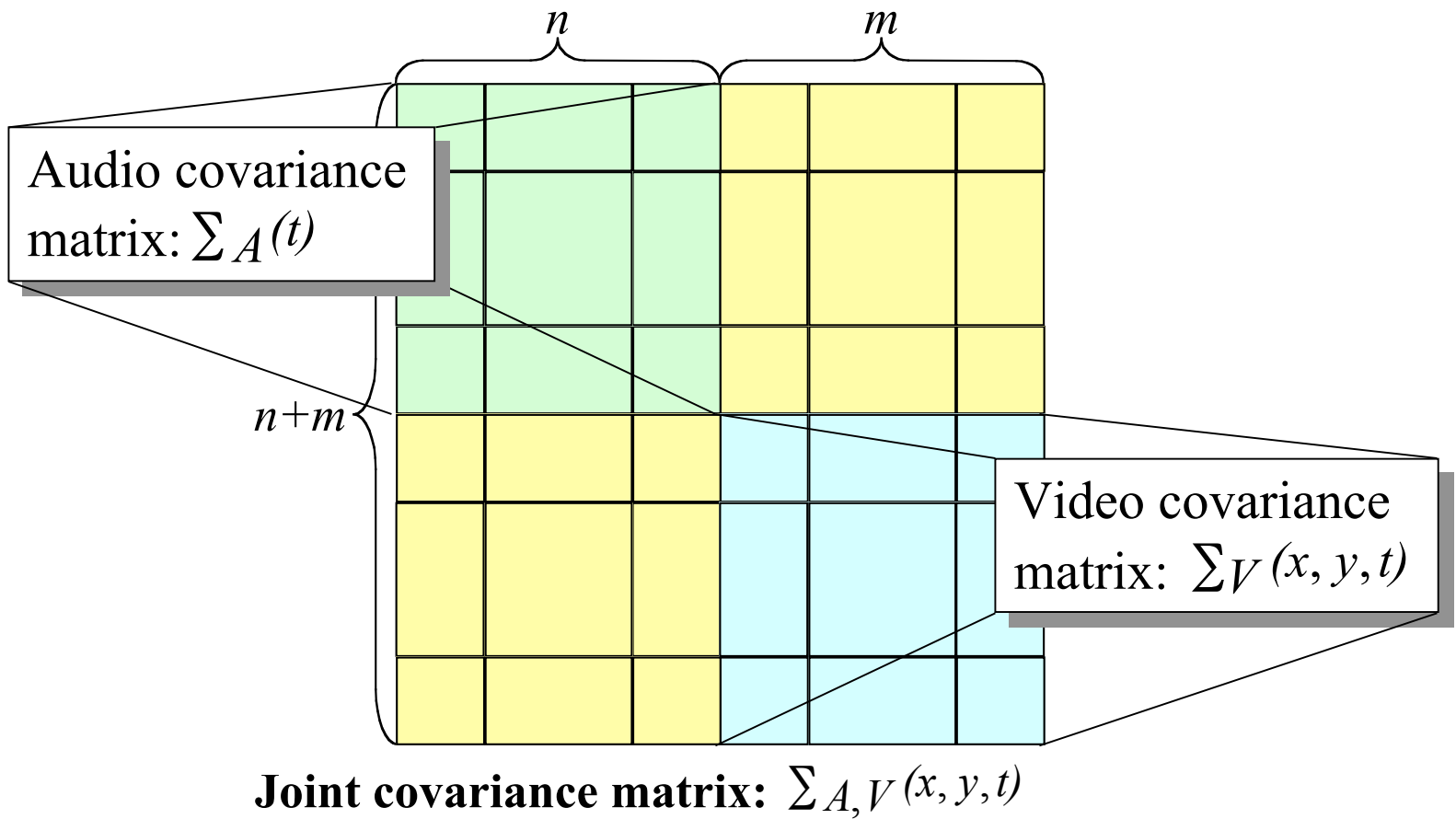
The calculation of these matrices is performed by considering the audio and visual data at some (x,y) location as one matrix of s rows and $n+m$ columns.

Combined audio-visual data matrix, D .

The i,j th element of the joint covariance matrix is then calculated as

$$\sum_{A,V} (x, y, t)_{i,j} = \frac{1}{s-1} \sum_{t=1}^s (d_{t,i} - \bar{d}_i)(d_{t,j} - \bar{d}_j)$$

Where \bar{d}_x is the mean of column x of D .

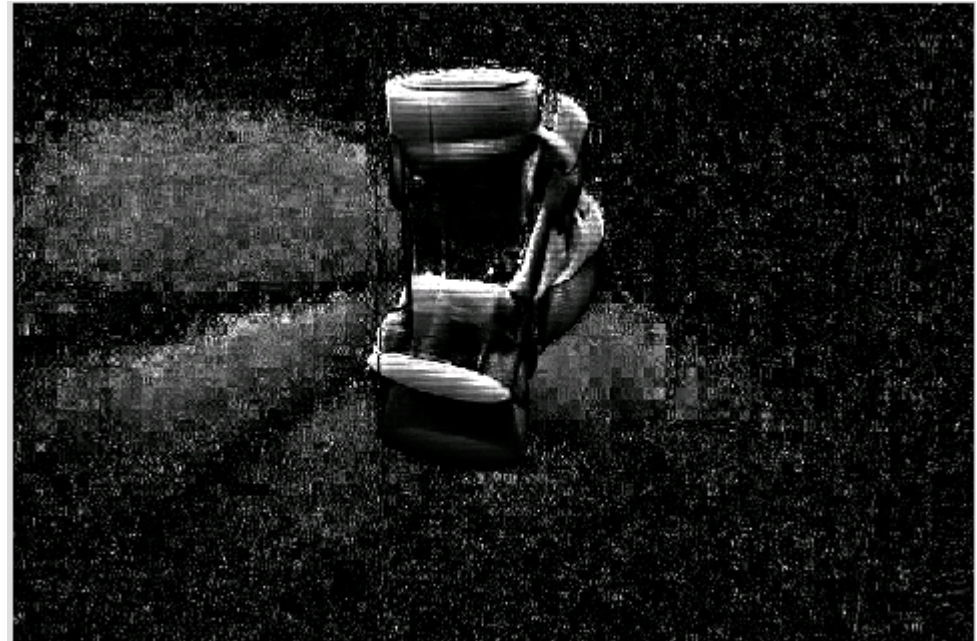


The mutual information is calculated at each (x,y) coordinate within a visual frame, resulting in a mutual information map of dimension $h \times w$ corresponding to that of the input visual data.

Each element within this mutual information map has been termed a *mixel* (mutual information pixel), and the map itself a *mixelgram*.



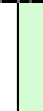
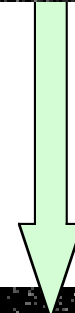
A mutual information image (mixelgram) of the author speaking.



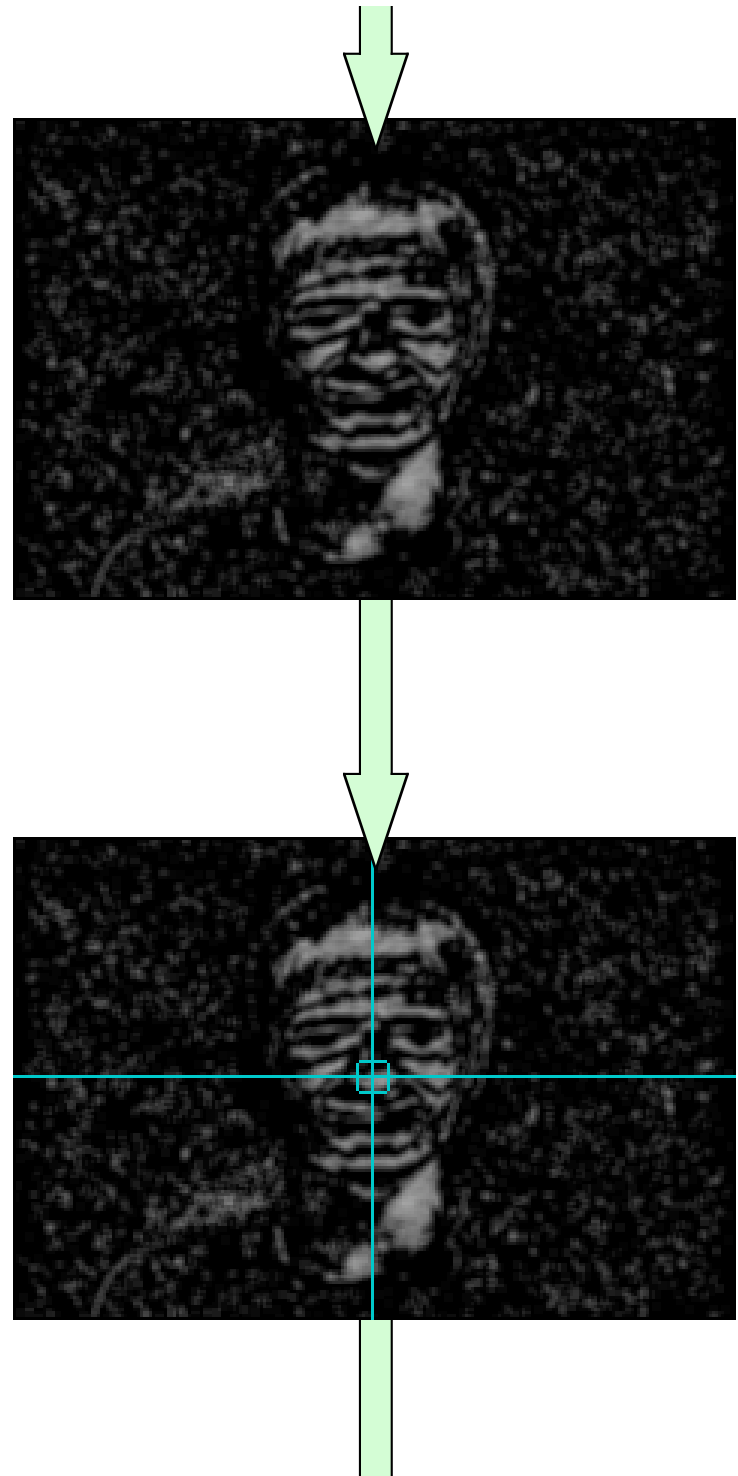
A mutual information image (mixelgram) of an object being moved in synchrony with a word utterance.

Segmentation

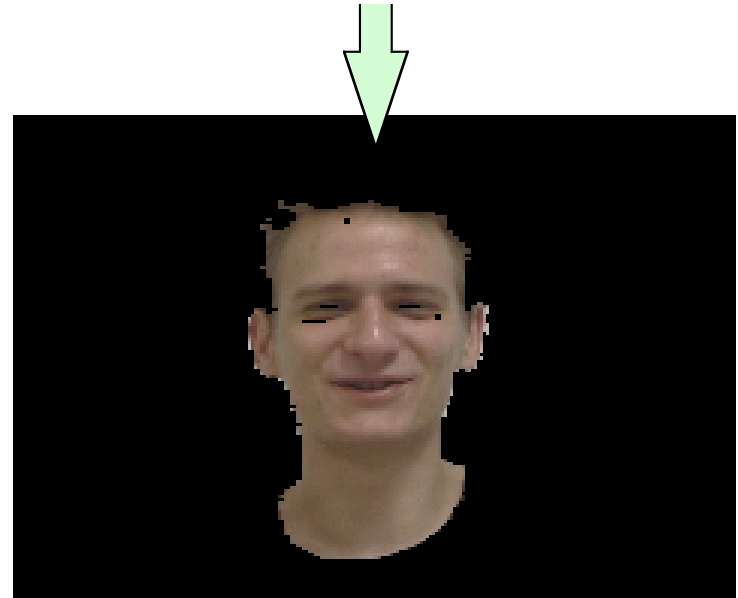
Mixel data from the mutual information processing is thresholded, with all values less than one standard deviation above the mean of all mixels in a frame being zeroed.



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 the centroid as an estimate of the location of the audio source within the visual field, pixels in the input image near the estimated location are sampled to produce a range of 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.



Evaluation

To evaluate the function of the facial segmentation, the software was run on several video files of people speaking.

- The segmentation method was most effective when the person speaking was the only significant actor (source of sound and motion).
- As no adjustment is made to processing for portions of the data in which there is no speech, the segmentation produced unreliable results for these intervals.
- With multiple concurrent sound sources, the centroid estimate of location, and hence the facial segmentation, is largely ineffective.

Testing on deviant datasets indicated limitations of a Gaussian mutual information approach.

- 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.
- The mutual information calculation is based on the assumption of a Gaussian distribution of the input data.
- Sensitive to very low-level variations in the input signals over time.
- General trend is higher mutual information, and a more ordered mixelgram, when there is audio-visual synchrony.

Related and Ongoing Research

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 accepted paper (Prince, Hollich, Helder, Mislivec, Reddy, Salunke, & Memon, accepted), 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.

Two related UROP projects currently in progress will make use of *SenseStream* for synchrony detection (Memon, in progress; Pollack, in progress). The goal of these projects is to have a pan-tilt robot discriminate between self-motion and motion in the world (“other” motion).

References

- 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.
- Memon, N. (in progress). *Detecting Environmental Synchrony Using a Robotic Camera: Hardware-Software Interfacing*. Undergraduate Research Opportunity Project, University of Minnesota Duluth.

- 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.
- Pollak, T. (in progress). *Detecting Environmental Synchrony Using a Robotic Camera: Hardware Development*. Undergraduate Research Opportunity Project, University of Minnesota Duluth.
- Prince, C. G., Hollich, G. J., Helder, N. A., Mislivec, E. J., Reddy, A., Salunke, S., & Memon, N. (accepted). Taking Synchrony Seriously: Comparing Infants With A Perceptual-Level Model. Paper accepted 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.

Further Information

More information on *SenseStream* and related projects can be found at:
<http://www.d.umn.edu/~cprince/Projects/KidCause/>

SenseStream itself is available online at:

<http://www.d.umn.edu/~cprince/projects/KidCause/SenseStream/>