

*AV-Detect Project:*  
Audio-Visual Synchrony Detection in  
Real-Time

Chris Prince,  
Nathan Helder,  
And the KidCause Team

University of Minnesota Duluth,  
Department of Computer Science

<http://www.cprince.com/PubRes/AV-Detect>

# Overview

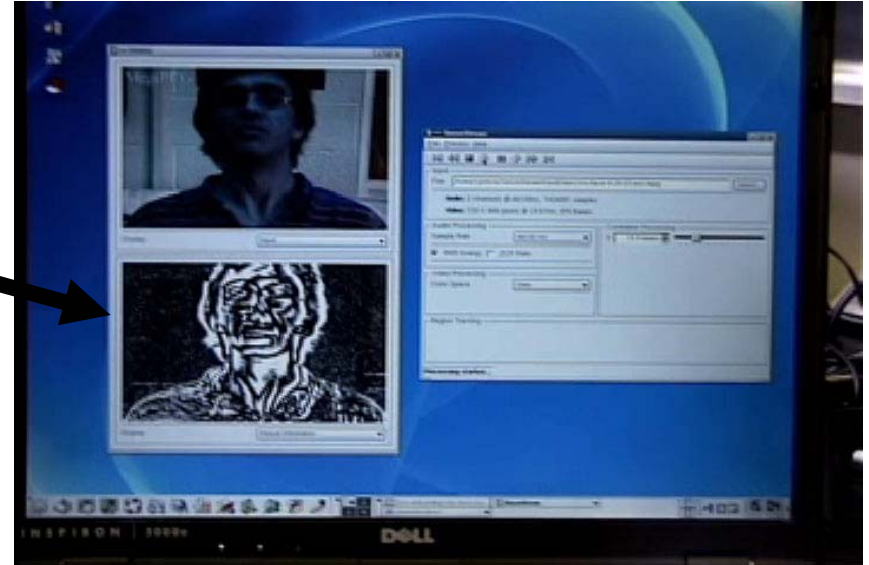
- Intro to Synchrony Detection
- Project Description Handout
- Demo of Detect Program
- Audio-Video Synchronization Issue
- Computing Mixelgrams
- Getting Started
- Advertising Plug
- Web Links & References

# Synchrony Detection - 1

- Motivation
  - We are building robotic models of young infants behavioral development
  - Young infants use audio-visual synchrony to help them learn
  - E.g., speech-object synchrony to help learn words (Gogate & Bahrick, 1998, 2001)
- We are building perceptual models of synchrony detection

# Synchrony Detection - 2

- Hershey and Movellan (2000) algorithm
  - Outputs *mixelgram* displays
  - Each pixel of the display is a *mixel*, a ***mutual information pixel***
  - Mixels computed from mutual information between two input channels



- SenseStream program: Mixels computed from mutual information between visual and audio (Mislivec, 2004)

***Perceptually relevant mixelgrams typically indicate synchrony between the two input channels*** (Vuppla, in preparation)

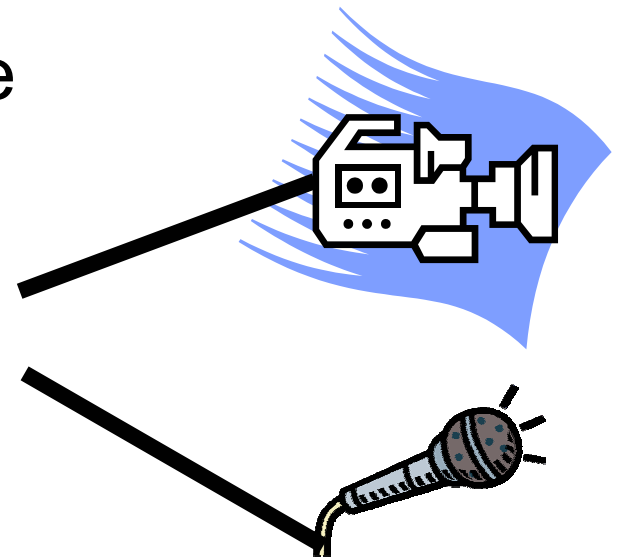
# Project Description Handout

# Demo of Detect Program

- Helder (2003)– Honors project at UMD, with extensions
- Detects synchrony between a video camera data stream, and descriptions of commands used to animate shapes

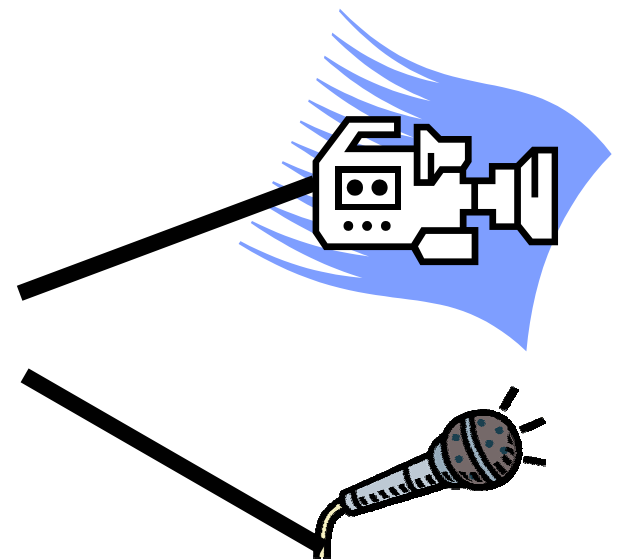
# Synchronization Issue - 1

- Consider writing a program to capture visual data and audio data from a camera and microphone, to save the results into a video data file
  - i.e., a home-brew digital camcorder
- Presumably, first need to acquire some data from devices
  - E.g., a visual frame and 1/15 of a second of audio



# Synchronization Issue - 2

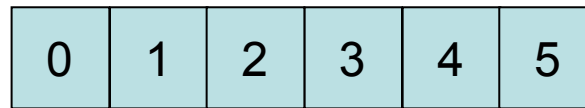
- Notice that the two devices are independent
  - With independent data sampling properties
- Separate streams of data from each device
  - Frames of visual data (size  $h*w$ ) from camera
    - @ say, 15 frames per second
  - Samples of audio from microphone
    - @ say, 44100 samples per second



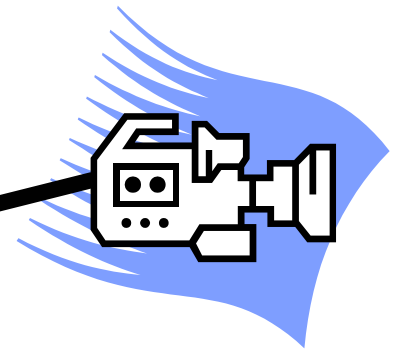
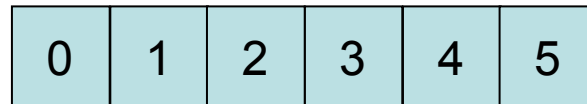
# Synchronization Issue - 3

- The streams of data will need to be “lined up” in time, i.e., synchronized
- Consider the first 6 visual frames  
6/15 of a second @ 15 frames per second

Visual frames (size h\*w)



Audio “frames” (length  
2940 samples; 44100  
samples per second)

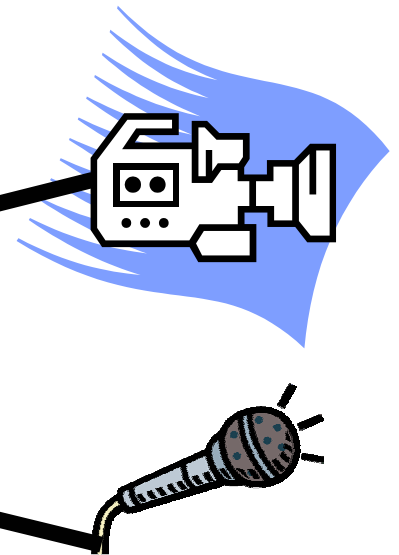
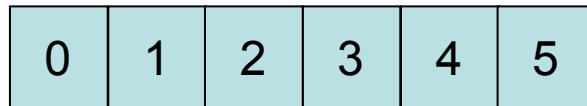
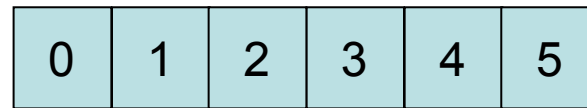
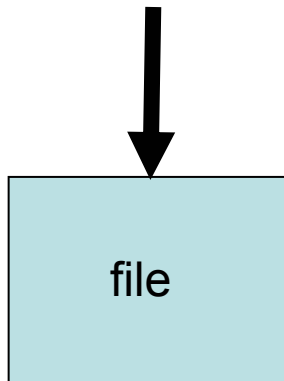


$$(44100/15 = 2940)$$

# Synchronization Issue - 4

- For an application writing visual and audio data to a file, you will likely need to interleave the visual frames and audio frames

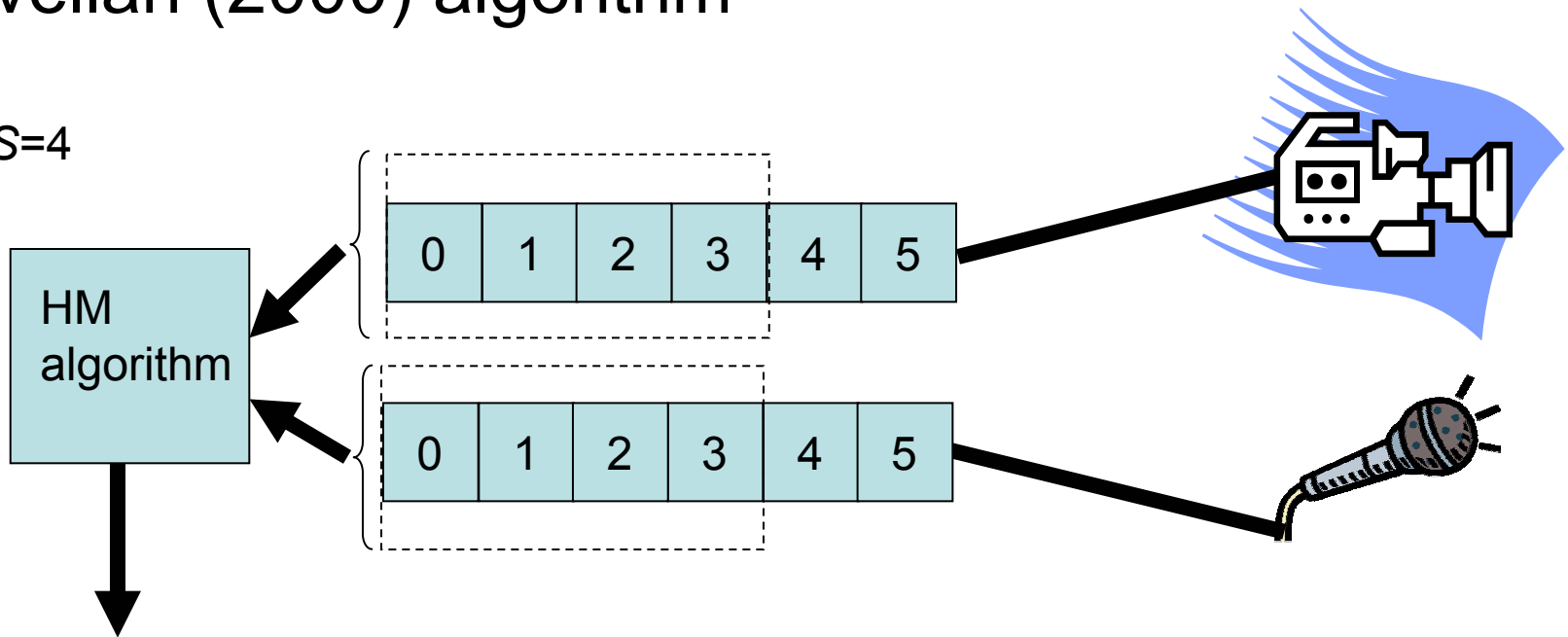
Write current  
video frame &  
current audio  
frame to file



# Synchronization Issue - 5

- In AV-Detect, a sequence of  $S$  visual and audio frames will be analyzed using the Hershey and Movellan (2000) algorithm

E.g.,  $S=4$



Mixelgram to computer display

# Computing Mixelgrams Includes:

- 1) Maintaining queues of audio and visual frames (each with length  $\geq S$ )
  - When a new audio or visual frame is obtained, and queue has already been filled, the frame at head is removed, and new frame inserted into queue
  - Compute Hershey and Movellan (2000) equation when queues have at least  $S$  audio and visual frames
- 2) Computing audio and visual features
  - RMS audio, RGB pixels
- 3) Computing the Hershey and Movellan (2000) equation (Equation 3 from their paper)

# RMS Audio Equations

- Let  $A_1, A_2, \dots, A_{NSamples}$  be raw audio (amplitude samples) obtained for one audio frame

$$NSamples = \frac{AudioSampleRate}{VideoFrameRate}$$

$$RMS = \sqrt{\frac{\sum_{i=1}^{NSamples} A_i^2}{NSamples}}$$

# Equations: Mixel Computation - 1

$$M(x, y) = \frac{1}{2} \log_2 \frac{|\sum A| |\sum V(x, y)|}{|\sum A, V(x, y)|}$$

# Equations: Mixel Computation - 2

$$M(x, y) = \frac{1}{2} \log_2 \frac{|\sum A| |\sum V(x, y)|}{|\sum A, V(x, y)|}$$

mixel

Covariance  
(matrix  
operation)

matrix  
determinant

# Equations: Mixel Computation - 3

$$M(x, y) = \frac{1}{2} \log_2 \frac{|\sum A| |\sum V(x, y)|}{|\sum A, V(x, y)|}$$

- $M(x, y)$ 
  - mixel computed from “column” of pixels located at position  $(x, y)$  on current  $S$  frames of visual data, and corresponding audio data
- $x$  ranges over  $h$  (height)
- $y$  ranges over  $w$  (width)

# Terms - 1

- $A$ :  $S \times n$  matrix ( $S$  rows,  $n$  columns)
  - Summarized (e.g., RMS) audio data corresponding to most recent  $S$  frames of visual data
  - Each row represents audio that corresponds to one visual frame
  - When using RMS audio,  $n=1$ 
    - i.e.,  $A$  is an  $S \times 1$  row vector

# Terms - 2

- $V(x,y)$ :  $S \times m$  matrix ( $S$  rows,  $m$  columns)
  - Visual data for pixel  $(x, y)$  from most recent  $S$  visual data frames
  - Each row represents one pixel
    - E.g., for RGB,  $m$  can be 3,  $V(x, y)$  would then be a  $S \times 3$  matrix
- $A, V(x,y)$ :  $S \times (n+m)$  matrix ( $S$  rows,  $n+m$  columns)
  - Rows have
    - data from  $A$  row ( $n$  elements)
    - then data from  $V(x,y)$  row ( $m$  elements).
  - Hence each row has  $n+m$  elements

# Terms - 3

- $\Sigma A$  is covariance of matrix  $A$ 
  - dimensionality of  $\Sigma A$  is  $n \times n$
- $\Sigma V(x,y)$  is covariance of matrix  $V(x,y)$ 
  - dimensionality of  $\Sigma V(x,y)$  is  $m \times m$
- $\Sigma A, V(x,y)$  is covariance of matrix  $A, V(x,y)$ 
  - dimensionality of  $\Sigma A, V(x,y)$  is  $(n+m) \times (n+m)$

# Getting Started

- Hardware
  - You will need a USB webcam
  - Need Windows 2000 or XP system, with USB device
    - How many students have one, personally?
- Software
  - Webcam drivers
  - Java installations
- Group organization
  - 1, 2, 3, 4 structure of Formal Requirements may be a good plan; correspond to four team members

# Advertising Plug

- Partially funded by the UROP program, and



- Join the KidCause Team!
  - Next lab meeting this week Thurs, 5:15pm, conference room, CS Dept. Heller Hall

# References - 1

- Gogate, L. J. & Bahrick, L. E. (1998). Intersensory redundancy facilitates learning of arbitrary relations between vowel sounds and objects in seven-month-old infants. *Journal of Experimental Child Psychology*, 69, 13-149.
- Gogate, L. J. & Bahrick, L. E. (2001). Intersensory redundancy and 7-month-old infants' memory for arbitrary syllable-object relations. *Infancy*, 2, 219-231.
- Helder, N. A. (2003). *A real-time, computational model of perceptually-based contingent behavior detection*. Honors project, University of Minnesota Duluth, Department of Computer Science. Internet: <http://www.cprince.com/projects/KidCause/Detect/>

# References - 2

- Hershey, J. & Movellan, J. (2000). Audio-vision: Using audio-visual synchrony to locate sounds. In S. A. Solla, T. K. Leen, & K. -R. Müller (Eds.), *Advances in Neural Information Processing Systems 12* (pp. 813-819). Cambridge, MA: MIT Press.  
Internet:  
<http://www.cprince.com/Projects/KidCause/contingency/AudioVision.pdf>  
<http://www.cprince.com/Projects/KidCause/contingency/HersheyAndMovellan.rm>
- Mislivec, E. J. (2004). *Audio-visual synchrony for face location and segmentation*. Undergraduate research opportunity project, University of Minnesota Duluth.  
Internet: <http://www.cprince.com/PubRes/SenseStream>
- Vuppla, K. (in preparation). *Evaluation of Two Synchrony Detection Implementations*. Masters Thesis, University of Minnesota Duluth, Computer Science Department.  
Internet: <http://www.cprince.com/PubRes/VupplaThesis04>