Automatic Recognition of Hand Raising Gestures & Voice Requests for a Remote Learning Application

Bill Kapralos[1,3], Alexander Barth[2], Jacky Ma[1], Michael Jenkin[1,3]

{billk, jenkin}@cs.yorku.ca

[1]Department of Computer Science, York University, Toronto Ontario, Canada M3J 1P3
[2]Dept. Of Computer Science, Bonn-Rhein-Sieg University, St. Augustin, Germany
[3]Centre for Vision Research, York University, Toronto Ontario, Canada M3J 1P3

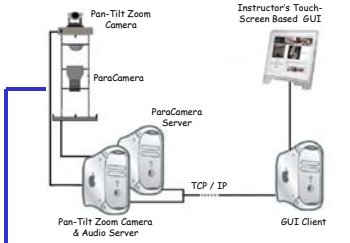
Introduction

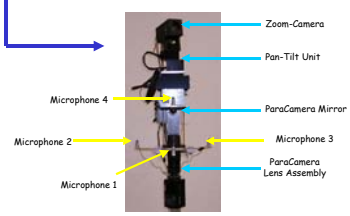
- **Synchronous Distance Learning (SDL):**
  - Permits for live discussion & immediate feedback
  - Can provide expert instructors to a geographically dispersed set of students
  - Enables new educational opportunities which were only a dream a few years ago!
- **Central Issue in Developing an SDL System:**
  - Enabling interaction between instructor & students of remote classes:
    - How do students signal their intent to interact with the instructor?
      - → Hand raising, speaking aloud
    - How does instructor select & attend to a student?

Project Goals

- **Develop an SDL System Integrating Audio & Video Cues:**
  - In a multiple student setting, automatically attend to a student wishing to interact with the instructor:
    - Students may speak or raise their hand to attract instructor's attention
    - Permit for dialogue between students & instructor as in "normal" classroom setting

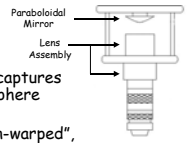
System Architecture





Combined Audio & Video Sensor

Video System



- **Cyclovision's ParaCamera:**
  - Omni-directional video sensor captures view of the entire visual hemisphere from a single viewpoint
  - Hemispherical view is easily "un-warped", allowing for multiple dynamic views of the scene
- **Pan-Tilt Mounted Zoom Camera:**
  - Low Resolution ParaCamera Image:
    - Provides quick overview of scene
    - Detect students wishing to interact
  - "Traditional" Zoom Camera Mounted on Pan-Tilt Unit (PTU):
    - Automatically steered in direction of potential speaker

PTU Mounted Camera

Audio System

- **Microphone Array:**
  - Four omni-directional microphones mounted in a static pyramidal shape about the ParaCamera
  - Beamforming → "Steer" array in some direction:
    - Appropriately delaying the signal of each mic. ensures desired signal is reinforced, while noise & sound coming from other directions is attenuated
  - Sound Localization Techniques:
    - Correlation of microphone signals to detect novel acoustic events

Instructor's Touch-Screen GUI

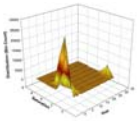
Potential Speakers

Overview

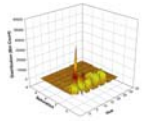
- **At Remote (Student) Sites:**
  - Detect potential people present in ParaCamera view wishing to interact with instructor (hand raising gestures or speech)
  - Estimate their "real-world" direction
  - Present this info. to the instructor's GUI
- **At Instructor's Site:**
  - By "clicking" on GUI, interaction can take place
    - → Pan-tilt camera & audio array focused on student

Detecting Hand Raising Gestures

- **Color Cues:**
  - HSV color models for both skin and non-skin color classes:
    - Constructed by manually classifying portions of ParaCamera images as either skin or non-skin



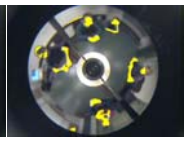
Skin Histogram



Non-Skin Histogram

  - Bayes' rule is used to classify the pixels of each image as either skin or non-skin:
    - Group "connected skin pixels" into separate regions
    - Separate skin regions into clusters

Original Image



Detected Skin Regions

  - "Real-world" estimate of direction to each skin region can be made
- **Motion Cues:**
  - Restrict skin color pixel classification to regions in image where motion occurs:
    - Image differencing over one or more frames
      - → Background subtraction or
      - → Subtraction of image at time τ and $\tau - n$
- **Segmentation of the Raising Hand/Arm:**
  - Group together spatially close skin color regions which have "moved":
    - Convex Hull
- **Sample Output:**
  - Sequence of three consecutive images of a hand raising gesture:
    - Yellow outline denotes raising hand/arm as determined by the system

Time