MIT2000-05
A Multi-Cue Vision Person Tracking Module

Trevor Darrell

Next generation intelligent environments and interfaces require low-cost, easily configurable person tracking systems to provide perceptual awareness of users. We propose to build a robust multi-cue vision module that will provide these services.

Most approaches to vision-based perceptual interfaces have relied on a single visual modality to detect, track, and recognize people. For example, face recognition usually relies on a statistical or neural network model of the intensity variation in a particular class of patterns [Pentland, Poggio, Rowley]. Research on body tracking often uses either motion analysis with an optic flow estimation algorithm, or a color-based background modeling and elimination technique [Wren et al.'s Pfinder, Grimson, Bregler, and many more]. Each of these techniques is often successful in the intended domain, and in laboratory experiments. However they often have difficulty in complex, unpredictable environments such as those with non-static background, changing illumination, and/or variable viewing geometry or object pose.

Our previous research demonstrated the value of using multiple visual modalities to achieve robust real-time person detection, tracking, and recognition. We combined stereo-based shape analysis, face pattern detection, and color analysis of skin, hair, and clothing regions. When run alone, each module had significant error modes and often only modest frame rate (1Hz). The integrated system was able to take advantage of the relative strengths of each module, e.g., the speed and view-invariance of the color module, vs. the low false positive rate of the pattern module. Each module contributed to detection (in a single frame), tracking (finding a corresponding detection in subsequent frames), and recognition of returning visitors.

The major drawbacks of our prototype system were the hardware resources and calibration efforts needed to run the system. Real-time stereo processing hardware provided frame-rate stereo depth images, but required a dedicated 8-chip FPGA board and associated control processor. Face pattern detection was implemented using the CMU neural network, one of the fastest systems reported in the literature, which nonetheless required a dedicated 500Mhz processor to run at 1Hz frame-rate on NTSC

images. The remaining processing and control were run in a third processor. Each processor ran as a separate PC and was loosely coupled via fast Ethernet. Installation of the system was difficult due to calibration procedures for the stereo system, which required manual alignment of the cameras, and for the color module, which required specification via example of the current illumination condition to learn the skin color class.

We propose to develop the next generation of this system. We would like to engineer a lightweight, easily deployable version of the existing approach, and add an infrared imaging option and a foreground motion analysis capability in the pattern module. We also would like to integrate head and body pose tracking explicitly into the new system.

Advances in general purpose CPU power in the last few years have made real-time implementations of stereo correspondence practical, so it is now possible to propose a robust multi-cue person tracking system with real-time stereo using general-purpose CPU and DSP components.

Our next generation system would include the following advances:

- Tighter coupling of inter-module control and search to reduce computational demand
- Integration of pattern decection/tracking/recognition steps into single statistical framework (e.g., Cootes and Taylor et al.) for improved performance
- Automatic multi-view/stereo calibration (e.g., Faugeras et al.)
- Skin-class and illuminant modeling via face-dectection bootstrap
- Perceptual output standards (XML? specification of user's physical state)
- Integrated head and body pose tracking
- Foreground pattern motion tracking
- TM1000 / DSP implementation of preprocessing steps (stereo correspondence, network convolution, subspace projection, and color indexing).
- Lightweight COTS stereo head (e.g., Visage, Inc.)
- Optional IR sensing modalities for outdoor environments.

Rather than require a triple-headed computer system with hundreds of watts of power, we would implement this system on a single motherboard system. We expect to utilize dual 800Mhz-1Ghz processors, possibly with a DSP coprocessor board (the Trimedia environment looks particularly promising). We would utilize cheap COTS stereo camera heads which are approx 1in x 4in x 1in, and IR imaging systems which are available in

standard digital camera form factors. Our system would require no user calibration, and would have a wireless LAN connection so only a single physical electrical connection would be required– a power plug. When configured without an IR sensor, we would hope to build these modules for on the order of $1K in parts. (The IR sensors are currently $10K, and so would be best relevant for environments where cost is not a factor.)

Development of this Robust Multi-cue Person Tracking System is an engineering challenge that will require significant algorithmic innovation (e.g., the first seven bullet items above.) Successful deployment of these modules will catalyze research on intelligent environments at MIT and elsewhere, allowing visual information about people to become a standard user-interface commodity that vision non-specialists can utilize.