Networks of Multi-cue Vision Tracking Modules
MIT2000-05
Trevor Darrell

Our work to date on the NTTMIT collaboration has focused on providing next generation intelligent environments and perceptual interfaces with low-cost, easily configurable person tracking systems to provide a visual awareness of users. We proposed, and are building, a robust multi-cue vision module that tracks and identifies users using range, color and pattern information.

Our system is based on networks of stereo vision sensors and optimal trajectory estimation in a spatio-temporal plan-view representation. (See the recent technical report, *Plan-view trajectory estimation with dense stereo views*, AI Memo 2001-01, to appear at the International Conference on Computer Vision this summer.) We have installed this system in the E21 meeting room at the AI Lab, and will use it to guide an active microphone array for enhanced audio input, as well as providing context information to applications: e.g., who is in the room, what devices they are near, what activity they are performing, etc.

For 2001-2002, we would like to continue research on this system, focusing on the following enhancements:

1. Integration of stereo tracking with multiple-viewpoint rendering (VVR/VisualHull).
2. Gesture and Expression tracking
3. Lightweight deployment and automatic calibration

Integration with multiple-viewpoint rendering

Previous work in the NTT-MIT collaboration examined algorithms for rendering images of scenes given a set of silhouette views. The “VVR” system of Profs. Grimson and Viola constructed a voxel representation from a set of silhouette views. Recently, Prof. Lenoard McMillian has extended this work using the idea of an efficient volumetric representation based on the visual hull, and demonstrated a real-time algorithm. A major problem with both the VVR and Visual Hull systems is they require static background scenes. We would like to combine a variable-viewpoint rendering system with our network of range sensors, and be able to render arbitrary views of objects without requiring a fixed background.

Gesture and Expression tracking

We would like to add kinematic and non-rigid shape processing and recognition to our tracking system. After segmentation into clusters of novel points in plan view, range data from each person should be further parsed into arm, body, and face regions. Pointing and other articulated gestures would be recognized, as well as the direction of facial gaze. This would support smart meeting rooms that would like information about how participants are interacting.

Lightweight deployment and automatic calibration

In addition to extending the functionality of our system as reflected in the previous two items, we wish to continue to engineer the system towards a low-cost, easily configured implementation. This was an issue in last year's proposal, but we have not made as much progress as we had expected. Currently we are prototyping nodes based on $3K special purpose cameras and desktop workstations with wired network connections; the next iteration will be to use laptop computers, wireless Ethernet, and pairs of commodity 1394 web cams ($200/ea). The following iteration would be to build a fully embedded node, with low power CPU and DSP or ASIC for stereo processing.