Person Tracking with Stereo Range Sensors

John Viloria, David Demirdjian, Neal Checka & Trevor Darrell

Artificial Intelligence Laboratory
Massachusetts Institue Of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

The Problem: The goal of this project is to design and build a person tracking system using stereo cameras. Our system, which stands in an ordinary conference room, can track multiple people and understand pointing gestures.

Motivation: Systems which can track and understand people have a wide variety of commercial applications. It is predicted that computers of the future will interact more naturally with humans than they do now. Instead of the desktop computer paradigm with humans communicating by typing, computers of the future will be able to understand human speech and movements. Our system demonstrates the capabilities of a solely vision-based system for these ends.

Previous Work: Several systems have been created which perform some of the above tasks. Krumm, Harris, and Meyers et al [1] used stereo cameras to track people as they moved about a room. Their system worked with multiple people, but required each person to enter and exit a specific point in the room to maintain the number of people in the room; our system counts the number of people in real time. Jojic, Brrummitt, and Meyers et al [2] used stereo cameras to detect people pointing and estimate the direction of their pointing. We will use a similar method for our gesture recognition.

Approach: Three camera modules, each consisting of a stereo camera and a computer, are situated in the room. The cameras are arranged to view the entire room, and continually estimate 3D-point clouds of the objects in the room. The computers associated with each camera calculate background models, and are able to identify the points associated with novel objects in the room. The foreground points are passed to an integration module, which clusters the points into blobs which represent people. From these blobs, features such as person location and posture are extracted. A gesture recognition module analyzes the posture and infers what the person is communicating. The current implementation of the tracking system uses three stereo cameras, but it is easily extensible to use as many cameras modules as needed.

Difficulty: There are several types of difficulties which can degrade performance at different levels in the system. At the stereo module stage, the stereo cameras can fail to produce depth data if there is not enough texture or if parts of the image are not different enough to be identifiable. Because the points are sent to the integration module by each camera module, some effort must be made to synchronize the incoming foreground point data from each module.

Impact: Once the techniques for tracking and recognizing gestures are mastered, visual stereo techniques can be infused into all sorts of applications - video surveillance, games, or other intelligent objects.

Future Work: (This project began in September 2000.)

Research Support: Project Oxygen, NTT

References:

[1] R. Krumm, S. Harris, B. Meyers et al, "Multi-Camera Multi-Person Tracking for EasyLiving," Third International Workshop on Visual Surveillance, 2000.

[2] N. Jojic, B. Brumitt, B. Meyers et al, "Detection and Estimation of Pointing Gestures in Dense Disparity Maps," Conference on Automatic Face and Gesture Recognition, 2000.