A Multi-Cue Vision Person Tracking Module
MIT2000-05

Progress Report: January 1, 2001—June 30, 2001

Trevor Darrell, Hiroshi Murase

Project Overview

The goal of this project is to develop a robust person tracking module that use multiple visual cues, processing modalities, and viewpoints. Initial development efforts are focusing on color and stereo range processing modalities, for deployment as a component module in an intelligent environment tracking system.

Progress Through June 2001

We have developed a multi-view stereo range-based person tracking system. Tracking people in known environments has recently become an active area of research in computer vision. Several person tracking systems have been developed to detect the number of people present as well as their 3-D position over time. These systems generally used a combination of foreground/background classification, clustering of novel points, and trajectory estimation in one or more camera views. Many color-based approaches to background modeling had considerable difficulty with fast illumination variation due to changing lighting and/or video projection. To overcome this, we have explored the use of background shape models based on stereo range data.

Unfortunately, the background models built by these systems are often sparse, due to the many regions of uniform brightness where stereo estimation fails in a typical background training sequence. Additionally, these systems are based on exhaustive stereo disparity search, which can be slow. Most recently, we have developed new techniques for dense, fast range-based tracking with modest computational complexity. There are three steps to our method. We first recover dense depth data using multiple-gain imaging and long-term observation approaches. We then match uniform but unoccluded planar regions in the scene and interpolate their interior range values. We finally apply ordered disparity search techniques to prune most of the disparity search computation during foreground detection and disparity estimation, yielding a fast, illumination-insensitive 3-D tracking system.

When objects are moving on a ground plane and are observed from multiple widely-separated viewpoints, rendering an orthographic vertical projection of foreground activity is useful. A “plan-view” image facilitates correspondence in time since only 2D search is required. Typically, previous systems would segment foreground data into regions prior to projecting into a plan-view, followed by region-level tracking and integration, potentially leading to sub-optimal segmentation and/or object fragmentation.

Instead, we have developed an approach which avoids any early segmentation of the foreground data. We merge the plan-view images from each viewpoint and estimate over time a set of trajectories that best accounts for the integrated foreground density. Trajectory estimation is performed using a dynamic programming-based algorithm, which can optimally estimate the position over time as well as the entry and exit locations of an object. This contrasts previous approaches, which generally used instantaneous measures, and/or specific object creation zones to decide on the number of objects per frame.

A technical report describes our recent progress in more detail. The report MIT-AIM-2001-001, available on the NTTMIT web site, describes our new algorithm for computing dense range-based foreground estimates and for fast estimation of foreground disparities. It also introduces the plan-view tracking representation and our algorithm for optimally estimating trajectories with limited temporal extent. It shows how this method can accurately detect the entry and exit of objects without constraints on the spatial location of such events, and includes a discussion of the overall system's performance and implications, as well as possible avenues for future work

Research Plan for the Next Six Months

In the next six months, our efforts will focus on improving the robustness and performance of the current visual tracking system, demonstrating its application in active camera and microphone tracking, and integrating new capabilities for virtual view generation and view-invariant face/gait recognition.

To improve robustness and performance of the current system, we are pursing several issues:

- Improve speed of range foreground estimation

We have implemented a fast, predictive range estimation algorithm that prunes the disparity search space at background pixels. The current implementation was done in C++ and was faster than an optimized global (non-pruned) search implementation. We are implementing our predictive scheme in optimized machine code, and expect further speed improvements.

- Automate geometric calibration of multiple camera system

Currently we perform an offline calibration procedure to find the relative position and orientation of each stereo rig. This is simple and suitable for research, but would not be appropriate for a deployed system. We are working on a fully automatic, online calibration system version that estimates and refines camera position from observed person trajectories.

- Add texture/color classification to disambiguate intersecting trajectories

The current version of the system has no ability to disambiguate users who are momentarily adjacent. We are adding texture and color features to discriminate users, and to match portions of their trajectory before and after they become adjacent.

- Add virtual background constraints from multiple views

Stereo background models often suffer from undefined range values. Our current system alleviates this through the use of long-term and variable gain/illumination imaging conditions. Recently, we've developed a new method that uses visibility constraints from multiple stereo views to detect foreground points. We are incorporating this method in our real-time system through the use of a virtual background that is constructed from these freespace visibility constraints evaluated along epipolar lines in the other stereo rigs.

We plan to demonstrate the utility of the tracking system to guide high-resolution sensors, including

- Active camera control

We have already implemented a simple system to point an active camera at the first user moving in the space. Future work is to decide how to plan which people to follow in the general case of N people and M<N cameras.

- Focusing microphone array on one or more speakers (and attenuating other speakers)

In our test environment we have a 32 element microphone array, which can be electronically steered to focus regions of audio enhancement (and attenuation) to various locations. We are connecting this to our person tracking system, so that regions of audio enhancement can be focused on the location of people in the room.

Finally, we plan to extend the current system to include recent algorithms developed at MIT for virtual view generation (such as the *VVR* and *Visual Hull* systems):

- View-invariant recognition for identification and trajectory disambiguation

Given a set of silhouette images computed from multiple views, a textured visual hull can be used to render new images for recognition in canonical pose. We've separately developed a system to do this from a set of monocular images with color background models. We plan to implement it with our set of multiple stereo sensors, which would allow visual hull extraction despite non-constant background.

Model generation for virtual teleconferencing

These textured visual hull models can also be used as 3-D models in virtual teleconferencing or virtual reality applications.