Multi-Person Tracking with Stereo Range Sensors

David Demirdjian, Neal Checka & Trevor Darrell

Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

The Problem: The goal of this project is to design and build a multi-person tracking system using a network of stereo cameras. Our system, which stands in an ordinary conference room is able to track people, estimate their trajectories as well as different characteristics (*e.g.* size, posture, ...).

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: 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 3D position over time. These systems use a combination of foreground/background classification, clustering of novel points, and trajectory estimation over time in one or more camera views [3].

Color-based approaches to background modeling have difficulty with illumination variation due to changing lighting and/or video projection. To overcome this problem, several researchers have supported the use of background models based on stereo range data [3]. Unfortunately, most of these systems are based on computationally intense, exhaustive stereo disparity search.

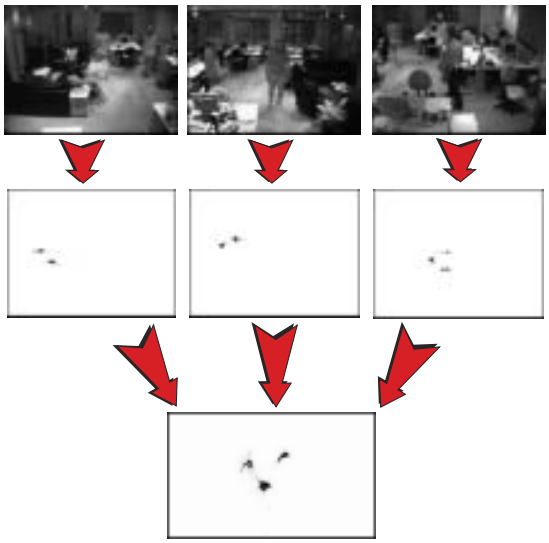

Figure 1: Detecting locations of users in a room using multiple views and plan-view integration. Three people are standing in a room, though not all are visible to each camera. Foreground points are projected onto a ground plane. Ground plane points from all cameras are then superimposed into a single data set before clustering the points to find person locations.

Approach: We have developed a system that can perform dense, fast range-based tracking with modest computational complexity. We apply ordered disparity search techniques to prune most of the disparity search computation during foreground detection and disparity estimation, yielding a fast, illumination-insensitive 3D tracking system. Details of our system are presented in [2].

When tracking multiple people, we have found that rendering an orthographic vertical projection of detected foreground pixels is a useful representation (see also [4]). A "plan view" image facilitates correspondence in time since only 2D search is required. 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 develop a technique that altogether avoids any early segmentation of foreground data. We merge the plan-view images from each view and estimate over time a set of trajectories that best represents 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 [1] .

Future Work: While the results are appealing, a problem remains: when the trajectory of two objects or people overlap, it is not possible from a foreground density representation to disambiguate trajectories if they subsequently separate. Appearance information can resolve this. Unfortunately, including this in the dynamic programming optimization would greatly increase the size of the state space of locations at each time frame, making the solution for the optimal trajectory impractical. Resolving this is a topic of ongoing and future work. We plan to use a trajectory-level correspondence process that uses a graph based on the overall trajectory data, computes aggregate appearance information along each edge (e.g., using color histograms), and then matches these features to resolve identity along each edge.

References:

[1] D.J. Beymer. Person counting using stereo. In *Worshop on Human Motion*, 2000.

[2] T. Darrell, D. Demirdjian, N. Checka, and P.F. Felzenszalb. Plan-view trajectory estimation with dense stereo background models. Technical Report AI Memo 2001-001, MIT Artificial Intelligence Laboratory, February 2001.

[3] T. Darrell, G.G. Gordon, M. Harville, and J. Woodfill. Integrated person tracking using stereo, color and pattern detection. *IJCV*, 2(37):175–185, June 2000.

[4] R. Krumm, S. Harris, and B. Meyers et al. Multi-camera multi-person tracking for easyliving. In *Third Workshop on Visual Surveillance*, 2000.