DETECTION AND TRACKING OF PEOPLE USING STEREO

Liang Zhao* and Larry Davis
Computer Vision Laboratory
University of Maryland
College Park, Maryland 20742

ABSTRACT

We present a real-time people detection and tracking system using sparse stereo ranging. The goal of our project is to provide the navigation component of an unmanned ground vehicle (UGV) navigation system with timely and accurate estimates of the locations and trajectories of people and vehicle in the vicinity of the UGV. It is intended to promote the survivability of the platform and of the people who will interact with it.

1. INTRODUCTION

We present a real-time people detection and tracking system using sparse stereo ranging. The goal of our project is to provide the navigation component of an unmanned ground vehicle (UGV) navigation system with timely and accurate estimates of the locations and trajectories of people and vehicles in the vicinity of the UGV. It is intended to promote the survivability of the platform and of the people who will interact with it. Additionally, the same technology can be used to provide unmanned perimeter security for temporary or permanent bases and storage facilities.

In our system people are first detected through a combination of background subtraction (Horprasert, Harwood and Davis, 2000) and then periodicity analysis of the changing shapes of objects as they are tracked through the scene. In our initial work, we assume that the vehicle is stationary. Other research in our laboratory addresses the problem of detecting people from moving platforms, when background subtraction cannot be used as an initial detection mechanism. Fig. 1(b), below, shows typical results from the background subtraction module of the system. The system does have a shadow suppression capability, but in this particular example shadows are cast on a gray surface, and shadow suppression generally requires color information.

2. SPARSE STEREO RANGING

The use of stereo provides more robust detection and tracking through the use of 3D information, and sparse stereo ranging (i.e., computing stereo correspondences at only a small subset of the image positions) is faster and on average more accurate than dense stereo ranging because of its use of only distinct features that typically can be correctly matched. In order to perform sparse stereo ranging, distinct features are first selected from the foreground region in the left image (see Fig. 1(c)). In our system, these features correspond to edge points with a high variation in gradient directions in their local neighborhoods. Their corresponding features in the right image are searched along the epipolar line based on a classical dissimilarity measure — sum-of-squared-differences (SSD) (Matthies, L., Szeliski, R., and Kanade T., 1989). The corresponding 3D position of each pair of matching feature points is then calculated based on their disparity and a prior exterior calibration of the stereo camera pair. Fig. 1(c) shows the points chosen from the detected person in Fig. 1(b) as interest points, and Fig. 1(d) shows the matching points found by the stereo matching algorithm.

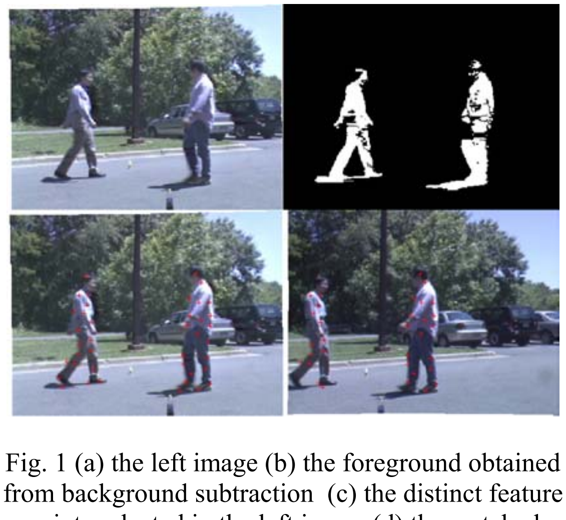

Fig. 1 (a) the left image (b) the foreground obtained from background subtraction (c) the distinct feature points selected in the left image (d) the matched feature points in the right image

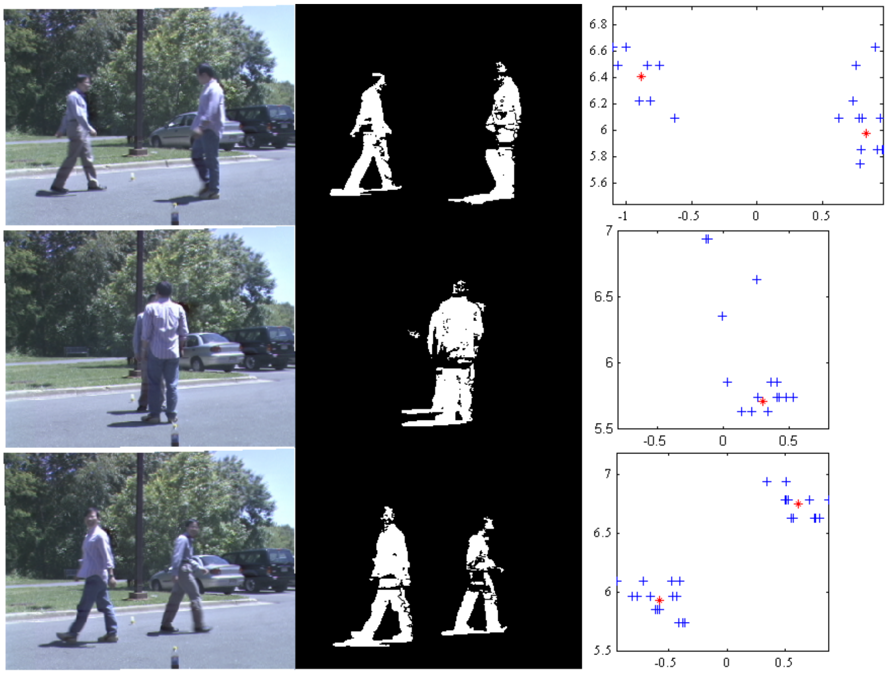

Fig. 2 Results of sparse stereo ranging

3. EXPERIMENTAL RESULTS

Examples of 3D localization are shown in Figure 2. Column 1 shows three images, one from each of three stereo pairs, and the second column shows the results of background subtraction. The third column in Figure 2 shows the projection of the recovered 3D person positions onto the ground plane. People are located and tracked by clustering the ground plane projections of the three dimensional positions of individual body points. Generally, we would like to estimate smooth trajectories for moving people, but simply taking the average 3D position of feature points from the silhouette will not provide smooth tracks because of the biasing effects of the extremities – arms and legs. In our current implementation, we simply compute the median position of the ground projected 3D points, but our current research aims to filter out the position estimates from points on the extremities. Fig. 3 shows the ground plane trajectories for the complete sequence.

The trajectories of people recovered from tracking (see Fig. 3) are then used to limit the search band of stereo matching in the next frame and to further reduce false matches and computational complexity. Experiments on outdoor scenes demonstrate that our system can detect people and track their 3D positions in real-time (20fps on Intel Xeon 1.40Ghz).

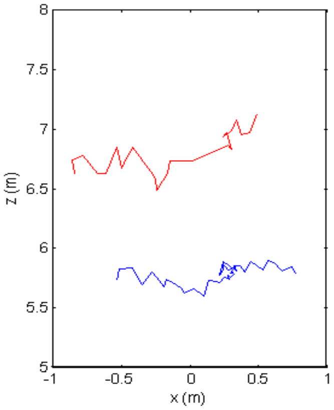

Fig. 3 The overhead map of trajectories of the tracked people

REFERENCE

Horprasert, T., Harwood D. and Davis L., “A Robust Background Subtraction and Shadow Detection,” *Proc. of Asian Conf. on Computer Vision*, 2000.

Matthies, L., Szeliski, R., and Kanade, T., “Kalman Filter-Based Algorithms for Estimating Depth from Image Sequences,” Int’l J. of Computer Vision, pp.209-236, 1989.