Stereo Person Tracking with Adaptive Plan-View Statistical Templates

Michael Harville
Hewlett-Packard Laboratories, 1501 Page Mill Rd., Palo Alto, CA 94304
harville@hpl.hp.com

Abstract

As the cost of computing per-pixel depth imagery from stereo cameras in real time has fallen rapidly in recent years, interest in using stereo vision for person tracking has greatly increased. Methods that attempt to track people directly in these "camera-view" depth images are challenged by their substantial amounts of noise and unreliable data. Some recent methods have therefore found it useful to first compute overhead, "plan-view" statistics of the depth data, and then track people in images of these statistics. We describe a new combination of plan-view statistics that better represents the shape of tracked objects and provides a more robust substrate for person detection and tracking than prior plan-view algorithms. We also introduce a new method of plan-view person tracking, using adaptive statistical templates and Kalman prediction. Adaptive templates provide more detailed models of tracked objects than prior choices such as Gaussians, and we illustrate that the typical problems with template-based tracking in camera-view images are easily avoided in a plan-view framework. Our results indicate superior tracking through phenomena such as complex inter-person occlusions and close interactions.

1. Introduction

While most real-time vision-based person tracking methods operate primarily on color or grayscale video, interest in augmenting this input space with depth or disparity imagery has grown as hardware and software for computing this data from stereo cameras has recently become much faster and cheaper [7, 9, 11]. Depth data has great potential for improving person tracking systems' performance, because it

- Is a powerful cue for foreground segmentation
- Provides shape and metric size information that can be used to distinguish people from other foreground objects
- Allows occlusions of people by each other or by background objects to be detected and handled more explicitly
- Permits the quick computation of new types of features for matching person descriptions across time
- Provides a third, disambiguating dimension of prediction in tracking

Several person detection and tracking methods that make use of real-time, per-pixel depth data have been described in recent years. Most of these analyze and track features, statistics, and patterns directly in the depth images themselves [1, 4, 5, 8]. This methodology is not as fruitful as one might hope, however, because today's stereo cameras produce depth images whose statistics are far less clean than those of standard color or monochrome video. For multi-camera stereo implementations, which compute depth by finding small area correspondences between image pairs, unreliable measurements often occur in image regions of little visual texture, as is often the case for walls, floors, or people wearing uniformly-colored clothing, so that much of the depth image is unusable. Also, it is difficult to find correct correspondences in regions, usually near depth discontinuities in the scene, that are visible in one input image but not the other. This results in additional regions of unreliable data, and causes the edges of an object in a depth image to be noisy and poorly aligned with the object's color image edges. All of these problems are evident in the typical color and depth image pair of Figure 2.

Even at pixels where depth measurements are typically informative, the sensitivity of the stereo correspondence computation to very low levels of imager noise, lighting fluctuation, and scene motion leads to substantial depth noise. For apparently static scenes, the standard deviation of the depth value at a pixel over time is commonly on the order of 10% of the mean - much greater than for color values produced by standard imaging hardware.

To combat these problems, some very recent person tracking methods have been based not on analysis of the raw depth images, but instead on images of depth statistics that are more conducive to the tracking task. Specifically, these methods have used the metric shape and location information inherent in the original "camera-view" depth images to compute statistics of the scene as if it were observed by an overhead, orthographic camera. In Section 2, we describe and motivate the use of these "plan-view" statistics for person tracking, and we begin to introduce a new combination of plan-view statistics that better preserves object shape information than prior approaches. We believe this choice of statistics, whose computation is described in more detail in Section 3, offers a superior basis for robust person tracking.

In Section 4, we describe a person detection and tracking method that has not previously been applied to plan-view images of any kind. The method uses Kalman prediction on adaptive statistical templates, and provides a more detailed

description of tracked people than the models used by prior plan-view methods. These improved person models allow for better tracking through complex inter-person occlusions and close interactions, among other benefits. We also show how the typical problems with adaptive template tracking in camera-view images are easily side-stepped in a plan-view framework. Section 5 discusses the high-quality tracking results obtained by applying these more flexible, detailed person models to our richer plan-view statistical basis.

2. Plan-View Statistics

The motivation behind using plan-view statistics for person tracking begins with the observation that, in most situations, people usually do not have significant portions of their bodies above or below those of other people. We might therefore expect to separate people more easily, and to reduce occlusion problems, by mounting our cameras overhead and pointing them toward the ground. However, methods based on monocular video that exploit this idea usually either must continue to deal with significant occlusion problems in all but the central portion of the image (particularly if wide-angle lenses are used), or must accept a somewhat limited field of view (particularly if the ceiling is relatively low). Furthermore, when mounted overhead, the cameras used for tracking are not suitable for extracting images of people's faces, which are desired in many applications that employ vision-based person tracking.

With a stereo camera, we can produce orthographically projected, overhead views of the scene that better separate people than the perspective images produced by a monocular camera. In addition, we can produce these images even when the stereo camera is not mounted overhead, but instead at an oblique angle that maximizes viewing volume and preserves our ability to see faces. All of this is possible because the depth data produced by a stereo camera allows for the partial 3D reconstruction of the scene, from which new images of scene statistics, using arbitrary viewing angles and camera projection models, can be computed. Plan-view images are just one possible class of images that may be constructed, and are discussed in greater detail below.

Every reliable measurement in a depth image can be back-projected, using camera calibration information and a perspective projection model, to the 3D scene point responsible for it. By back-projecting all of the depth image pixels, we create a 3D point cloud representing the portion of the scene visible to the stereo camera. If we know the direction of the "vertical" axis of the world - that is, the axis normal to the ground level plane in which we expect people to be well-separated - we can discretize space into a regular grid of vertically oriented bins, and then compute statistics of the 3D point cloud within each bin. A plan-view image contains one pixel for each of these vertical bins, with the value at the pixel being some statistic of the 3D points within the corresponding bin. This procedure effectively builds an orthographically projected, overhead view of some property of the 3D scene. Figure 1 illustrates this idea.

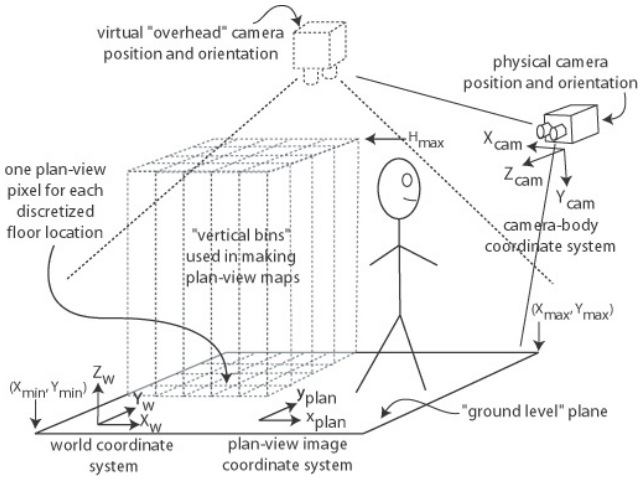

Figure 1. Concepts important to building plan-view maps.

Plan-view projection of per-pixel depth from stereo has been applied to person detection and tracking by Beymer [2], Darrell et. al. [3], and by researchers at Interval Research Corp. [12]. All of these methods chose to image the same statistic of the 3D points within the vertically oriented bins, namely the *count* of points in each bin. In the resulting images, referred to as plan-view "occupancy" or "density" maps, people appear as "piles of pixels" that can be tracked as they move around the ground. Although powerful, this representation discards virtually all object shape information in the vertical dimension. In addition, the occupancy map representation of a person will show a sharp decrease in saliency when the person is partially occluded by another person or object, as far fewer 3D points corresponding to the person will be visible to the camera.

To address these shortcomings, we image a second plan-view statistic, namely the height above the ground-level plane of the highest point within each vertical bin. This image, which we refer to as a "plan-view height map", is effectively a simple orthographic rendering of the shape of the 3D point cloud when viewed from overhead. The notion of applying plan-view height maps to person tracking has been explored preliminarily by Interval researchers [12], and plan-view height maps from stereo have been used in other contexts such as path-planning for the Mars rover [10]. Height maps preserve about as much 3D shape information as is possible in a 2D image, and therefore seem better suited than occupancy maps for distinguishing people from each other and from other objects. This shape data also provides richer features than occupancy for accurately tracking people through close interactions and partial occlusions. Furthermore, when the stereo camera is mounted in a high position at an oblique angle, the heads and upper

Figure 2. Example camera-view input. Left to right: (a) Color, (b) Depth (unreliable data shown in black), (c) Foreground color.

bodies of people often remain largely visible during inter-person occlusion events, so that a person's height map representation is usually more robust to partial occlusions than his occupancy map statistics.

Use of the point with maximal, rather than, say, 90th percentile, height within each vertical bin allows for fast computation of the height map, but makes the height statistics very sensitive to depth noise. In addition, the movement of relatively small objects at heights similar to those of people's heads, such as when a book is placed on an eye-level shelf, can appear similar to person motion in a height map. We address both of these problems by using occupancy and height statistics together, and by using the occupancy statistics to refine the height map statistics. This novel approach circumvents many of the problems of using either statistic in isolation, and is described in greater detail in Section 3.

3. Building Maps of Plan-View Statistics

The input to our method is a video stream of "color-with-depth"; that is, the data for each pixel in the video stream contains three color components and one depth component. Color and depth from one frame of such a stream is shown in Figures 2**a** and 2**b**. We restrict our attention to the "foreground" in the scene, as in Figure 2**c**. We extract foreground using a method that models both the color and depth statistics of the scene background with Time-Adaptive, Per-Pixel Mixtures Of Gaussians (TAPPMOGs), as detailed in [6].

Each foreground pixel with reliable depth is used in building a 3D point cloud. For a binocular stereo pair with horizontal separation b, horizontal and vertical focal lengths f_u and f_v, and image center of projection (u_o, v_o), we can project the disparity $disp$ at camera-view pixel (u, v) to a 3D location $(X_{cam}, Y_{cam}, Z_{cam})$ in the camera body coordinate frame (see Figure 1) as follows:

$$Z_{cam} = \frac{bf_u}{disp}, X_{cam} = \frac{Z_{cam}(u - u_o)}{f_u}, Y_{cam} = \frac{Z_{cam}(v - v_o)}{f_v} \quad (1)$$

We then transform these camera frame coordinates into the (X_W, Y_W, Z_W) world space, where the Z_W axis is aligned with the "vertical" axis of the world and the X_W and Y_W axes describe a ground level plane, by applying the rotation $\mathbf{R_{cam}}$ and translation $\vec{t}_{cam}$ relating the coordinate systems:

$$\begin{bmatrix} X_W & Y_W & Z_W \end{bmatrix}^T = -\mathbf{R_{cam}} \begin{bmatrix} X_{cam} & Y_{cam} & Z_{cam} \end{bmatrix}^T - \vec{t}_{cam} \quad (2)$$

Before building plan-view maps from the 3D point cloud, we must choose a resolution δ_{ground} with which to quantize 3D space into vertical bins. We would like this resolution to be small enough to represent the shapes of people in detail, but we must also consider the limitations imposed by the noise and resolution properties of our depth measurement system. In practice, we typically divide the $X_W Y_W$-plane into a square grid with resolution δ_{ground} of 2-4cm.

After choosing the bounds $(X_{min}, X_{max}, Y_{min}, Y_{max})$ of the ground level area within which we will restrict our attention, we can map 3D point cloud coordinates to their corresponding plan-view image pixel locations as follows:

$$\begin{aligned} x_{plan} &= \lfloor (X_W - X_{min})/\delta_{ground} + 0.5 \rfloor \\ y_{plan} &= \lfloor (Y_W - Y_{min})/\delta_{ground} + 0.5 \rfloor \end{aligned} \quad (3)$$

Plan-view height and occupancy maps, denoted as $\mathcal{H}$ and $\mathcal{O}$ respectively, can be computed in a single pass through the foreground data. To do so, we first set all pixels in both maps to zero. Then, for each pixel classified as foreground, we compute its plan-view image location (x_{plan}, y_{plan}), Z_W-coordinate, and Z_{cam}-coordinate using equations (1), (2), and (3). If the Z_W-coordinate is greater than the current height map value $\mathcal{H}(x_{plan}, y_{plan})$, and if it does not exceed H_{max}, where H_{max} is an estimate of how high a very tall person could reach with his hands if he stood on his toes, we set $\mathcal{H}(x_{plan}, y_{plan}) = Z_W$. We next increment the occupancy map value $\mathcal{O}(x_{plan}, y_{plan})$ by $Z_{cam}^2/f_u f_v$, which is an estimate of the real area subtended by the foreground image pixel at distance Z_{cam} from the camera. The plan-view occupancy map will therefore represent the total physical surface area of foreground visible to the camera within each vertical bin of the world space. The plan-view height and occupancy maps corresponding to the foreground of Figure 2 are shown in Figures 3**a** and 3**b**, respectively.

Because of the substantial noise in these plan-view maps, we denote them as $\mathcal{H}_{raw}$ and $\mathcal{O}_{raw}$, and we smooth them

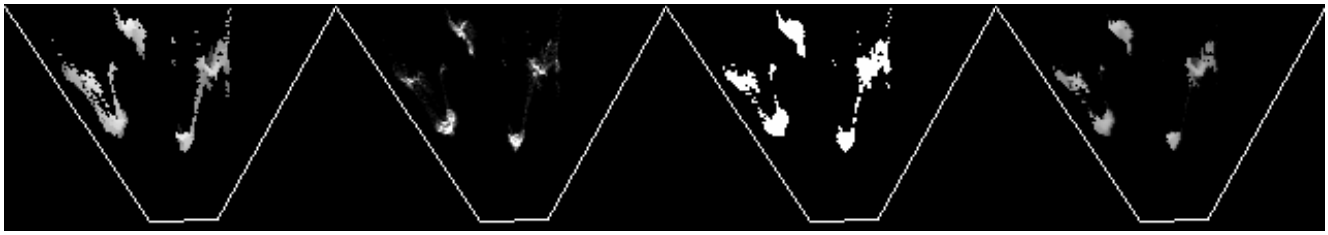

Figure 3. Plan-view maps of the foreground of Figure 2. The stereo camera's plan-view location is just outside the bottom of the images, and the lines indicate the camera's field of view. Left to right: (a) Raw height map $\mathcal{H}_{raw}$, (b) Raw occupancy map $\mathcal{O}_{raw}$, (c) Bitmap showing where smoothed occupancy exceeds threshold θ_{occ}, (d) Masked, smoothed height map $\mathcal{H}_{masked}$.

prior to further analysis. The smoothed maps $\mathcal{H}_{sm}$ and $\mathcal{O}_{sm}$ are generated by convolution with a Gaussian kernel whose variance in plan-view pixels, when divided by the map resolution δ_{ground}, corresponds to a physical size of 1-4cm. This reduces depth noise in person shapes, while retaining gross features like arms, legs, and heads.

Although the shape data provided by $\mathcal{H}_{sm}$ is very powerful, it is not prudent to give all of it equal weight. We propose to use the smoothed height map statistics only in floor areas where we believe that something "significant" is present, as indicated by the amount of local occupancy map evidence. We therefore prune $\mathcal{H}_{sm}$ by setting it to zero wherever the corresponding pixel in $\mathcal{O}_{sm}$ is below a threshold θ_{occ}. This helps us discount foreground noise that appears to be located at "interesting" heights, and helps us ignore the movement of small, non-person foreground objects, such as a book or sweater that has been placed on an eye-level shelf by a person.

Figure 3**c** shows the mask obtained by applying the threshold θ_{occ} to the smoothed occupancy map of the foreground of Figure 2. The result of applying this mask to $\mathcal{H}_{sm}$ is shown in Figure 3**d**. This masked height map $\mathcal{H}_{masked}$, along with the smoothed occupancy map $\mathcal{O}_{sm}$, provide the basis for the tracking algorithm described in Section 4.

4. Tracking and Adapting Templates of Plan-View Statistics

Our method uses classical Kalman filtering to track patterns of plan-view height and occupancy statistics over time. The Kalman state maintained for each tracked person is the three-tuple $\left\langle \vec{x}, \vec{v}, \vec{S} \right\rangle$, where $\vec{x}$ is the two-dimensional plan-view location of the person, $\vec{v}$ is the two-dimensional plan-view velocity of the person, and $\vec{S}$ represents the body configuration of the person. While one might think it preferable to parameterize body configuration in terms of joint angles or other pose descriptions, we find that simple templates of plan-view height and occupancy statistics provide an easily computed but powerful shape description. Hence, we update the $\vec{S}$ component of the Kalman state directly with values from subregions of the $\mathcal{H}_{masked}$ and $\mathcal{O}_{sm}$ images, rather than first attempt to infer body pose from these statistics, which is likely an expensive and highly error-prone process. Our Kalman state may therefore more accurately be written as $\langle \vec{x}, \vec{v}, \mathcal{T}_H, \mathcal{T}_O \rangle$, where $\mathcal{T}_H$ and $\mathcal{T}_O$ are a person's height and occupancy templates, respectively. The observables in our Kalman framework are the same as the state; that is, we assume no hidden state variables.

For Kalman prediction, we use a constant velocity model, and we assume that person pose varies smoothly over time. At high system frame rates, we therefore would expect little change in a person's template-based representation from one frame to the next. For simplicity, we predict no change at all. Because the template statistics for a person are highly dependent on the visibility of that person to the camera, we are effectively also predicting no change in the person's state of occlusion between frames. These predictions will obviously not be correct in general, but they will become increasingly accurate as the system frame rate is increased. Fortunately, the simple computations employed by our method are well-suited for high-speed implementation, so that it is not difficult to construct a system that operates at a rate where our predictions are reasonably approximate.

For the measurement step, we search in the neighborhood of the predicted person position $\vec{x}_{pred}$ for the location where the current plan-view statistics best match the predicted ones. This consists largely of comparing the person's $\mathcal{T}_H$ and $\mathcal{T}_O$ templates to the current $\mathcal{H}_{masked}$ and $\mathcal{O}_{sm}$ images. The area in which to search is centered at $\vec{x}_{pred}$, with a rectangular extent determined from the Kalman uncertainty in the positional state $\vec{x}$. For person i, a match score $\mathcal{M}$ is computed as follows at all locations $\vec{x}$ in the search zone:

$$\begin{aligned}\mathcal{M}(\vec{x}) = &\alpha * SAD(\mathcal{T}_H, \mathcal{H}_{masked}(\vec{x})) + \beta * SAD(\mathcal{T}_O, \mathcal{O}_{sm}(\vec{x})) \\ &+ \gamma * dist\left(\vec{x}_{pred}, \vec{x}\right) + \epsilon * \sum_{j<i} \eta\left(\vec{x}_j, W_{avg}, \vec{x}\right) \qquad (4)\end{aligned}$$

SAD refers to "sum of absolute differences", but averaged over the number of pixels used in the differencing operation so that all matching process parameters are independent of the template size. For the height SAD, we use a height difference of $H_{max}/3$ at all pixels where $\mathcal{T}_H$ has been masked to zero but $\mathcal{H}_{masked}$ has not, or vice versa. The third term in equation (4) encourages the matching of a person to his

Kalman-predicted location. In the last term, $\eta(\ldots)$ denotes the Gaussian function evaluated at location $\vec{x}$, where the Gaussian has a mean equal to the position estimate $\vec{x}_j$ of a person previously tracked in this frame, and a variance equaling the average person torso width W_{avg}. This final term discourages the matching of multiple people to nearly the same plan-view location. α, β, γ, and ϵ are chosen so that all terms have roughly equal influence, except for the distance term, which is given a lesser weight.

If the best (minimal) match score found falls below a threshold θ_{track}, we update the Kalman state with new measurements. The location $\vec{x}_{best}$ at which $\mathcal{M}(\vec{x})$ was minimized serves as the new position measurement, and the new velocity measurement is the inter-frame change in position divided by the time difference. The statistics of $\mathcal{H}_{masked}$ and $\mathcal{O}_{sm}$ surrounding $\vec{x}_{best}$ are used as the new body configuration measurement for updating the templates. A relatively high Kalman gain is used in the update process, so that templates adapt quickly. If the best match score is above θ_{track}, we do not update the Kalman state with new measurements, and we report $\vec{x}_{pred}$ as the person's location.

People are tracked by this method one at a time, in reverse order of their positional uncertainty. When a person is tracked successfully, plan-view data at the tracked location is cleared before tracking of the next person is attempted. After we have attempted to track all known people, we detect potential new people at plan-view locations with high occupancy and a local height map maximum above some plausible minimum height for people. If any newly detected people are sufficiently near (e.g. within 2 meters) the last predicted or measured location of a "lost" person who has not been tracked successfully for a short time (e.g. less than 4 seconds), we decide that the new person is in fact the lost person, and we update the lost person's state vector with that of the new person. Otherwise, a potential new person must be tracked successfully in some minimum number of consecutive frames before we decide that he is not just noise. Any "lost" person not tracked for a long enough time (e.g. 4 seconds) is removed from the list of tracked people.

Most template-based tracking methods that operate on camera-view images encounter difficulty in selecting and adapting the appropriate template size for a tracked object, because the size of the object in the image varies with its distance from the camera. In the plan-view framework described above, however, we are able to obtain good performance with a template size that remains *constant across all people and all time*. This is possible because plan-view representations of people are largely invariant to floor location relative to the camera, and because the spatial extents of most upright people, when viewed from overhead, fall within a limited range. People spend almost all their waking time in a predominantly upright position, even when sitting, so we employ templates whose pixel width and height correspond to a real plan-view size of $2 * W_{avg}$, which is twice an estimate of the average torso width of people. We use $W_{avg} \approx 40$cm. For people of unusual size or in unusual postures, this template size is not ideal, but still works well.

Templates that are updated over time with current image values inevitably "slip off" the tracked target, and begin to reflect elements of the background. This is perhaps the primary reason that adaptive templates are seldom used in current tracking methods, and our method as described thus far suffers from this problem as well. However, with our plan-view statistical basis, it is relatively straightforward to counteract this problem in ways that are not feasible for other image substrates. Specifically, we are able to virtually eliminate template slippage through a simple "re-centering" scheme, detailed below, that is applied on each frame after tracking has completed.

For each tracked person, we first examine the quality of the current height template $\mathcal{T}_H$. If the fraction of non-zero pixels in $\mathcal{T}_H$ has fallen below a threshold $\theta_{HTcount}$ (around 0.3), or if the centroid of these non-zero pixels is more than a distance $\theta_{HTcentroid}$ (around $0.5 * W_{avg}$) from the template center, we decide that the template has slipped too far off the person. We then search, within a square of width $2 * W_{avg}$ centered at the person's current position estimate, for the location $\vec{x}_{occmax}$ in $\mathcal{O}_{sm}$ of the local occupancy maximum. New templates $\mathcal{T}_H$ and $\mathcal{T}_O$ are then extracted from $\mathcal{H}_{masked}$ and $\mathcal{O}_{sm}$ at $\vec{x}_{occmax}$. Also, the person location in the Kalman state vector is shifted to $\vec{x}_{occmax}$, without changing the velocity estimates or other Kalman filter parameters. This procedure usually keeps templates solidly situated over the plan-view statistics representing a person, despite depth noise, partial occlusions, and other factors.

5. Results

We have implemented our method in C++ on a PC platform. Live color and depth video input, at 320x240 resolution and with subpixel disparity interpolation, is provided by a Point Grey Triclops stereo module [9]. With little attempt at optimization of our code, the overall system runs at 8Hz on a dual 750MHz-processor PC. Good tracking performance is obtained at this frame rate, but because of our method's underlying assumption of slow inter-frame evolution of plan-view statistics, we expect our tracking results to improve as the system frame rate is increased. This frame rate can obviously be increased through use of faster processors and better-optimized code, but one should also note that the Triclops' software computation of depth is the most computationally expensive component of our system. Use of a stereo camera head with hardware-assisted depth computation, such as that available from Tyzx Inc. [11], would therefore dramatically improve the system speed.

We have quantitatively evaluated our method on several color-with-depth video test sequences captured at 12-15Hz

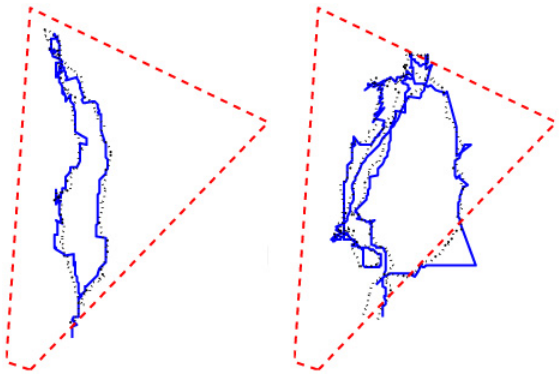

Figure 4. Comparisons of two individual result tracks with estimated ground truth. Results are shown in solid lines, ground truth in dotted lines. Stereo camera field of view is shown in dashed lines. Plan-view position of stereo pair is near upper-right corner of the images.

and 320x240 resolution. The stereo camera was typically mounted about 2.2m above the ground, with a view like that of Figure 2. Tracking results were examined for "significant" errors, defined as any of 1) losing track of a person, 2) failing to detect a person, 3) swapping the identities of two tracked people, and 4) tracking a non-person object. The test sequences, totaling about 10 minutes in duration, are very challenging: they contain dozens of majority or complete occlusions of people by one or more other people or static objects, and contain many close inter-personal interactions of extended duration. Rolling chairs are pushed around the room and left in new places, other objects are deposited into or removed from the scene, and large dark shadows appear when people stand in certain places. Some people walk behind cubicle walls so that only their heads are visible, others sit down on chairs or on the floor, and another performs a cartwheel. Many unusual postures and gestures are observable.

Despite these challenges, and without requiring extensive parameter tuning, our method made only one significant error: the tracks of two people were swapped when they interacted closely for a long period and were briefly joined by a third, occluding person. This performance is remarkable considering that it was obtained with no long-term person appearance models and no use of color beyond the foreground segmentation stage. We compared our method's performance against that obtained when either plan-view height or occupancy is omitted. Without occupancy, height maps are not pruned, and height alone is used to detect new people. Without height, occupancy alone is relied upon to decide when to re-center templates and to reject non-person objects. In both cases, over the same test sequences, performance was much poorer than when height and occupancy are combined: 17 significant errors when occupancy was not used, and 25 such errors when height was not used.

We have also evaluated the positional accuracy of our method's tracks, by comparing result tracks to manually estimated ground truth. For 8 person tracks that did not contain significant errors (e.g. the person was not lost, or swapped with another person), the measured track was differenced on a point-by-point basis with the ground truth estimate of the plan-view location of the person's head. Track sections for which the ground truth was outside the field of view of the camera - for instance, for a person who briefly exited the scene - were excluded from the comparison. The mean positional error was found to be 18cm, with a standard deviation of 14cm. Two example tracks are compared with ground truth in Figure 4.

6. Conclusions

An important contribution of this paper is its methods for combining and refining plan-view statistical maps to produce an excellent substrate for person detection and tracking. We introduce a novel template-based scheme for tracking in plan-view that takes great advantage of the detail in these plan-view maps, and we demonstrate that the typical difficulties with adaptive template tracking are easily avoided in our plan-view framework. The resulting method is highly amenable to real-time implementation, and exhibits robust performance under challenging conditions.

Acknowledgments

The author would like to thank John Woodfill, Gaile Gordon, Harlyn Baker, Trevor Darrell, and Ali Rahimi for inspiration and technical insights gained while working with them and others at Interval Research Corp. on the the first real-time person tracking system to use multiple stereo cameras and plan-view occupancy maps (to be described in a forthcoming paper).

References

[1] D. Beymer, K. Konolige. "Real-time tracking of multiple people using continuous detection." In *ICCV Frame-rate Workshop*, 1999.

[2] D. Beymer. "Person counting using stereo." In *Workshop on Human Motion 2000*, Dec 2000.

[3] T. Darrell, D. Demirdjian, N. Checka, P. Felzenszwalb. "Plan-view trajectory estimation with dense stereo background models". In *ICCV'01*, July 2001.

[4] T. Darrell, G. Gordon. M. Harville, J. Woodfill. "Integrated person tracking using stereo, color, and pattern detection." In *CVPR'98*.

[5] I. Haritaoglu, D. Harwood, L. Davis. "W^4S: A real-time system for detecting and tracking people in $2\frac{1}{2}$D." In *ECCV'98*, 1998.

[6] M. Harville. "A framework for high-level feedback to adaptive, per-pixel, mixture-of-gaussian background models." In *ECCV'02*.

[7] K. Konolige. "Small Vision Systems: hardware and implementation". *8th Int. Symp. on Robotics Research*, 1997.

[8] J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. Hale, S. Shafer. "Multi-camera multi-person tracking for EasyLiving". In *Workshop on Visual Surveillance 2000*, July 2000.

[9] Point Grey Research, http://www.ptgrey.com

[10] M. Snorrason, J. Norris, P. Backes. "Vision based obstacle detection and path planning for planetary rovers." In *Proc. of SPIE Vol. 3693, 13th AeroSense conference*, pp. 44-54, April 1999.

[11] Tyzx Inc. http://www.tyzx.com

[12] Unpublished work, Interval Research Corp., including real-time multiple person tracker using multiple stereo camera heads. First demonstration June 1999. Contributors include: Tim Allen, Harlyn Baker, Michael Coffey, Trevor Darrell, Gaile Gordon, Michael Harville, Lieven Leroy, Ali Rahimi, Paul Regier, John Woodfill.