Stereo Person Tracking with Adaptive Plan-View Statistical Templates

Michael Harville
Mobile and Media Systems Laboratory
HP Laboratories Palo Alto
HPL-2002-122
April 29th , 2002*

E-mail: harville@hpl.hp.com

person tracking, plan-view statistics, stereo depth images, adaptive template, Kalman filter

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 confronted 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. We compare results of our method with those for techniques using different plan-view statistics or person models, and find our method to exhibit superior tracking through challenging phenomena such as complex inter-person occlusions and close interactions. Reasonable values for most system parameters may be derived from physically measurable quantities such as average person dimensions.

* Internal Accession Date Only
Approved for External Publication
To be published in the Journal of Image and Vision Computing, 2003.

1 Introduction

Many methods for real-time multi-person detection and tracking with video cameras have been described in the literature. Unfortunately, few of these, if any, produce reliable results for long periods of time in unconstrained environments. This poor performance stems from the many difficult challenges that commonly beset the problem, among the most significant of which are:

- Segmenting the novel or dynamic objects ("foreground") in the video from the rest of the scene ("background")
- Distinguishing people from other foreground objects such as cars, shopping carts, or curtains blowing in the wind
- Avoiding distraction and confusion due to lighting-related scene appearance changes such as shadows, inter-reflections, and global illumination variation
- Tracking people through temporary occlusions, either in part or in full, by other people or by static objects in the scene
- Maintaining track integrity when people engage in close interactions, accelerate rapidly, or quickly change their body pose or appearance

Per-pixel depth or disparity imagery from stereo cameras offers much promise for dealing with these issues. For example, the distance information inherent in these images allows for straightforward assessment, in comparison with techniques based on monocular video, of the 3D locations of tracked objects. In addition, depth data

- Is a powerful cue for foreground segmentation
- Is relatively insensitive to lighting effects such as shadows and global illumination changes
- 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

In recent years, as hardware and software for computing depth imagery from stereo cameras has become increasingly fast and cheap [2–4,1,5], several person detection and tracking methods that make use of real-time depth data have been presented. Most of these analyze and track features, gradients, and smoothly connected regions directly in the depth images themselves [6–9]. When the depth images are accompanied by a spatially- and temporally-registered color or grayscale video stream, the results of the depth-based analysis are easily integrated with those extracted from the color or luminance data.

Fig. 1. Example of color-with-depth video input, obtained using the Point Grey Triclops camera [1]. In the depth image, brighter pixels indicate greater distance from the camera, and invalid (unreliable) depth data is shown in black.

Many of the traditional frameworks for tracking in monocular views may then be applied, but to the much richer per-pixel feature space of appearance (color or luminance) plus shape (depth).

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. This usually causes much of a depth image to be unusable. Also, it is not possible to find the correct correspondences in regions, usually near depth discontinuities in the scene, that are visible in one stereo 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 1.

Even at scene locations where depth measurements are 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. This noise makes it difficult to apply typical image analysis and tracking methods to depth data with the same confidence to which we are accustomed for color or monochrome video. For instance, a single person in a depth image is commonly split into multiple image regions not just by partial occlusions, but also by patches of unreliable depth data. In addition, the depth image gradients

dividing closely-spaced people are not unlike those that occur, for instance, between a person's hand and his own body when he directs his arm toward the camera. Methods that attempt to segment people based on depth gradients therefore often have trouble separating one person from another without also splitting individuals into pieces.

To combat all 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 these "plan-view" images, the representations of people are highly amenable to accurate spatial localization and tracking under diverse and challenging conditions. Section 2 describes the computation of plan-view statistics from depth data in greater detail, motivates their use in person detection and tracking, and outlines the context of previous work in this area.

In this paper, we introduce a new combination of plan-view statistics that better preserves object shape information than prior approaches, and therefore provides superior features for tracking. We also present 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, which provide a more detailed description of tracked people than the models used by prior plan-view methods. The improved person models allow for better tracking through complex inter-person occlusions and close interactions, among other benefits. We also illustrate how the typical problems with adaptive template tracking in camera-view images are easily avoided in our plan-view framework. Our method enables high-quality, multi-person tracking with a single, compact, static stereo camera unit that may be mounted on walls, ceilings, furniture, or desktop computers. However, the method is also well-suited for incorporation into a multi-unit system that tracks people throughout arbitrarily large, complex spaces. The method operates in an online mode, rather than as a batch process, and therefore is appropriate for real-time applications.

The remainder of the paper proceeds as follows. Section 2 leads up to the introduction of our plan-view statistical substrate, whose computation is described in Section 3. Section 4 details an approach for detecting and tracking people in this substrate, using Kalman filtering on adaptive statistical templates. In Section 5, we highlight a means of dealing with common adaptive template issues such as drift and choice of template size. Section 6 discusses the high-quality tracking results obtained by our method, and compares these results to those obtained when more limited statistics or person models are used.

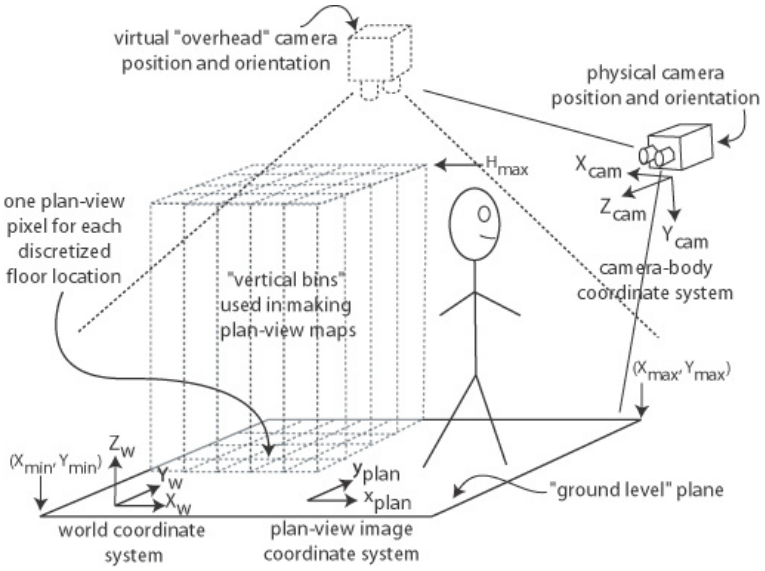

Fig. 2. To make a plan-view map, we project foreground into a point cloud in the 3D camera body coordinate system, and then rotate our view of this point cloud to a virtual, overhead camera position. The space is then divided into vertical bins aligned with an axis pointed toward the sky, and one plan-view pixel value is computed for each vertical bin.

2 Plan-View Statistics

The motivation behind using plan-view statistics for person tracking begins with the observation that, in general, people tend to not overlap much in the dimension normal to the ground. That is, in most contexts in which person-tracking might be used, people typically 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. Several person tracking systems that rely on monocular video exploit this idea and are designed to operate with cameras mounted in this way [10,11]. However, such methods 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. When tracking is required in an outdoor environment, an overhead camera mount may not even be feasible.

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 overhead camera. In addition, we can produce these

"plan-view" 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 different viewing angles and camera projection models, can be computed. Plan-view images are just one possible class of images that may be constructed, but are among those with the greatest utility for person tracking.

Figure 2 illustrates the basic principles and coordinate systems underlying the transformation of camera-view depth images into plan-view statistical images. 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, in the $X_{cam}Y_{cam}Z_{cam}$-coordinate frame of the stereo camera, representing the portion of the scene visible to the camera. We would like to analyze the 3D point cloud in terms of a world $X_WY_WZ_W$-coordinate system in which the Z_W-axis is aligned with the "vertical" axis of the world - that is, the axis from the center of the Earth toward the sky, normal to the X_WY_W-ground-level-plane in which we expect people to be well-separated. (For simplicity, one may assume in this discussion that the true scene ground is planar, but in Section 3 we discuss a simple modification of our method that compensates for scenes in which it is not.) We therefore select such a coordinate system (which involves choosing a ground level plane, a world origin in the plane, and the directions of the X_W- and Y_W-axes in the plane), and then measure the stereo camera's location and orientation within it. This effectively tells us how to move the real, physical stereo camera into a virtual, overhead, downwardly-directed configuration, thereby aligning the 3D point cloud in the frame of the stereo camera with the world coordinate frame.

Although one might implement a person tracking algorithm that operates directly on this point cloud, we prefer to reduce the data dimensionality by finding 2D views or projections of it that are well-suited for person tracking. We therefore discretize the 3D world space into vertical bins extending along the Z_W-axis and intersecting the X_WY_W-plane in a regular grid. 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.

Plan-view projection of per-pixel depth from stereo has been applied to person detection and tracking by Beymer [12], Darrell et. al. [13], and by researchers at Interval Research Corp. [14]. 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 plane. Although powerful, this representation discards virtually all object shape information in the vertical (Z_W) 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 [14], and plan-view height maps from stereo have been used in other contexts such as automatic military target recognition [15] and path-planning for the Mars rover [16]. 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 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.

Like occupancy maps, however, height maps are susceptible to some problems. For instance, the movement of relatively small objects at heights similar to those of people's heads, such as when a person places a book on an eye-level shelf, can appear similar to person motion in a height map. Also, use of the highest point within each vertical bin, rather than height-rank-filtering for the point at perhaps the 90th-percentile, allows for fast computation of height maps, but makes these maps very sensitive to depth noise. The effects of depth noise are often severe enough to counteract any benefits of constructing plan-view height maps.

We largely overcome both of these problems, as well as those associated with occupancy maps, through the novel strategy of 1) using occupancy statistics to refine the height map, and 2) using occupancy and height statistics together for tracking. This approach, described in Section 3, creates a rich and robust plan-view basis for tracking. Sections 4 and 5 go on to describe a tracking method, based on Kalman prediction on adaptive statistical templates, that seeks to leverage as much of the detail in our plan-view maps as possible for use as features in tracking.

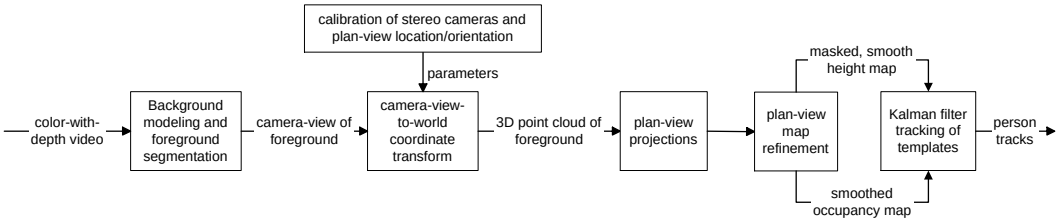

Fig. 3. Overview of person tracking algorithm.

3 Building Maps of Plan-View Statistics

An overview of our full person tracking method is shown in Figure 3. In this section, we describe all the steps prior to the rightmost block. These steps transform and refine camera-view color and depth data to create the plan-view image input upon which the tracking methods of Sections 4 and 5 are based.

3.1 Camera Setup and Video Input

The input to the 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 4**a** and 4**b**. We use depth instead of disparity for two reasons. First, operations carried out in a metric depth space are usually easier to understand, are more directly related to physically determinable parameters, and are often simpler to compute than equivalent operations in a disparity data space. This should become clearer as we explain our method. Second, we would like our method to be applicable in systems that compute depth not only by image area-matching techniques, but also by methods based on other means, such as lidar, that do not produce disparities.

When multi-camera stereo is used to provide color and depth, calibrations must be performed to determine three mappings between coordinate systems that will be used in constructing plan-view images:

1. Calibration of each individual camera's intrinsic parameters and lens distortion function is needed to map each camera's raw, distorted input to images that are suitable for stereo matching.
2. Stereo calibration and determination of the cameras' epipolar geometry is required so that the set of individual cameras may be treated as a single virtual camera head producing color-with-depth video. More specifically, we must find the parameters for mapping disparity image values $(u, v, disp)$, using perspective back-projection, to 3D coordinates $(X_{cam}, Y_{cam}, Z_{cam})$ in

Fig. 4. Example camera-view input. From left to right, (a) Current color, (b) Current depth, (c) Foreground color. Unreliable (low confidence) depth data is shown in black.

the frame of the camera body. The perspective equations relating these coordinate systems are given in Section 3.3, but we note here that the parameters required from this calibration step are the camera baseline separation b, the virtual camera horizontal and vertical focal lengths f_u and f_v (for the general case of non-square pixels), and the image location (u_o, v_o) where the virtual camera's central axis of projection intersects the image plane.

3. The rigid transformation relating the $X_{cam}Y_{cam}Z_{cam}$ camera body coordinate system to the $X_WY_WZ_W$ world space must be determined so that we can align the 3D point cloud properly with the vertically-oriented bins and the ground level plane. We specifically seek the rotation matrix $\mathbf{R_{cam}}$ and translation vector $\vec{t}_{cam}$ required to move the real stereo camera into alignment with an imaginary stereo camera located at the world origin and with X_{cam}-, Y_{cam}-, and Z_{cam}-axes aligned with the world coordinate axes.

Many standard methods exist for accomplishing these calibration steps, and any two or more of the above steps can be combined into a single parameter optimization process. Since calibration methods are not our focus here, we do not describe particular techniques, but instead set forth the requirements that, whatever methods are used, they result in the production of distortion-corrected color-with-depth imagery, and they determine the parameters $b, f_u, f_v, (u_o, v_o), \mathbf{R_{cam}}$, and $\vec{t}_{cam}$.

To maximize the volume of viewable space without making the system overly susceptible to occlusions, we prefer to mount the stereo camera at a relatively high location, with the central axis of projection roughly midway between parallel and normal to the X_WY_W-plane. Although our stereo camera typically consists of a planar array of imagers with a stereo baseline on the order of 10-20cm, our method is applicable for any positioning and orientation of the monocular cameras, provided that the above calibration steps can be performed accurately. We use lenses with as wide a field of view as possible, provided that the lens distortion can be well-corrected.

3.2 Foreground Segmentation

Rather than use all of the image pixels in building plan-view statistical maps, we restrict our attention to objects in the scene that are novel or that move in ways that are atypical for them. These objects are segmented in the camera-view space via a relatively sophisticated technique for background estimation and removal, as detailed in [17,18]. In this method, the recent history of observations is modeled independently at each pixel, using Time-Adaptive, Per-Pixel Mixtures Of Gaussians (TAPPMOGs) in a four-dimensional pixel observation space of depth, luminance, and two chroma components. An online approximation to Expectation-Maximization is used to adapt the Gaussian mixture parameters as new image observations arrive, with observations corresponding to older images receiving less weight in the modeling process. A subset of the Gaussians in each pixel's mixture model is selected at each time step to represent the background. At each pixel where the current color and depth are well-described by that pixel's background model, the current video data is labeled as background. Otherwise, it is labeled as foreground.

Connected components analysis is used to remove small, isolated foreground regions and to fill small foreground holes, but we have found in practice that this is not critical to the success of our person tracking method. Figure 4**c** shows an example result of foreground extraction. The imprecision in the foreground edges is due to depth noise, but does not significantly affect person tracking performance.

"High-level" feedback is also used to further refine and guide the pixel-level segmentation. The background model is not updated where the person tracker believes that people are present, so that people who remain relatively static for long time periods are not slowly incorporated into the background model. Most of the remaining foreground is assumed to not correspond to people, and is regarded as foreground segmentation "errors". Rapid lighting changes are also detected, and the foreground associated with these changes are also regarded as errors. The distribution of all such errors over time are, like the observation history, modeled using a TAPPMOG scheme. The error TAPPMOG and observation history TAPPMOG are periodically merged, as described in [17], so that the background model better reflects these errors, thereby decreasing the likelihood that they will be segmented as foreground again in the future. This error-correction via high-level feedback allows for rapid recovery (in less than 2 seconds) from rapid lighting changes, and allows quick adaptation of the background model to uninteresting scene changes (such as the moving of the chair), so that few of these types of events provide substantial distraction to the person tracking system.

3.3 Plan-View Map Construction

All camera-view image foreground pixels with reliable depth measurements are used in building our plan-view statistical maps. This process begins with construction of a 3D point cloud representing the foreground in the scene. For a binocular stereo pair, two coordinate transforms are needed to convert disparity image data to a point cloud aligned with our $X_W Y_W Z_W$ world coordinate system. First, we 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:

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

These equations assume a disparity image coordinate system in which the u- and v-axes are oriented left-to-right along image rows and top-to-bottom along image columns, respectively. The monocular cameras are assumed to be separated along the image u-axis. In the camera body coordinate frame, the origin is at the camera principal point, the X_{cam}- and Y_{cam}-axes are coincident with the disparity image u- and v-axes, and the Z_{cam}-axis points toward the scene imaged by the camera.

Next, we transform these camera frame coordinates into the (X_W, Y_W, Z_W) world space by applying the rotation $\mathbf{R_{cam}}$ and translation $\vec{t}_{cam}$ relating the two coordinate systems:

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

Before building plan-view maps from the 3D point cloud, we must choose a resolution at which to quantize the world 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 use vertical bins that intersect the $X_W Y_W$-ground-level-plane so as to divide it into a square grid with resolution δ_{ground} of 2-4cm/pixel.

We must also choose the bounds $(X_{min}, X_{max}, Y_{min}, Y_{max})$ of the ground level area within which we will restrict our attention. This can be done by first intersecting the stereo camera's field of view with the full volume of space within which portions of people might be present, and then projecting this intersection volume onto the $X_W Y_W$-plane. The volume in which people may exist is restricted by the floor and walls of the environment, as well as by the surface at height H_{max} above the ground, where H_{max} is an estimate of the maximum height at which we might expect to see human body parts (e.g. how

high a very tall person might reach with his hands if he stands on his toes).

The ground plane discretization and bounds determine the size of our plan-view images, and allow us to map 3D point cloud coordinates to their corresponding plan-view image pixel locations (x_{plan}, y_{plan}) as follows:

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

We can now describe how to compute plan-view height and occupancy maps, denoted as $\mathcal{H}$ and $\mathcal{O}$ respectively, in a single pass through the camera-view foreground data. Specifically, after setting all pixels in both plan-view maps to zero, we do the following for each camera-view foreground pixel:

1. Compute the pixel's $(X_{cam}, Y_{cam}, Z_{cam})$-coordinate via equation (1).
2. Compute the pixel's (X_W, Y_W, Z_W)-coordinate via equation (2).
3. Compute the pixel's plan-view image location (x_{plan}, y_{plan}) via equation (3).
4. If $Z_W > \mathcal{H}(x_{plan}, y_{plan})$ and $Z_W < H_{max}$, then update the height map: $\mathcal{H}(x_{plan}, y_{plan}) = Z_W$.
5. Increment $\mathcal{O}(x_{plan}, y_{plan})$ by $Z_{cam}^2/f_u f_v$, which is an estimate of the real area subtended by the pixel at distance Z_{cam} from the camera.

The scaling of the increments to the occupancy map compensates for the dependence on distance of an object's size in a camera-view image. If our plan-view occupancy map contained simply the count of pixels in each vertical bin in the world, the "pile of pixels" representing an object would grow in size as the object more closely approached the camera, since the object would subtend more pixels. By instead using the sum of $Z_{cam}^2/f_u f_v$ over the points in each vertical bin, we make an object's representation in the occupancy map relatively independent of distance to the camera, and we tie occupancy map values to a physically meaningful quantity - namely, the total metric surface area of foreground visible to the camera within each vertical bin.

Plan-view height and occupancy maps corresponding to the foreground of Figure 4 are shown in Figures 5**a** and 5**b**, respectively. The plan-view location of the stereo camera is at the middle of the bottom of these images, and the lines indicate the bounds of the camera's field of view when projected onto the ground plane. In the occupancy map, five "blobs" corresponding to the five people in the foreground image are clearly visible. The two people facing each other and talking to each other correspond to the two lowermost blobs in the maps. In the height map, the data is less clearly separable into five distinct blobs.

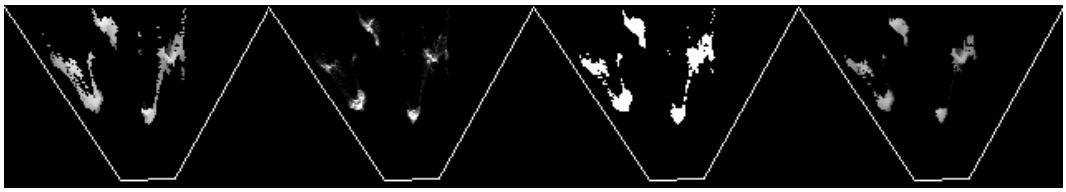

Fig. 5. Plan-view maps corresponding to the extracted foreground of Figure 4. The plan-view location of the stereo camera is near the bottom middle of the image, and the lines indicate the camera field of view. From left to right, images are **(a)** Raw height map $\mathcal{H}_{raw}$, **(b)** Raw occupancy map $\mathcal{O}_{raw}$, **(c)** Bitmap indicating where smoothed occupancy is above threshold θ_{occ}, **(d)** Masked, smoothed height map $\mathcal{H}_{masked}$.

If the true ground in the scene is not planar, we use the depth data associated with the scene background model to construct a "height-offset" map $\mathcal{H}_o(x_{plan}, y_{plan})$. This map estimates the deviation from planarity of the ground at all locations in the plan-view images. When building the plan-view height map for a given frame of foreground, we first subtract $\mathcal{H}_o(x_{plan}, y_{plan})$ from each Z_W value before comparing it against H_{max} and $\mathcal{H}(x_{plan}, y_{plan})$.

3.4 Plan-View Map Refinement

Given the rather substantial noise that is typical of depth imagery, it should not be surprising that the plan-view maps constructed as described above are also quite noisy. We therefore denote these maps as the "raw" data $\mathcal{H}_{raw}$ and $\mathcal{O}_{raw}$, and we convolve them with a Gaussian kernel to produce "smooth" maps $\mathcal{H}_{sm}$ and $\mathcal{O}_{sm}$. Because plan-view map coordinates correspond to a metric space (measured in centimeters in our case), and because we know the typical metric size of the objects (people) in which we are interested, we can make an intelligent choice of spatial extent for the Gaussian smoothing operator. In particular, we choose a Gaussian variance in pixels that, when multiplied by the map resolution δ_{ground}, corresponds to a physical size on the order of 1-4cm. This smooths depth noise in person shapes, while retaining gross features like arms, legs, and heads.

Although simple Gaussian smoothing produces relatively clean plan-view occupancy maps, it is inadequate for dealing with the substantial noise in our raw height maps. Our $\mathcal{H}_{raw}$ contains the maximum height statistic for each vertical bin; while this statistic may be computed very efficiently, it is also very sensitive to depth image noise. When this noise is located at "interesting" heights, such as those typical of heads of upright people, it can severely disrupt tracking. However, even if a more robust, rank filtered height statistic were used, and even if this statistic succeeded in completely eliminating height map noise, tracking in plan-view height maps would still be confused by the

movement of small, non-person foreground objects at these same "interesting" heights. For example, when a person places a sweater on an eye-level shelf, its appearance in a height map can resemble that of the person who placed it there, since the height map gives only a partial indication of whether or not there is a human body beneath any head-level object. These problems help explain why no prior person tracking methods attempt to make use of plan-view maps based on only height statistics.

A critical innovation in our method, therefore, is its use of the much less noisy occupancy map statistics to refine the height maps so they can be used in tracking with an acceptable level of confidence. Specifically, we propose to use the smoothed height map statistics only in floor areas where something "significant" is present, as indicated by the amount of local occupancy map evidence. In other words, data in the height map is largely ignored where the corresponding region of the occupancy map fails to indicate that a large, person-sized foreground object is in the vicinity. This criterion may be implemented in a number of ways, but here we use the relatively simple technique of pruning $\mathcal{H}_{sm}$ (setting it to zero) wherever the corresponding pixel in the smoothed occupancy map $\mathcal{O}_{sm}$ is below a threshold θ_{occ}. We set the threshold θ_{occ} to a relatively low value, because we do not want to remove from the height map all evidence of people who are substantially occluded. Even with a low value of θ_{occ}, the refinement of the height map is significant. One formula for θ_{occ} based on physically measurable quantities is described in the Appendix.

Figure 5**c** shows the mask obtained by applying the threshold θ_{occ} to the smoothed occupancy map of the foreground of Figure 4. The result of applying this mask to $\mathcal{H}_{sm}$ is shown in Figure 5**d**. This masked height map $\mathcal{H}_{masked}$, along with the smoothed occupancy map $\mathcal{O}_{sm}$, provide an excellent new basis on which to build person detection and tracking algorithms. In Sections 4 and 5, we describe an example method that, as indicated by the results of Section 6, is particularly well-tailored to this basis.

4 Tracking and Adapting Templates of Plan-View Statistics

In the plan-view person tracking method of Beymer [12], people are modeled with Gaussians applied to occupancy maps. Darrell et. al. [13] use an even simpler model, namely the integral of the occupancy data within some plan-view support region, and implement tracking as a batch process that relies on dynamic programming to produce results for an entire sequence of video frames at once. In this section, we describe a novel technique that employs much more descriptive, template-based person models that can adapt quickly over time. Furthermore, our method allows for online, real-time detection and tracking of people with good accuracy. The detection and tracking components

of our approach are detailed in Sections 4.1 and 4.2, respectively.

4.1 Person Detection

We begin to detect a new person in the scene by looking for a significant "pile of pixels" in the occupancy map that has not been accounted for by tracking of people found in previous frames. More precisely, after tracking of known people has been completed, and after the occupancy and height evidence supporting these tracked people has been deleted from the plan-view maps, we convolve the occupancy map $\mathcal{O}_{sm}$ with a box filter and find the maximum value of the result. If this peak value is above a threshold θ_{newOcc}, we regard its location as that of a candidate new person. The box filter size is again a physically-motivated parameter, with width and height equal to an estimate of twice the average shoulder-to-shoulder torso width W_{avg} of adult people. We use a value of W_{avg} around 40cm.

We apply additional tests at the candidate person location to better verify that this is a person and not some other type of object. Currently, we require that two simple tests be passed:

1. The highest value in $\mathcal{H}_{masked}$ within a square of width $2 * W_{avg}$ centered at the candidate person location must exceed some plausible minimum height θ_{newHt} for people.
2. Among the camera-view foreground pixels that map to the plan-view square of width $2 * W_{avg}$ centered at the candidate person location, the fraction of those whose luminance has changed significantly since the last frame must exceed a threshold θ_{newAct}.

These tests ensure that the foreground object is physically large enough to be a person, and is more physically active than, for instance, a statue. However, these tests sometimes exclude small children or people in unusual postures, and sometimes fail to exclude large, non-static, non-person objects such as foliage in wind. We are therefore in the process of implementing more sophisticated tests to be applied to the shape data available in the plan-view images. In addition, some of these errors might be avoided by restricting the detection of people to certain entry zones in the plan-view map.

Whether or not the above tests are passed, we delete, after the tests have been applied, the height and occupancy map data within a square of width $2 * W_{avg}$ centered at the location of the box filter convolution maximum. We then apply the box filter to $\mathcal{O}_{sm}$ again to look for another candidate new person location. This process continues until the convolution peak value falls below θ_{newOcc}, indicating that there are no more likely locations at which to check for newly occurring people.

In detecting a new person to be tracked, our philosophy is that we would like to see him without substantial occlusion for a few frames before we officially add him to our "tracked person" list. We therefore aim to set the new person occupancy threshold θ_{newOcc} so that half of an average-sized person must be visible to the stereo pair in order to exceed it. This is approximately implemented using $\theta_{newOcc} = \frac{1}{2} * \frac{1}{2} * W_{avg} H_{avg}$, where W_{avg} and H_{avg} denote average person width and height, and where the extra factor of $\frac{1}{2}$ roughly compensates for the non-rectangularity of people and the possibility of unreliable depth data. We also do not allow the detection of a candidate new person within some small plan-view distance (e.g. $2 * W_{avg}$) of any currently tracked people, so that our box filter detection mechanism is less susceptible to exceeding θ_{newOcc} due to contribution of occupancy from the plan-view fringes of more than one person. Finally, after a new person is detected, he remains only a "candidate" until he is tracked successfully for some minimum number of consecutive frames. No track is reported while the person is still a candidate, although the track measured during this probational period may be retrieved later.

4.2 Tracking with Plan-View Templates

Kalman filtering is used to track patterns of plan-view height and occupancy statistics over time. Much prior work on multi-target tracking with Kalman filters exists both within and outside the computer vision literature; see, for example, [19–21]. Also, Beymer [12] applies a Kalman framework to tracking multiple Gaussian models of people in plan-view occupancy maps. Our focus here is not on innovation in Kalman filtering, but rather on adapting the standard methodology to our plan-view statistical substrate. In the future, we are interested in exploring more powerful tracking methods such as particle filtering, which can estimate non-Gaussian, non-linear dynamic processes and which has recently been extended to track multiple targets [22,23].

The state representation and dynamical models underlying our Kalman filters are described in Section 4.2.1. Section 4.2.2 details the core of the algorithm for tracking people from one frame to the next. Finally, in Section 4.2.3, we outline simple "long-term" tracking methods that help compensate for possible failures in inter-frame tracking.

4.2.1 Kalman State and Prediction

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.

4.2.2 Kalman Measurement and Update Steps

The measurement step of the Kalman process is carried out for each person individually, in order of our confidence in their current positional estimates. This confidence is taken to be proportional to the inverse of $\sigma^2_{\vec{x}}$, the variance for the Kalman positional estimate $\vec{x}$. To obtain a new position measurement for a person, we search in the neighborhood of the predicted person position $\vec{x}_{pred}$ for the location at which the current plan-view image statistics best match the predicted ones for the person. The area in which to search is centered at $\vec{x}_{pred}$, with a rectangular extent determined from $\sigma^2_{\vec{x}}$. A match score $\mathcal{M}$ is computed at all locations within the search zone, with lower values of $\mathcal{M}$ indicating better matches.

The value of the match score $\mathcal{M}$ for the ith person at some plan-view location $\vec{x}$ is linearly proportional to four metrics that are easily understood from a physical standpoint:

1. The difference between the shape of the tracked person when seen from overhead, as indicated by the ith person's height template $\mathcal{T}_H$, and that of the current scene foreground, as indicated by the masked height map

$\mathcal{H}_{masked}$, in the neighborhood of $\vec{x}$.

2. The difference between the amount of the tracked person's visible surface area, as indicated by the ith person's occupancy template $\mathcal{T}_O$, and that of the current scene foreground, as indicated by the smoothed occupancy map $\mathcal{O}_{sm}$, in the neighborhood of $\vec{x}$.
3. The distance between $\vec{x}$ and the predicted person location $\vec{x}_{pred}$.
4. The closeness of $\vec{x}$ to the measured locations of all people previously tracked in this frame.

The final component attempts to enforce the physical principle that two people cannot occupy the same space at the same time, and discourages the matching of more than one person to nearly the same location in the plan-view maps. For the ith person, we compute the above match score at location $\vec{x}$ as follows:

$$\begin{aligned}\mathcal{M}(i,\vec{x}) = \; & \alpha * SAD(\mathcal{T}_H, \mathcal{H}_{masked}(\vec{x})) + \\ & \beta * SAD(\mathcal{T}_O, \mathcal{O}_{sm}(\vec{x})) + \\ & \gamma * \sqrt{(x - x_{pred})^2 + (y - y_{pred})^2} + \\ & \epsilon * \sum_{j<i} \eta\left(\vec{x}_j, W_{avg}, \vec{x}\right)\end{aligned} \tag{4}$$

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. $\eta(\ldots)$ 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 previously tracked person, and a variance equaling the average person torso width. Appropriate relative weightings for the various terms in equation (4) can be determined from physical principles, as discussed in the Appendix.

When comparing a height template $\mathcal{T}_H$ to a portion of $\mathcal{H}_{masked}$ via the SAD operation, it is not desirable to include differences at pixels where either $\mathcal{T}_H$ or $\mathcal{H}_{masked}$ has been masked out but the other has not, as this might artificially inflate the SAD score. We should not simply ignore these pixels, however, because we would like the SAD to be high (bad) when a data-rich template is compared with a largely empty portion of $\mathcal{H}_{masked}$. Therefore, we modify the SAD process, for the height comparison only, to substitute an artificial height difference wherever either, but not both, of the corresponding pixels of $\mathcal{H}_{masked}$ and $\mathcal{T}_H$ are zero. The artificial difference is chosen to be $H_{max}/3$, which is the expected difference of two random variables uniformly distributed between 0 and H_{max}.

If the best (minimal) match score falls below a threshold θ_{track}, we update the person's 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 image values of $\mathcal{H}_{masked}$ and $\mathcal{O}_{sm}$ in the area surrounding $\vec{x}_{best}$ are used as the new body configuration measurements for updating the templates. (The extent of this area is discussed in Section 5.) This image data is cleared before tracking of another person is attempted. A relatively high Kalman gain is used in the template update process, so that templates adapt quickly. To reduce computation, one might simply copy the current plan-view map statistics into the person templates $\mathcal{T}_H$ and $\mathcal{T}_O$, rather than maintain Kalman covariance matrices for them and use the standard Kalman update equations. We use this simplification, and have found in practice that it causes little degradation in tracking performance.

If the best match score from equation (4) is above θ_{track} for some person, we do not update that person's Kalman state with new measurements, and we report $\vec{x}_{pred}$ as the person's location. The positional state variances are incremented, reflecting our decrease in tracking confidence for the person. The person is also placed on a temporary list of "lost" people, which is used as described in the next section.

4.2.3 Long-Term Tracking Techniques

When someone exits the scene temporarily or is substantially occluded for an extended time, he may remain on the "lost" person list for many frames, with his position being reported solely on the basis of Kalman prediction. When the person becomes visible again, it is often the case that the Kalman position prediction will be far from correct, especially when the person has been lost for significant time. The person's plan-view templates will also probably not match well with his new appearance. We therefore resort to secondary, "long-term" tracking methods, as described below, in order to link the tracks of lost people to their reappearances in the scene.

After tracking (as described in Section 4.2.2) and new person detection (as described in Section 4.1) have been completed, we determine, for each lost person, whether or not any newly detected person is sufficiently close in space (e.g. 2 meters) to the predicted location of the lost person or to the last place he was sighted. If so, and if the lost person has not been lost too long, we decide that the two people are a match. In the future, we will apply more sophisticated matching criteria based on person shape and appearance. If the new and lost people are declared a match, we set the lost person's Kalman state to be equal to that of the newly detected person. If a lost person cannot be matched with any newly detected person, we consider how long it has been since the person was successfully tracked. If it has been too long (above some time threshold such as 4 seconds), we decide that the person is permanently lost, and we delete him from the list of people we are trying to track.

5 Avoidance of Adaptive Template Problems

A variety of tracking schemes based on adaptive templates have been used in prior, "camera-view" tracking systems; recent examples include [24,25]. As models of tracked objects, adaptive templates would seem to offer, in theory, the advantages of simplicity, flexibility, and descriptive power. In practice, however, adaptive template descriptions of objects have proven sufficiently problematic to cause them to be abandoned, in most current systems, in favor of parameterized models such as "blobs", Gaussians, or linear combinations of basis shapes. In this section, we discuss how the typical problems with camera-view, adaptive template tracking are easily side-stepped in our plan-view approach.

5.1 Choice of Template Size

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*. Specifically, we employ square templates whose sides have a length in pixels that, when multiplied by the plan-view map resolution δ_{ground}, is roughly equal to $2 * W_{avg}$. As discussed in Section 4.1, W_{avg} is an estimate of the average torso width (from shoulder to shoulder) of adult people, and we use $W_{avg} \approx 40$cm.

Use of a constant template size is reasonable because of a combination of two factors. First, people spend almost all of their waking time in a predominantly upright position (even when sitting), and the spatial extents of most upright people, when viewed from overhead, are confined to a relatively limited range. Second, our plan-view representations of people are, ideally, invariant to their floor locations relative to the camera. In practice, the plan-view statistics for a given person become more noisy as he or she moves away from the camera, because of the smaller number of camera-view pixels that contribute to them. Nevertheless, some basic properties of these statistics, such as their typical magnitudes and spatial extents, do not depend on the person's distance from the camera, so that no change in template size is necessitated by the person's movement around the room.

Our template width of $2 * W_{avg} \approx 80$cm is large enough to accomodate the torsos of nearly all upright people, as well as much of their outstretched limbs, without being overly large for use with small or closely-spaced people. For

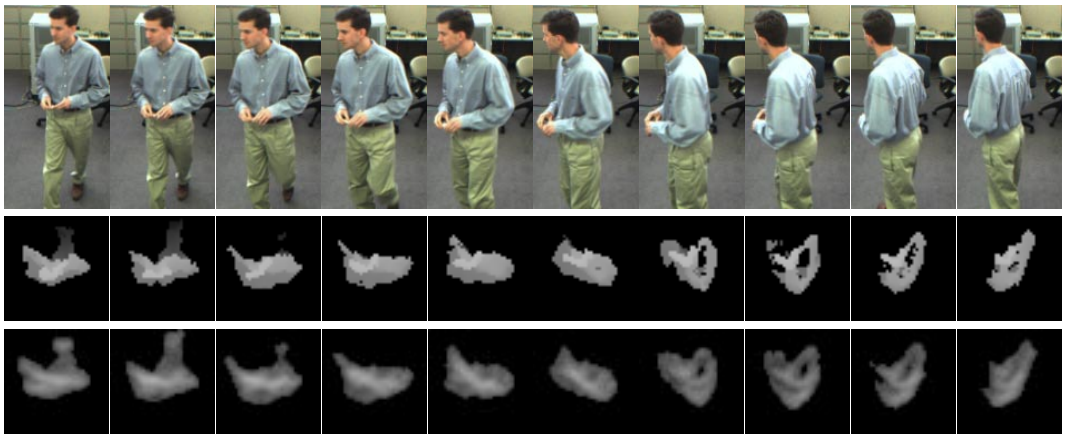

Fig. 6. Evolution of height and occupancy templates $\mathcal{T}_H$ and $\mathcal{T}_O$ for a single tracked person over the course of 1.5 seconds (every other frame shown). **Top row:** Color video of person taking a step toward camera and then turning to his right (depth not shown); **Middle row:** Extracted, re-centered height templates $\mathcal{T}_H$; **Bottom row:** Extracted, re-centered occupancy templates $\mathcal{T}_O$.

people of unusual size or in unusual postures, this template size still works well, although perhaps it is not ideal. As we develop more sophisticated methods of analyzing person shape and activity, we may allow our templates to adapt in size when appropriate.

5.2 Reduction of Template "Slippage"

Templates that are updated over time with current image values inevitably "slip off" the tracked target, and begin to reflect elements of the non-target image background. This is perhaps the primary reason that adaptive templates are seldom used in current tracking methods. Our method, as described thus far, also suffers from this template "slippage". However, with our plan-view statistical basis, it is relatively straightforward to counteract the problem in ways that are not feasible for other image substrates.

Specifically, we are able to virtually eliminate template slippage through a simple template "re-centering" technique that relies on our knowledge of typical person sizes in plan-view maps. On each frame, after tracking according to Section 4.2 has completed, we compute for each person the location in $\mathcal{O}_{sm}$ of the occupancy center-of-mass $\vec{x}_{com}$ for the pixels within a square of width W_{avg} centered at the person's estimated location. This provides an estimate for each person of where the bulk of his or her associated plan-view statistical data is currently centered. New templates $\mathcal{T}_H$ and $\mathcal{T}_O$ are then extracted from $\mathcal{H}_{masked}$ and $\mathcal{O}_{sm}$ at $\vec{x}_{com}$ for each person. Also, each person's location in the Kalman state vector is shifted to $\vec{x}_{com}$, without changing the velocity

estimates or other Kalman filter parameters. Re-centering is not applied to people whose best match score from equation (4) is above θ_{track}, and whose current location estimate is therefore simply the Kalman prediction.

We have found this re-centering technique to be very effective in keeping templates solidly situated over the plan-view statistics representing a person, despite depth noise, partial occlusions, and inter-person interactions. It also serves to well align the templates extracted for a given individual across successive frames. An example of the evolution in time of the extracted, re-centered height and occupancy templates $\mathcal{T}_H$ and $\mathcal{T}_O$ for a tracked person is shown in Figure 6. The robustness provided by the re-centering technique arises from its ability to use the average person size W_{avg} to constrain the search window for finding a corrected template location. Without this constraint, it is difficult to prevent templates for a given person from "jumping" to other nearby people and other foreground objects in the plan-view maps.

6 Experimental Results

We have implemented our method in C++ on a standard 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 [1]. 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. [5], would therefore dramatically improve the system speed, and thereby lead to gains in real-time person tracking performance.

The sophisticated background estimation and removal method of [17] is the next most costly system component, and hence the frame rate could also be increased by substituting simpler methods that, for instance, rely on depth alone [26,27], assume a static background, or employ simpler per-pixel background models. We have found our tracking results to be largely unaffected by minor foreground errors such as holes in foreground objects, imprecise object boundaries, shadows, and isolated regions of foreground noise, so use of simpler background removal methods is not detrimental in many situations. This is due largely to our use of plan-view image statistics rather than camera-view foreground analysis. However, simpler foreground estimation methods

Fig. 7. Example frames from tracking test sequences. Tracking was successful in the leftmost four frames, but failures occurred on the rightmost two. From left to right: **(a), (b):** Typical frames. **(c):** Three people lined up in the direction away from camera; note successful tracking of most distant person (inside square) behind cubicle wall. **(d):** Multi-person occlusions and non-person foreground object (chair). **(e):** Failure occurs where three people are in frame, but third is completely occluded by another (inside box, not visible) for a long period of time. **(f):** Failure occurs where person (inside box) crouches behind a newly opened cabinet door while multiple people pass in front and nearby.

often fail to adequately handle global lighting changes, slow-moving or stationary foreground objects, dynamic background objects such as foliage in wind, movement or addition of "background" objects such as chairs, and other phenomena. The resulting erroneous omission or inclusion of large foreground objects may substantially degrade the performance of our, and most other, tracking methods. Hence, for systems intended to run for long periods of time in real-world conditions, we recommend use of robust, adaptive foreground extraction, despite the extra computation this may entail.

We have quantitatively evaluated our method on several color-with-depth video sequences captured at 12-15Hz and 320x240 resolution. Across all sequences, the stereo camera was statically mounted between 2.2m and 3m above the ground, with a view like that shown in Figure 4. 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 actions are observable.

Despite these challenges, and without requiring extensive parameter tuning, our method made only two significant errors. In one case, a person who hid

Table 1: Performance Comparison for Tracking Methods

Person Model	*Plan-View Statistics Used*		
	Occ+Height	Height Only	Occupancy Only
Adaptive Template	2	10	13
Gaussian	17	33	28

Numbers indicate count of "significant" errors made, as defined in Section 6, on a set of challenging test sequences.

behind another was not correctly linked to her re-appearance after a long time; in the other, confusion occurred when a person crouched behind a cabinet door opened by another occluding person. Use of secondary, long-term tracking techniques more sophisticated than those described in Section 4.2.3 would likely remedy these problems. Nevertheless, 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. Several frames from the test sequences are shown in Figure 7, including frames indicating the two failures. An example movie of results may be found at http://www.hpl.hp.com/personal/Michael_Harville/movies/ptrack02_1.mpg.

To better assess the value of the combination of plan-view height and occupancy statistics for tracking, we compared our results against those obtained for modified versions of our method that omit either height or occupancy. Without occupancy data, only height templates are used in equation (4), and we cannot prune noise and small foreground objects from our height maps as described in Section 3.4. In addition, we must use height data alone to detect new people, and the template re-centering method of Section 5.2 is modified to move templates to the local (within W_{avg}) peak in the height data. Alternatively, when height data is omitted from tracking, only occupancy templates are used in equation (4), and the "minimum height" test of Section 4.2 is not used in distinguishing whether a plan-view "pile of pixels" is a person or some other type of foreground object.

The first row of Table 6 shows that tracking performance declines significantly when either occupancy or height data is neglected. For the same challenging test sequences on which our method made only a single error, 10 significant errors occurred when height statistics alone were used, and 13 such errors occurred when only occupancy data was used. Many of these errors involved people far from the camera, where noise in the plan-view statistics becomes more significant. Other errors included confusion of rolling chairs with people when height data was not used to distinguish the two, and swapping of identities of closely interacting people when occupancy was not used to prune height map noise.

We also sought to evaluate the merit of our choice of template-based person models. To this end, we compared our results to those obtained when a simple parametric model of a person, namely a plan-view Gaussian, is used instead. Tracking of Gaussian mixtures on height and occupancy statistics was implemented primarily through replacement of the Kalman measurement step of equation (4) with a method for fitting competitive mixtures of Gaussians to plan-view data, similar to that used in [12]. For completeness, we applied this technique to occupancy data alone, height data alone, and to the combination of the two. When both height and occupancy were used, separate Gaussian fits were done on each statistic in isolation, and the mean of the two results was used as the new measurement.

The second row of Table 6 shows that the simpler, Gaussian person model does not perform as well as the adaptive template. The number of significant errors increased for all choices of plan-view statistical substrate, and the percentage of this increase was especially high for the combination of height and occupancy data. Many of the new errors occur when close interactions and substantial inter-person occlusions happen, as Gaussians for different people would sometimes swap places, or would attach to parts of the same person and other objects the person might be moving (such as a chair). It is possible that the Gaussian performance could be improved through modification of the fitting process, but we believe the template-based approach offers a simple, fast alternative.

Finally, we have attempted to estimate the positional accuracy of our method's tracks, by comparing result tracks to manually estimated ground truth. For each of 8 person tracks obtained from the above test sequences 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.

Figure 8 shows some example comparisons of result tracks against manually estimated ground truth. The rightmost example shows a very long (nearly 1 minute) track that maintained its integrity despite multiple inter-person occlusions, the person's brief scene exit, and his temporary, complete occlusion by a static object.

In general, across all testing, we found the method to be relatively insensitive to the quality of both the foreground segmentation and the depth data itself. We observed good tracking performance despite significant holes and noise in depth and foreground images. Also, the method proved very adept at linking

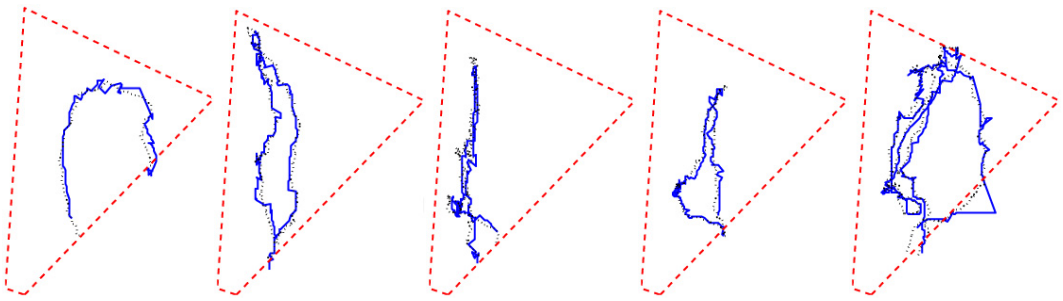

Fig. 8. Comparisons of five individual result tracks with estimated ground truth. Results are shown in solid lines, ground truth in dotted lines. Field of view of stereo pair at the $\mathcal{H}_{max}$ level is shown in dashed lines. Plan-view position of stereo pair is near lower-left corner of the images. Ground truth was obtained by manual labeling of raw plan-view height maps.

the tracks of "lost" people to the correct new sightings of these people when they re-emerge from a complete occlusion. This is due in large part to the success of Kalman trajectory prediction in the plan-view image space.

Much work has been done recently to compare the performance of person tracking systems operating on color or grayscale video from one or more widely spaced cameras. This has been made possible largely through the collection and publicization of standard monocular test image sequences, together with the organization of forums for presenting comparison results [28]. Unfortunately, at the present time, these standard data sets do not include color-with-depth test sequences, so it is difficult to compare our method directly to the best monocular techniques. Instead, we have attempted to demonstrate significant improvement over the best methods operating on color-with-depth video. Because monocular methods operate on only a subset of the data used by ours, we expect that the proper combination of techniques from the best monocular systems and from our method (or related plan-view trackers) would produce a new method whose performance significantly surpasses that of all. We plan to continue work in that direction.

7 Conclusions

We have presented a technique that allows for accurate measurement and tracking of the 3D spatial locations associated with events involving people. 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 believe that this transformation of the depth data is sufficiently clean and compelling that many standard tracking methods could be built upon it and would obtain state-of-the-art performance.

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 more robust performance under challenging conditions, such as in crowded environments with many inter-person occlusions and close interactions, than do methods that rely upon more limited scene statistics or person models. Also, little tuning of system parameters is required, as reasonable values for most can be computed from physically measurable values, such as depth sensor resolution or average person size, via simple formulae.

A Appendix: Relation of System Parameters to Physical Quaniti-ties

A common problem with many complex engineering systems, including vision-based person detection and tracking systems, is their dependence on and sensitivity to the values of a large number of tunable parameters. As the number of system parameters increases, the capabilities of the system grow, but so may the effort required to determine a set of parameter values that allow for satisfactory system operation in one or more contexts of interest. When adapting a system to operate in a new context, it is highly desirable that the process of selecting system parameters not require substantial human labor.

The preceding description of our person detection and tracking method enumerated its reliance on a number of tunable parameters. Unlike many other such tracking systems, however, our method's operation in a metric, plan-view space allows reasonable values for nearly all system parameters to be expressed in terms of simple formulae involving physically measurable quantities. Hence, when adapting our person tracking system to a new context, little work is required other than the sorts of physical camera calibration described in Section 3.1.

Our method for choosing θ_{occ}, the threshold for converting the smoothed occupancy plan-view map $\mathcal{O}_{sm}$ into a bitmap for masking out noise in the height map, illustrates our reliance on physical principles in setting system parameters. First, one should recall that the values in the occupancy map reflect the visible surface area of foreground objects within each vertical bin of world space. Next, we note that the surface area of one face of a vertical bin, from the floor to the maximum person height H_{max}, is simply $\delta_{ground} H_{max}$. Finally, we derive a formula for θ_{occ} by setting forth the requirement that, for us to consider the occupancy of some vertical bin to be "significant", at least some

fraction κ of the viewable area of the bin must be occupied by foreground:

$$\theta_{occ} = \kappa * \delta_{ground} H_{max} \tag{A.1}$$

Of course, κ itself is a tunable parameter, and we are now left with the task of choosing its value instead of that for the original parameter θ_{occ}. The above equation, however, provides us with a rational justification for reasonable values of κ, which we did not have for θ_{occ}. For instance, a choice of $\kappa = 0.05$ provides reasonable performance, and comes with the interpretation that we require a relatively low fraction of a vertical bin to be occupied before we consider this occupancy to be "significant".

The match score equation (4),

$$\begin{aligned}\mathcal{M}(i, \vec{x}) = & \alpha * SAD(\mathcal{T}_H, \mathcal{H}_{masked}(\vec{x})) + \\ & \beta * SAD(\mathcal{T}_O, \mathcal{O}_{sm}(\vec{x})) + \\ & \gamma * \sqrt{(x - x_{pred})^2 + (y - y_{pred})^2} + \\ & \epsilon * \sum_{j<i} \eta\left(\vec{x}_j, \frac{1}{2} W_{avg}, \vec{x}\right)\end{aligned}$$

used in tracking person templates from one frame to the next, also requires several parameters to be selected, namely the relative weightings among the different equation terms. After choosing an arbitrary value (e.g. 1) for α, we scale the other weights as follows:

$$\begin{aligned}\beta &= \alpha * \lambda_1 / \delta^2_{ground} \\ \gamma &= \alpha * H_{max} / 9\sigma_{\vec{x}} \\ \epsilon &= \alpha * H_{max} / 3\end{aligned} \tag{A.2}$$

Again, these formulae arise from criteria based on physical principles. Our formula for β arises from our desires to 1) give height and occupancy equal weighting in the matching process, and 2) make the scaling of α to β insensitive to our choice of δ_{ground} for plan-view spatial discretization. If we note that occupancy SAD levels increase with the square of δ_{ground}, while height SAD levels are relatively independent of it, we see that we should make the ratio of α to β proportional to δ^2_{ground}. The proportionality constant λ_1 must only be calibrated once for a given experimental setup (perhaps by comparing peak occupancy map values for people with their peak height map values, across different people and floor locations), and depends on factors such as the imager resolution and the depth noise level. In practice, calibration of λ_1 for a given stereo camera only needs to be done once, regardless of changes in its mounting or the external environment.

The expression for ϵ is derived from our desire that, when the proposed position $\vec{x}$ of the ith person is precisely that of $\vec{x}_j$ for some jth person previously tracked in this frame, the value of the fourth term is comparably large, and therefore comparably discouraging of a match, as is the first term in equation (4) when the height template match is extremely poor. As described in Section 4.2.2, the expected difference of two random variables uniformly distributed between 0 and H_{max} is $H_{max}/3$, and the SAD is averaged over the number of pixels used in creating the difference. Hence, the expected value of our SAD measure between two completely different templates is $H_{max}/3$. When $\vec{x} = \vec{x}_j$ for some Gaussian in the fourth term, the fourth term will evaulate to 1, and so we wish to scale this by $H_{max}/3$ to make it comparable to a bad height template match.

We also attempt to set γ such that when the person is "far" from his Kalman-predicted position, the contribution of the third term above is comparable to that supplied by the first when the height template match is very poor. Hence, we again wish to scale γ such that this term is around $H_{max}/3$ when the proposed location $\vec{x}$ is far the predicted location $\vec{x}_{pred}$. We define "far" dynamically for each person at each time step, as being 3 times the standard deviation $\sigma_{\vec{x}}$ in the Kalman-predicted position. Hence, we make γ proportional to $(H_{max}/3)/3\sigma_{\vec{x}}$. In practice, we sometimes lower this proportionality constant, so that this distance term has a lesser effect on tracking than do the template terms.

Acknowledgments

Much of this work was inspired by the author's experience in using the real-time stereo depth system developed by John Woodfill, Harlyn Baker, and others at Interval Research Corporation. In helping build, while at Interval, the first real-time person tracking system to use multiple stereo cameras and plan-view occupancy maps (to be described in a forthcoming publication), the author gained from his colleagues many insights that helped shape his understanding of concepts important to this work. The author would especially like to thank Gaile Gordon, John Woodfill, Ali Rahimi, Trevor Darrell, Harlyn Baker, Michael Coffey, Tim Allen, Paul Regier, and Lieven Leroy.

References

[1] Point Grey Research.
URL `http://www.ptgrey.com`

[2] R. Gvili, A. Kaplan, E. Ofek, G. Yahav, Depth keying, in: SPIE Electronic Imaging, Vol. 5006, Santa Clara, California, 2003, pp. 564–574.

[3] K. Konolige, Small vision systems: hardware and implementation, in: 8th International Symposium on Robotics Research, Hayama, Japan, 1997, pp. 111–116.

[4] P. Mengel, G. Doemens, L. Listl, Fast range imaging by CMOS sensor array through multiple double short time integration (MDSI), in: International Conference on Image Processing (ICIP), Vol. 2, Thessaloniki, Greece, 2002, pp. 169–172.

[5] Tyzx Inc.
URL `http://www.tyzx.com`

[6] D. Beymer, K. Konolige, Real-time tracking of multiple people using continuous detection, in: ICCV Frame-rate Workshop, Corfu, Greece, 1999.

[7] T. Darrell, G. Gordon, M. Harville, J. Woodfill, Integrated person tracking using stereo, color, and pattern detection, in: Computer Vision and Pattern Recognition (CVPR), Santa Barbara, California, 1998, pp. 601–608.

[8] I. Haritaoglu, D. Harwood, L. Davis, W^4S: a real-time system for detecting and tracking people in $2\frac{1}{2}$d, in: European Conference on Computer Vision (ECCV), Vol. 1, Freiburg, Germany, 1998, pp. 877–892.

[9] J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. Hale, S. Shafer, Multi-camera multi-person tracking for EasyLiving, in: Workshop on Visual Surveillance, Dublin, Ireland, 2000, pp. 3–10.

[10] M. Greiffenhagen, V. Ramesh, D. Comaniciu, H. Niemann, Statistical modeling and performance characterization of a real-time dual camera surveillance system, in: Computer Vision and Pattern Recognition (CVPR), Vol. 2, Hilton Head, South Carolina, 2000, pp. 335–342.

[11] S. Intille, J. Davis, A. Bobick, Real-time closed-world tracking, in: Computer Vision and Pattern Recognition (CVPR), San Juan, Puerto Rico, 1997, pp. 697–703.

[12] D. Beymer, Person counting using stereo, in: Workshop on Human Motion, Austin, Texas, 2000, pp. 127–133.

[13] T. Darrell, D. Demirdjian, N. Checka, P. Felzenszwalb, Plan-view trajectory estimation with dense stereo background models, in: International Conference on Computer Vision (ICCV), Vol. 2, Vancouver, Canada, 2001, pp. 628–635.

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

[15] J. Verly, D. Dudgeon, R. Lacoss, Model-based automatic target recognition system for the UGV/RSTA LADAR, in: Image Understanding Workshop, Vol. 1, Monterey, California, 1994, pp. 559–583.

[16] M. Snorrason, J. Norris, P. Backes, Vision based obstacle detection and path planning for planetary rovers, in: 13th AeroSense Conference, SPIE, Vol. 3693, 1999, pp. 44–54.

[17] M. Harville, A framework for high-level feedback to adaptive, per-pixel, mixture-of-gaussian background models, in: European Conference on Computer Vision (ECCV), Vol. 3, Copenhagen, Denmark, 2002, pp. 543–560.

[18] M. Harville, G. Gordon, J. Woodfill, Foreground segmentation using adaptive mixture models in color and depth, in: IEEE Workshop on Detection and Recognition of Events in Video, Vancouver, Canada, 2001, pp. 3–11.

[19] Y. Bar-Shalom, X. Li, Multitarget-multisensor tracking: principles and techniques, YBS Publishing, 1995.

[20] D. Focken, R. Stiefelhagen, Towards vision-based 3-d people tracking in a smart room, in: International Conference on Multimodal Interfaces (ICMI), Pittsburgh, Pennsylvania, 2002.

[21] D. Schulz, D. Fox, J. Hightower, People tracking with anonymous and ID-sensors using Rao-Blackwellised particle filters, in: International Joint Conference on Artificial Intelligence (IJCAI), Acapulco, Mexico, 2003.

[22] J. MacCormick, A. Blake, A probabilistic exclusion principle for tracking multiple objects, in: International Conference on Computer Vision (ICCV), Corfu, Greece, 1999, pp. 572–578.

[23] D. Schulz, W. Burgard, D. Fox, A. Cremers, Tracking multiple moving objects with a mobile robot, in: Computer Vision and Pattern Recognition (CVPR), Vol. 1, Kauai, Hawaii, 2001, pp. 371–377.

[24] I. Cohen, G. Medioni, Detecting and tracking moving objects for video surveillance, in: Computer Vision and Pattern Recognition (CVPR), Vol. 2, Ft. Collins, Colorado, 1999, pp. 319–325.

[25] A. Lipton, H. Fujiyoshi, R. Patil, Moving target classification and tracking from real-time video, in: IEEE Workshop on Applications of Computer Vision, Princeton, New Jersey, 1998, pp. 8–14.

[26] C. Eveland, K. Konolige, B. Bolles, Background modeling for segmentation of video-rate stereo sequences, in: Computer Vision and Pattern Recognition (CVPR), Santa Barbara, California, 1998, pp. 266–271.

[27] Y. Ivanov, A. Bobick, J. Liu, Fast lighting independent background subtraction, International Journal of Computer Vision 37 (2) (2000) 199–207.

[28] PETS2002: Third IEEE Workshop on Performance Evaluation of Tracking and Surveillance, Datasets and additional information are available at http://pets2002.visualsurveillance.org.