Stereo Person Tracking with Short and Long Term Plan-View Appearance Models of Shape and Color

Michael Harville
Mobile and Media Systems Laboratory
HP Laboratories Palo Alto

HPL-2005-135
September 16, 2005*

person tracking,
plan-view, stereo,
surveillance

In prior work, we introduced adaptive plan-view height and occupancy templates, derived from stereo camera data, for person tracking and activity recognition. These templates efficiently capture current details of each tracked person's body pose, thereby enabling good tracking performance even when multiple people occlude and interact with each other. However, the templates ignore useful color information, and their rapid evolution makes them poorly suited for recognizing the same person at well-separated times. In this paper, we seek to remedy both of these shortcomings, by 1) adding novel plan-view color templates to our short-term, template-based models of person appearance, and 2) augmenting our person descriptions with longer-term models that describe invariants of each person's shape and color. We demonstrate how each of these improves our real-time tracking performance on challenging, multi-person sequences.

* Internal Accession Date Only

Presented at the IEEE International Conference on Advanced Video and Signal-Based Surveillance, 15-16 September 2005, Como, Italy
Approved for External Publication

Stereo Person Tracking with Short and Long Term Plan-View Appearance Models of Shape and Color

Michael Harville
Hewlett-Packard Laboratories
1501 Page Mill Rd., Palo Alto, CA 94304 United States

Abstract

In prior work, we introduced adaptive plan-view height and occupancy templates, derived from stereo camera data, for person tracking and activity recognition. These templates efficiently capture current details of each tracked person's body pose, thereby enabling good tracking performance even when multiple people occlude and interact with each other. However, the templates ignore useful color information, and their rapid evolution makes them poorly suited for recognizing the same person at well-separated times. In this paper, we seek to remedy both of these shortcomings, by 1) adding novel plan-view color templates *to our short-term, template-based models of person appearance, and 2) augmenting our person descriptions with longer-term models that describe invariants of each person's shape and color. We demonstrate how each of these improves our real-time tracking performance on challenging, multi-person sequences.*

1. Introduction

As methods for producing real-time dense depth imagery have advanced [4, 5, 12, 13, 14], an increasing number of person tracking systems have chosen to rely upon them for input. While some researchers have tracked people in depth images directly, others have found it useful to create "plan-view" projections of the input depth data, in which the 3D data inherent in a depth image is re-rendered as if viewed from an overhead, orthographic camera [1, 2, 3, 8]. Because people typically do not overlap much in the dimension normal to the ground, plan-view projections of depth data allow people to be more easily separated and tracked than in the original "camera-view" depth images. Plan-view images may be obtained from stereo cameras mounted to the side of a scene, rather than above it, so that large viewing volumes are obtained and the ability to see faces is preserved.

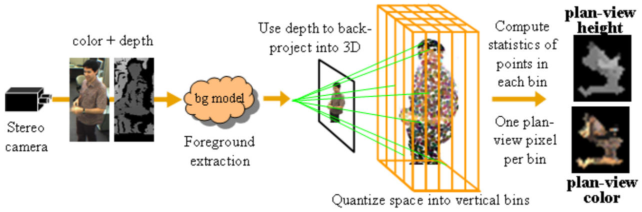
color + depth
Stereo camera
bg model
Foreground extraction
Use depth to back-project into 3D
Quantize space into vertical bins
Compute statistics of points in each bin
One plan-view pixel per bin
plan-view height
plan-view color

Figure 1. Concepts for plan-view map construction, and example plan-view height and color templates.

Figure 1 illustrates the concept of plan-view map formation from depth data. After foreground segmentation (based on color and/or depth), every foreground image pixel with reliable depth can be back-projected, using camera calibration data and a perspective model, to its corresponding 3D scene point. Back-projection of all such foreground pixels creates a 3D point cloud representing the portion of the foreground surface visible to the stereo camera. If 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 - is known, space may be discretized into a regular grid of vertically oriented bins, and then statistics of the 3D point cloud may be computed within each bin. A plan-view image contains one pixel for each vertical bin, with the value at the pixel being some statistic of the 3D points within the corresponding bin.

In previous work [7], we demonstrate real-time reporting of 3D person tracks, based on two plan-view statistics: 1) *occupancy*, which reflects the number of points in each bin, and 2) *height*, which was computed as the height above ground of the highest point within each bin. The plan-view occupancy map provides an estimate of the "amount" of foreground at each floor location, while the plan-view height map indicates the shape of foreground objects as viewed from above. We employed person appearance models consisting of *templates*, or small images, extracted from the plan-view maps at the estimated person locations. After each tracking step, each person's templates were updated with data from his newly estimated plan-view map location.

The method of [7] was tested on many challenging video sequences, containing multiple people engaging in a variety of behaviors and interactions, and was shown to provide superior performance to other known stereo-based methods. However, significant error modes still remained. First, the tracker occasionally swapped identities of closely interacting people, despite large differences in the color of their clothing. The method made no use of color beyond the foreground extraction stage, and hence lacked important cues for avoiding such errors. Second, as with most tracking methods based on adaptive templates, the templates sometimes "drifted" off of moving people and onto neighboring plan-view image regions that are varying less rapidly over time. These relatively static regions might correspond to a non-moving person, or to a non-person object that temporarily appears in the foreground because it was moved by a person or falls in his shadow. Although the plan-view image substrate afforded effective strategies for largely eliminating template drift (see [7], and Section 4), drift was still the greatest source of tracking errors.

In this paper, to address both of the above issues as well as improve other aspects of tracking, we extend the method of [7] in two significant ways. First, we augment the template-based, person appearance models with color information. Specifically, we introduce *plan-view color maps and templates*, which are enabled through back-projection of color foreground pixels together with corresponding depth pixels, during creation of the 3D point cloud as described above. Color statistics of points in each vertical spatial bin are computed to produce plan-view color maps, from which per-person color templates may be extracted and tracked. We believe that plan-view color statistics have not previously been used in video-based surveillance.

Second, we complement the rapidly evolving, template-based person appearance models with "long-term appearance models" that seek to capture relatively invariant shape and color statistics for each person. By developing a tracking algorithm to match new image data to both the short and long term models of each person, we are better able to prevent template drift from a person onto dissimilar, nearby objects, and we more accurately connect tracks of people before and after their temporary occlusion.

In Section 2, we review our methods for plan-view height and occupancy map creation, and show how plan-view color map construction may be added straightforwardly. Section 3 describes our person appearance models, consisting of short-term, template-based components as well as new long-term components. In Section 4, we discuss how the tracking method of [7] is modified to incorporate both the new plan-view color data and the long-term appearance models. These modifications result in significant tracking performance improvement, as illustrated in Section 5.

2. Plan-View Map Construction

The input to our method is a single video stream of "color-with-depth" data, where each pixel contains three color components and one depth component. Color and depth for one frame of such a stream are shown in Figures 2(**a**) and 2(**b**). The substantial noise, imprecise object borders, and large regions of unreliable data evident in the depth image are typical of the input we used. Color-with-depth video may be obtained from a pair of closely-spaced cameras operating as a single, stereo unit. The scene "foreground" is segmented within the input video, as in Figure 2(**c**), using color and/or depth data. We employ the technique of [6], which is based on Time-Adaptive, Per-Pixel Mixtures of Gaussians (TAPPMOGs) in a joint color-with-depth observation space. However, we also obtain good results with adaptive methods based on color or depth alone.

Foreground pixels with reliable depth are back-projected to form a 3D point cloud in the coordinate frame of the camera, using a standard perspective model. Prior calibration of the stereo unit provides its location and orientation within an XYZ-world coordinate system, in which the XY-plane is aligned with the ground, and Z represents height above ground. The 3D point cloud is rotated and translated from the camera coordinate frame to the XYZ-space. This space is divided into Z-aligned "vertical" bins, for plan-view map construction. We typically use bins intersecting the XY-plane in a square grid, with 2-4cm extent.

We build plan-view maps in two steps: 1) construction of raw maps, followed by 2) map refinement. The raw maps directly image some statistic of the points in each vertical bin. In prior work [7], we image two statistics, which we call "height" and "occupancy", and denote the corresponding maps as $\mathcal{H}_{raw}$ and $\mathcal{O}_{raw}$. $\mathcal{H}_{raw}$ contains the height of the highest point in each vertical bin that is above ground level and below a reasonable maximum height H_{max} at which to find human body parts (e.g. 230cm). $\mathcal{O}_{raw}$ displays weighted counts of the points in each bin, where the weights effectively convert the counts to estimates of the physical surface area (in cm^2) of foreground objects visible to the camera within each vertical bin.

$\mathcal{H}_{raw}$ and $\mathcal{O}_{raw}$ may be computed efficiently in a single pass through the input depth image. Specifically, for each depth pixel, the weighted count is incremented in the corresponding plan-view pixel of $\mathcal{O}_{raw}$, while the corresponding $\mathcal{H}_{raw}$ pixel is set to the Z-value (height above ground) for this depth pixel if it exceeds the current value stored there. The noisy height map thus produced is then improved via the map refinement stage. First, both $\mathcal{O}_{raw}$ and $\mathcal{H}_{raw}$ are smoothed with a Gaussian kernel whose plan-view extent is a fraction of that expected for people. Next, the height map is set to zero wherever the smoothed occupancy falls below a threshold. The resulting "masked" height map $\mathcal{H}_{masked}$ thus omits low-confidence regions where nothing "significant", as measured by the smoothed occupancy statistic, is present. This refinement is critical to producing useful height maps for tracking. Examples of smoothed occupancy and masked height maps are shown in Figures 2(**d**) and 2(**e**).

We build plan-view color maps in a manner analogous to that of height maps. First, a raw color map, denoted $\mathcal{C}_{raw}$, is constructed in the same single pass through the depth data needed to build $\mathcal{O}_{raw}$ and $\mathcal{H}_{raw}$. We experimented with maps representing two different color statistics: 1) the color of the highest point in each bin, and 2) the mean color of all points in the bin. To compute the former, we simply update a plan-view pixel in $\mathcal{C}_{raw}$ with the color of a camera-view

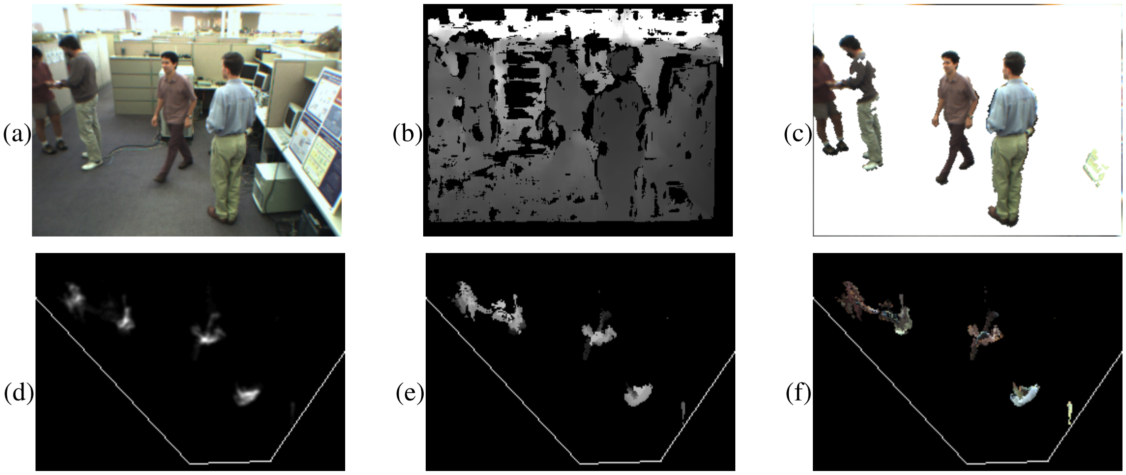

Figure 2. (a)-(c): **Example camera-view input:** (a) **Color,** (b) **Depth (unreliable data shown in black),** (c) **Foreground color.** (d)-(f) **Plan-view maps of foreground, where camera's plan-view location is just outside the bottom of the images, and the lines indicate field of view:** (d) **Smoothed occupancy** $\mathcal{O}_{sm}$**,** (e) **Height map** $\mathcal{H}_{masked}$**,** (f) **Color map** $\mathcal{C}_{masked}$**.**

foreground pixel whenever the corresponding $\mathcal{H}_{raw}$ plan-view pixel is updated with a new maximum height value obtained from a foreground camera-view depth pixel. To compute the mean color for a bin, we instead accumulate color values in the appropriate pixels of $\mathcal{C}_{raw}$, using the corresponding occupancy values as weights, and then divide the total accumulated colors by the final corresponding $\mathcal{O}_{raw}$ location values. In either case, we set the color map to zero wherever plan-view occupancy is low, as was done for the height maps, to produce $\mathcal{C}_{masked}$.

Figure 2(**f**) shows an example $\mathcal{C}_{masked}$, containing the color of the highest point in each bin. Images of this statistic proved to be very similar, in general, to those of the mean color in each bin. Due to the much greater computational expense of computing the mean, we use maps of the color of the highest point in our tracking experiments. This form of $\mathcal{C}_{masked}$ may be considered an approximation of the appearance of foreground objects as if viewed from above. While we believe plan-view color maps are novel for surveillance applications, they are related to the outputs of some methods for textured surface reconstruction from stereo video, such as [9]. Also, Mittal and Davis [10] use stereo data to develop height-stratified color appearance models for tracked people, but the models are formed in the camera-view space rather than in plan-view.

3. Person Appearance Models

Together, plan-view occupancy, height, and color maps provide a very powerful substrate on which to build tracking applications. Height maps preserve about as much 3D shape information as is possible in a 2D image, while occupancy maps provide a sense of the total mass of tracked objects. Our novel color maps add important appearance information. In this section, we describe person appearance models that are designed to take advantage of the details in the plan-view maps, in order to better track people through partial occlusions and close interactions. We divide our appearance model in short-term and long-term components, described in the next two sections, respectively.

3.1. Short-Term Appearance Models

Let $\vec{x}_i^t$ denote the state associated with the ith tracked person at time t. Time superscripts are omitted hereafter in this section except where needed for clarity. We decompose $\vec{x}_i$ into short and long term components $\vec{S}_i$ and $\vec{L}_i$, respectively. Person position, velocity, and body pose are all quantities that may vary rapidly over time, and hence we include these in $\vec{S}_i$. Specifically, we use $\vec{S}_i = \left\langle \vec{p}_i, \vec{v}_i, \vec{B}_i \right\rangle$, where $\vec{p}_i = (p_{x,i}, p_{y,i})$ is plan-view location, $\vec{v}_i = (v_{x,i}, v_{y,i})$ is plan-view velocity, and $\vec{B}_i$ represents body pose.

While $\vec{B}_i$ might be parameterized in terms of an articulated body model and joint angles, we instead find that simple templates of plan-view image data, extracted as rectangular image patches at estimated person locations, provide more easily-computed, but still powerful, pose descriptors. In fact, we demonstrate in [7] that these templates are sufficiently powerful to enable reliable determination of the orientation of a person's body, and to discriminate in real-time between poses such as standing, sitting, bending over, crouching, and reaching. In prior work, our tracking used templates derived from both the plan-view occupancy and height maps. Here, we use height and color templates, as the addition of occupancy templates to our pose descriptors was found to have negligible improvement on tracking accuracy. We thus can rewrite our our short-term appearance model as $\vec{S}_i = \langle \vec{p}_i, \vec{v}_i, \mathcal{T}_{H,i}, \mathcal{T}_{C,i} \rangle$, where $\mathcal{T}_{H,i}$ and $\mathcal{T}_{C,i}$ are the ith person's height and color templates, respectively.

Example color and height templates extracted during

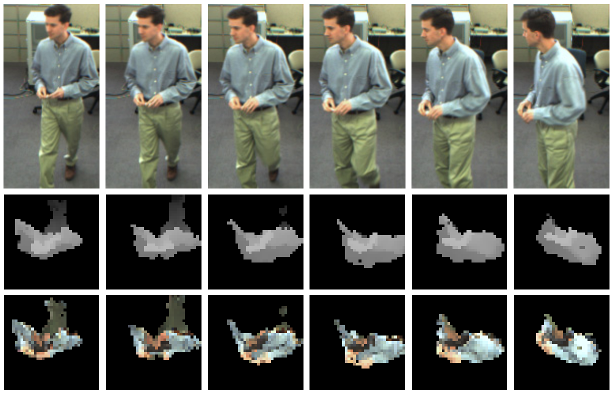

Figure 3. Templates for tracked person over 0.9 seconds (every other frame shown). Row 1: Color input (depth not shown); Row 2: Height templates $\mathcal{T}_H$; Row 3: Color templates $\mathcal{T}_C$. Note that head, hands, and legs are visible.

tracking are shown in Figure 3. At each time step, after tracking completes, the template models are replaced with the current plan-view image data at locations centered on the estimated person locations. Due to the limited range of torso widths of people, together with the strong tendency of people to remain in a predominantly upright position (even when sitting), we are able to obtain good tracking performance with a template size that remains constant across all people and all time. We employ square templates that are $2 * W_{avg}$ on a side, where W_{avg} is an estimate of the average torso width of people. We use $W_{avg} \approx$ 40cm. The template size in pixels depends on the plan-view map resolution: with vertical bins of 4cm extent, our templates are about 20 pixels on a side.

3.2. Long-Term Appearance Models

Each person's long-term appearance model should describe features that are relatively independent of body pose and activity over the duration of tracking. We would like these features to take good advantage of the multiple modalities provided by our plan-view maps. Hence, we express the ith person's long-term model as $\vec{L}_i = \left\langle \vec{L}_{S,i}, \vec{L}_{C,i} \right\rangle$, where $\vec{L}_{S,i}$ and $\vec{L}_{C,i}$ are shape and color components.

While $\vec{L}_{S,i}$ might include measurements of many different parts of an articulated body, we currently use a model consisting solely of an estimate of the person's standing height. A new such estimate h is derived at each time step from the person's current plan-view height template, as the 90%-ile height value in the template. The new estimate is used to update $\vec{L}_{S,i}$ via a simple recursive process:

$$\vec{L}^t_{S,i} = \alpha * h + (1 - \alpha) * \vec{L}^{t-1}_{S,i} \qquad (1)$$

The "learning rate" α is set very low, equivalent to a time constant of several minutes, to enable $\vec{L}_{S,i}$ to adapt to persistent posture changes (e.g. when someone sits down), without varying greatly due to temporary occlusions or brief posture deviations (e.g. bending over to pick something up).

For $\vec{L}_{C,i}$, we employ a color histogram, quantized coarsely (e.g. 4 bins) in each color channel. To mitigate color variation due to shadows and illumination, we use a normalized color space of $Y = R + G + B$, $normR = R/Y$, and $normG = G/Y$. At each time step, the raw RGB colors in a person's plan-view template are converted to the normalized space, divided into histogram bins, and then normalized by the total template pixel count. Each component of $\vec{L}_{C,i}$ is then updated independently, using a recursive process analogous to that of Equation (1), with the corresponding component of this normalized histogram.

4. Probabilistic Multi-Person Tracking

Many techniques for multi-person tracking have been proposed; see [11] for a recent survey. Tracking in plan-view occupancy maps has been described by several researchers [1, 2, 3, 8], using relatively simple person appearance models such as Gaussian occupancy densities. While these methods perform quite well, we demonstrate significant tracking improvement in [7] by using plan-view height data and template appearance models. In this section, we adapt the maximum *a posteriori* (MAP) tracking method of [7] to incorporate the plan-view color templates and long-term appearance models introduced in Section 3.

Let $X^t = \left(\vec{x}^t_1, \ldots, \vec{x}^t_{m_t}\right)$ denote the state at time t associated with a configuration of m_t tracked people. Further, let Z^t be our measurements at time t, consisting of the current plan-view height and color maps, denoted here as Z^t_H and Z^t_C, respectively. In a MAP probabilistic tracking framework, we seek the X^t that maximizes $P\left(Z^t|X^t\right) P\left(X^t\right)$.

As in [7], the prior distribution $P\left(X^t\right)$ over multi-person configurations at a given time step is a product of two parts. The first part is based on dynamical prediction applied to the estimated configuration X^{t-1} from the previous time step, while the second part is an inter-person "exclusion" term that discourages estimation of two people at very similar locations. For prediction, we assume a constant velocity model, with no change in either our short or long term appearance models. These are reasonable predictions at sufficiently high system frane rates. If we denote the predicted position of the ith person at time t as $\widetilde{\vec{p}^t_i}$, and if we assume a Gaussian probability density $\eta\left(\cdot\right)$ for each person's new location centered at his respective $\widetilde{\vec{p}^t_i}$, then we may express the prior over X^t as:

$$P\left(X^t\right) = \prod_i \eta\left(\widetilde{\vec{p}^t_i}, \frac{\alpha}{2}\Delta t^2, \vec{p}^t_i\right) * \prod_{i \neq j}\left(1 - \eta\left(\vec{p}^t_i, W_{avg}, \vec{p}^t_j\right)\right) \quad (2)$$

The first term is a product of Gaussians centered at predicted locations $\widetilde{\vec{p}^t_i}$ and evaluated at hypothesized locations $\vec{p}^t_i$, with variances $\frac{\alpha}{2}\Delta t^2$ equal to the positional error that would be produced from a reasonable maximum acceleration α of a person in the time Δt since the last measurement. Multiplication of these Gaussians assumes independence between person locations, which is corrected by the second term. The Gaussians in the second term have "person-size" variance W_{avg}, and are centered at the ith person's hypothesized location but evaulated at the jth person's hypothesized locations. Hence, if the hypothesized locations of two people are close together, one of the terms in the second part of Equation (2) will be near zero. Note that the prior of Equation (2) is a function of person locations only; we assume a uniform prior over all other components of our person state.

The measurement likelihood $P\left(Z^t|X^t\right)$ is approximated as a product of independent likelihoods conditioned on the individual person states $\vec{x}_i^t$, each of which is composed of likelihoods conditioned independently on the shape and color parts of our short and long term color appearance models. Omitting the t superscripts, we have:

$$\begin{aligned} P\left(Z|X\right) &= \prod\nolimits_i P\left(Z|\vec{x}_i\right) \qquad (3) \\ &= \prod\nolimits_i P\left(Z_C|\vec{p}_i, \mathcal{T}_{C,i}\right) P\left(Z_H|\vec{p}_i, \mathcal{T}_{H,i}\right) * \\ &\quad\ \prod\nolimits_i P\left(Z_C|\vec{p}_i, \vec{L}_{C,i}\right) P\left(Z_H|\vec{p}_i, \vec{L}_{S,i}\right) \end{aligned}$$

$P\left(Z_C|\vec{p}_i, \mathcal{T}_{C,i}\right)$ denotes the likelihood of the plan-view color map measurements given the hypothesis that person i, with predicted color appearance represented by template $\mathcal{T}_{C,i}$, is at plan-view location $\vec{p}_i$. To calculate $P\left(Z_C|\vec{p}_i, \mathcal{T}_{C,i}\right)$ at a given $\vec{p}_i$, we first align $\mathcal{T}_{C,i}$ with location $\vec{p}_i$ in the plan-view color map, and then compute the "sum of absolute differences" (SADs) over all pixel locations that are non-zero in each. We divide by the number of pixels used, and then transform the SADs to likelihoods via sigmoidal functions fitted to training data. $P\left(Z_H|\vec{p}_i, \mathcal{T}_{H,i}\right)$ is defined and calculated analogously.

$P\left(Z_C|\vec{p}_i, \vec{L}_{C,i}\right)$ denotes the likelihood of the plan-view color map measurements given the hypothesis that person i, with long-term color appearance model $\vec{L}_{C,i}$, is located at $\vec{p}_i$. We calculate this likelihood by building a histogram of the plan-view map colors in a $2W_{avg}$-sized region about $\vec{p}_i$, computing its L1 norm difference from $\vec{L}_{C,i}$, and then transforming the difference to a likelihood via a learned sigmoidal function. A similar technique is applied to compute $P\left(Z_H|\vec{p}_i, \vec{L}_{H,i}\right)$, using the absolute difference between $\vec{L}_{S,i}$ (an estimate of the person's height) with the maximal value within a W_{avg}-sized region about $\vec{p}_i$.

To find the X^t that maximizes the MAP probability that is the product of Equations (2) and (3), we use the technique of [7], which shares some foundation with particle filtering but is adapted for real-time operation in template-based tracking. In brief, we first evaluate the likelihood of Equation (3) exhaustively within regions $\mathcal{R}_i$ centered at predicted person locations $\widetilde{\vec{p}_i}$ and large enough to account for reasonable prediction errors and inter-frame person acceleration. Next, we select several random orderings in which to estimate new locations for currently tracked people. For each ordering, we first find the position $\vec{p}_1$, by exhaustive search within $\mathcal{R}_1$, that maximizes the product of Equations (2) and (3) considering only the first person in the ordering. We then fix $\vec{p}_1$, and search $\mathcal{R}_2$ for the location $\vec{p}_2$ that maximizes the MAP estimate considering only the first two people in the ordering. This continues sequentially for all people in the ordering. Once this is completed for all orderings, that with highest MAP probability is selected, and the locations $\vec{p}_i$ used to obtain it are the new person locations. A post-tracking, template "re-centering" step is applied to shift each $\vec{p}_i$ onto the local plan-view occupancy center-of-mass within a $2W_{avg}$-sized window (see [7] for discussion). Person states $\vec{x}_i$ are then updated using plan-view map data centered at these locations, as described in Section 3.

Our method allows the possibility that, at a given time step, one or more currently tracked people may be "lost", perhaps due to temporary occlusion or exiting the scene. Specifically, when performing the sequential MAP maximization described above for a given ordering, we ignore any person for whom the highest product of the four likelihoods in Equation (3) falls below a threshold. If this ordering is selected as the best, all people ignored by it are considered lost. For a short period of time (e.g. less than a minute) after initially losing track of a person, we attempt to match his long-term appearance model to that of any "new" person detected. New people are detected by first removing plan-view data around locations $\vec{p}_i$ of successfully tracked people, and then searching for other locations with a height above some reasonable minimum for people, with high local occupancy, and with sufficient motion in the corresponding camera-view image region. Long-term appearance models for new people are initialized using the current plan-view image data. The $\vec{L}_S$ shape and $\vec{L}_C$ color models of each new person are compared with those of any lost person whose predicted or last observed location is near that of the new person. If both L1 norm differences are below a threshold, the track of the lost person is continued at new person's location. Use of long-term appearance models to link interrupted tracks greatly improves upon our prior methods, which matched lost people to new ones based only on time and location information.

5. Experimental Results

A C++ implementation of our system runs on a dual 2.2GHz PC, with software computation of 320x240-resolution color-with-depth video provided by a Point Grey Triclops stereo module [12]. Software depth computation by the Triclops was the most expensive operation, so that our method can be expected to run equally fast on much more modest computers when hardware-assisted stereo (e.g. [4, 13, 14]) is employed. The current method is significantly slower than our prior work, mostly due to the

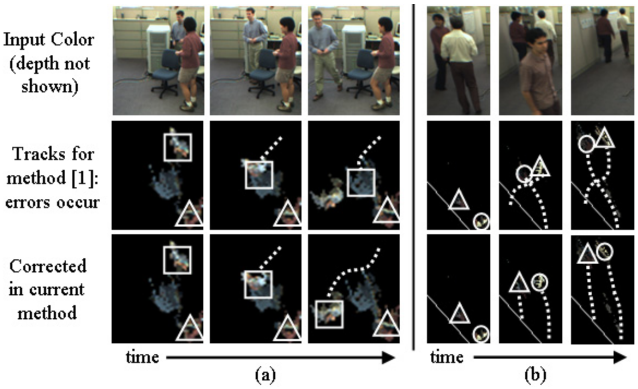
Input Color (depth not shown)
Tracks for method [1]: errors occur
Corrected in current method
time
(a)
time
(b)

Figure 4. Examples of tracking errors of [7] corrected by the present method. (a): Template drift from person onto recently moved chair, corrected by long-term appearance model. (b) Identity swap error of similarly sized people, due to brief occlusion as they both change location, corrected by addition of plan-view color.

new expense of computing and comparing color histograms at many plan-view image locations during the tracking process. The frame rate varies between 10-15Hz, tending toward the lower end as more people are tracked.

We quantitatively evaluated our method on 16 minutes of multi-person test sequences captured at 12-15Hz and 320x240 resolution, with the camera typically mounted with a view like that of Figure 2**(a)**. All sequences contain at least three people within a 10m x 20m space, with many occlusions, close inter-person interactions, moved objects, shadows, and unusual postures and behaviors. We examined our tracking results for "significant" errors, defined as any of 1) losing track of a person (e.g. failing to reconnect a track after occlusion), 2) failing to detect a person, 3) swapping the identities of two tracked people, and 4) tracking a non-person object. To demonstrate the improvements afforded by plan-view color maps and long-term appearance models, we compared the number of significant errors for our current method to those obtained when 1) color templates are replaced with occupancy, 2) long-term appearance models are omitted, or 3) both. The latter case is equivalent to the method of [7], which was shown to be superior to other known stereo-based tracking methods. Comparison to the best monocular methods is difficult because the community's standard test sequences provide no depth data.

The results of this comparison are shown in Table 1. By adding either color information or long-term appearance models, a significant reduction in swapped identities and tracking failures is achieved. When these improvements are combined, tracking becomes quite robust. Figure 4 shows examples of lost track (due to template drift) and identity swap errors, made by the method of [7], corrected here through use of long-term appearance models and plan-view color data, respectively. In our current system, the dominant error modes now appear to be 1) tracking non-person objects, and 2) losing fully-occluded or motionless people within the camera view. The former could likely be reduced by use of a stronger person model during new person detection, while the latter might be mitgated by a more intelligent understanding of the valid plan-view entry and exit points in the scene. We plan to address these shortcomings in future work. Nevertheless, we believe our current tracking performance to be state-of-the-art.

Table 1: Tracking Performance Comparison

Method Type		**Number of Errors**			
Plan-View Color	Long-Term Models	Track Lost	Detection Failure	Identity Swap	NonPerson Tracked
		11	2	6	7
X		9	2	4	5
	X	8	2	3	5
X	X	5	2	1	3

Counts of "significant" errors made on test sequences, for various methods. Top row is method of [7], bottom row is current work.

6. Conclusions and Future Work

We have demonstrated a new method for robust, real-time person tracking that well exploits the rich shape and color information available from stereo cameras. The addition of long-term appearance models significantly improved on prior performance, and we believe the novel concept of plan-view color maps may have many applications worth further exploration. We plan to develop more sophisticated long-term person models, perhaps including representations of each person's shape and color appearance in each of several different, observed poses. We also plan to work on algorithms to increase the speed of the method, particularly with regard to the extensive search processes and histogram computations used during tracking.

References

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

[2] N. Checka, et. al. "A probabilistic framework for multi-modal multi-person tracking." In *Wkshp. on Multi-Object Tracking*, 2003.

[3] T. Darrell, D. Demirdjian, et. al. "Plan-view trajectory estimation with dense stereo background models." In *ICCV'01*.

[4] S.B. Gokturk, H. Yalcin, C. Bamji. "A time-of-flight depth sensor - system description, issues, and solutions". In *Wkshp. on Real-Time 3D Sensors and Their Use*, 2004.

[5] R. Gvili, A. Kaplan, E. Ofek, G. Yahav. "Depth keying." In *SPIE Elec. Imaging*, Jan. 2003, Vol. 5006, pp. 564-574.

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

[7] M. Harville, D. Li. "Fast, integrated person tracking and activity recognition with plan-view templates from a single stereo camera." In *CVPR'04*.

[8] Interval Research Corp., unpublished work, June 1999.

[9] S.B. Kang, R. Szeliski. "3-D scene data recovery using omnidirectional multibaseline stereo." In *CVPR'96*.

[10] A. Mittal, L. Davis. "M_2Tracker: a multi-view approach to segmenting and tracking people in a cluttered scene." *Intl. J. Comp. Vis. (51)*, No. 3, pp. 189-203, 2003.

[11] Performance Evaluation of Tracking and Surveillance, Workshop Proceedings, 2004.

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

[13] J. Woodfill, G. Gordon, R. Buck. "Tyzx DeepSea high frame rate stereo vision system." In *Wkshp. on Real-Time 3D Sensors* , 2004.

[14] R. Yang, M. Pollefeys. "Multi-resolution real-time stereo on commodity graphics hardware." In *CVPR'03*.