Fast, Integrated Person Tracking and Activity Recognition with Plan-View Templates from a Single Stereo Camera

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

Dalong Li
Center for Signal and Image Processing
Georgia Inst. of Technology, Atlanta, GA 30332
dalong@ece.gatech.edu

Abstract

Plan-view projection of real-time depth imagery can improve the statistics of its intrinsic 3D data, and allows for cleaner separation of occluding and closely-interacting people. We build a probabilistic, real-time multi-person tracking system upon a plan-view image substrate that well preserves both shape and size information of foreground objects. The tracking's robustness derives in part from its "plan-view template" person models, which capture detailed properties of people's body configurations. We demonstrate that these same person models - obtained with a single compact stereo camera unit - may also be used for fast recognition of body pose and activity. Principal components analysis is used to extract plan-view "eigenposes", onto which person models, extracted during tracking, are projected to produce a compact representation of human body configuration. We then formulate pose recognition as a classification problem, and use support vector machines (SVMs) to quickly distinguish between, for example, different directions people are facing, and different body poses such as standing, sitting, bending over, crouching, and reaching. The SVM outputs are transformed to probabilities and integrated across time in a probabilistic framework for real-time activity recognition.

1. Introduction

Dense (per-pixel) depth or disparity imagery, obtained from stereo cameras, offers powerful new approaches for the difficult problems in vision-based perception of the presence, location, and activities of people. In recent years, many systems for producing real-time dense depth imagery have been described (for example, see [4, 12, 19, 24, 26]), and most are either currently inexpensive, or promise to become so as commercial demand grows. Moreover, most of these methods can be implemented with compact, pre-calibrated stereo camera units that are mounted and utilized much like standard monocular cameras, but which produce "color-with-depth" video. It therefore seems that cheap, compact stereo cameras will replace monocular cameras in many contexts in the near future. We develop here new methods for person perception using just one such "next-generation" camera.

Most person perception methods using dense depth video have analyzed and tracked features, statistics, and patterns directly in the depth images. Unfortunately, the substantial noise, imprecise object contours, and large regions of low-confidence data typical of currently available real-time depth imagery (see Figure 3**(b)**) make application of standard image analysis and tracking methods to such imagery less reliable than is customary for color or grayscale video. In person detection and tracking contexts, some researchers have therefore 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 [3, 6, 10, 15]. 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.

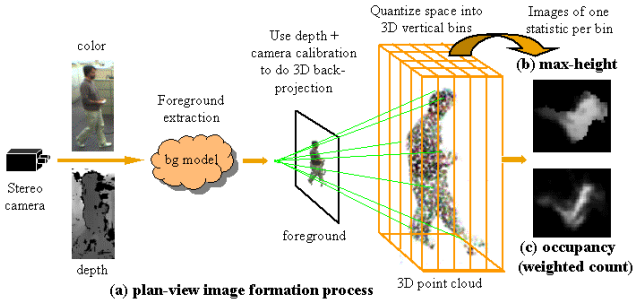

Figure 1. Concepts for plan-view map construction, and example plan-view height and occupancy templates.

A conceptual flow diagram for plan-view projection is shown in Figure 1. Every reliable depth image value can be back-projected, using camera calibration data and perspective projection, to its corresponding 3D scene point. Back-projection of all such depth image pixels creates a 3D point cloud representing the portion of the scene surfaces visible to the stereo camera. This back-projection may optionally be restricted to the scene "foreground", such as people or other moving objects. 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.

We compute two particularly useful statistics of the set of 3D points within each vertical bin: 1) *occupancy*, which reflects the number of points in each bin, and 2) *height*, which reflects the height of the highest point within each bin. When

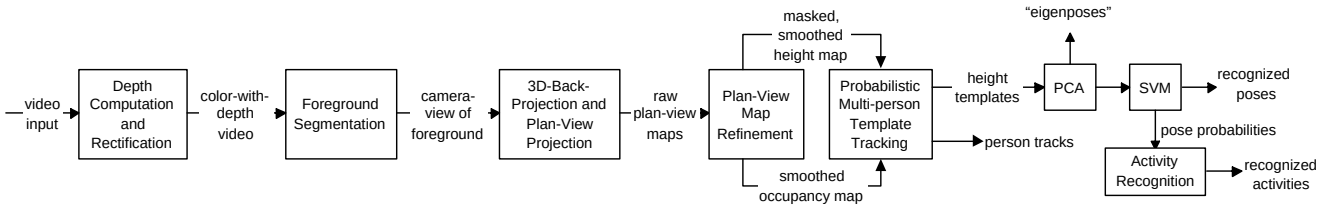

Figure 2. Overview of person perception system.

plan-view projection is restricted to the foreground in the scene, 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 the foreground objects as viewed from above. Casual inspection of this plan-view imagery reveals details of human body pose and limb positions. Furthermore, from sequences of such images, untrained observers can easily recognize dynamics of human motion and activity. Hence, one might speculate that plan-view projections could facilitate understanding not just of where people are, but also of what they are doing.

In this paper, we integrate person tracking and activity recognition into a single, real-time framework based on "plan-view template" person models. These models are allowed to adapt over time during tracking, to reflect changes in people's body configuration and visibility to the camera. While easy to construct and evolve, these models provide much richer descriptions of tracked people than simpler choices such as Gaussians or "blobs" used in prior plan-view tracking work [3, 10, 15], thereby enabling more accurate tracking of people through partial occlusions and close interactions. We also leverage these richer models to develop novel body pose and activity recognition techniques. Specifically, we formulate articulated body pose estimation as a classification task on the plan-view height templates extracted by our person tracker, and use support vector machines (SVMs) to quickly and reliably discriminate between poses such as sitting, standing, and reaching, or facing left rather than right. Compact pose representations may be obtained through re-description of plan-view templates in terms of plan-view "eigenposes", much like the "eigenface" work of Turk and Pentland [23]. Conversion of the SVM outputs to probabilities, and integration of these probabilities over time, allows for principled, stable detection of human activities and "events" such as "sitting down" or "turning to face left".

We have applied plan-view height and occupancy maps to person tracking in previous work [14]. The current work not only extends to pose and activity analysis, but also provides a new tracking method with stronger theoretical grounding and improved accuracy at similar computational cost.

Figure 2 provides an overview of our integrated person tracking and activity recognition system. In Section 2, we motivate and describe the construction of the plan-view maps that underlie our tracking and activity recognition methods. Further discussion and practical implementation details of these maps may be found in [14]. In Section 3, we discuss plan-view template person models, and we describe a novel Bayesian, maximum likelihood framework for tracking and updating these models over time to monitor the locations of multiple people. In Section 4, we explain how to re-purpose the extracted person models for human body pose and activity recognition, and we discuss experimental performance.

2. Plan View Height and Occupancy Maps

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 3**(a)** and 3**(b)**. The substantial noise, imprecise object borders, and large regions of unreliable data (indicated in black) evident in the depth image are typical of the input we used. We restrict our attention to the pixels associated with the "foreground" in the scene, as in Figure 3**(c)**. We extract foreground via the technique of [13], which is based on Time-Adaptive, Per-Pixel Mixtures of Gaussians (TAPPMOGs) in a joint color-with-depth observation space, but this choice is not critical.

Each foreground pixel with reliable depth is back-projected to build a 3D point cloud, which is then vertically binned as discussed in Section 1. In this paper, we use bins with 2x2cm plan-view extent; in practice we have used bin sizes up to 4x4cm. We next 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 bin. 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 single 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 weighting compensates for the smaller camera-view appearance of more distant objects. Our choice of Z^2/f (Z is the depth image value, f is the camera focal length) as the weighting factor causes $\mathcal{O}_{raw}$ to approximate the total physical surface area 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 (see [14]). Raw height and occupancy maps corresponding to the foreground of Figure 3**(c)** are shown in Figures 3**(d)** and 3**(e)**, respectively. In the occupancy map, five "blobs" corresponding to the five people in the foreground image are clearly visible, but noise causes this distinction to be less clear in $\mathcal{H}_{raw}$.

Our map refinement step is predicated on the idea that our confidence in the raw plan-view statistics should grow as the number of foreground points from which they are derived

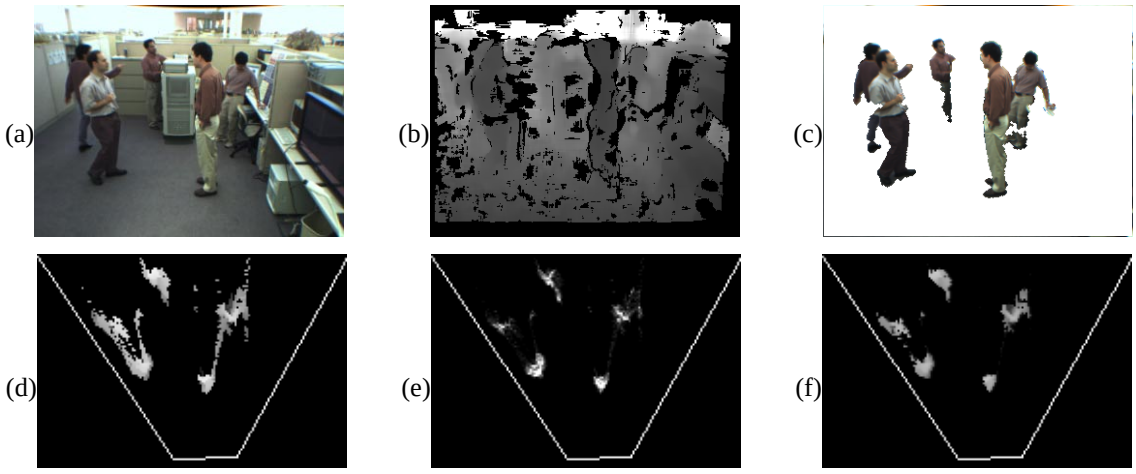

Figure 3. (a)-(c): Example camera-view input: (a) Color, (b) Depth (unreliable data shown in black), (c) Foreground color. (d)-(f) Plan-view maps of foreground, where camera's plan-view location is just outside the bottom of the images, and the lines indicate field of view: (d) Raw height map $\mathcal{H}_{raw}$, (e) Raw occupancy map $\mathcal{O}_{raw}$, (f) Masked, smoothed height map $\mathcal{H}_{masked}$.

increases. We attempt to build a confidence measure directly into the maps, by deprecating statistics with little support. One simple but effective means of achieving this is to first smooth the raw maps, and then set them to zero where the statistical support, as indicated by the local occupancy level, is below a threshold. This removes data at floor locations where nothing "significant", as measured by the smoothed occupancy statistic, is present. The masked, smoothed height map $\mathcal{H}_{masked}$ for the foreground of Figure 3**(c)** is shown in Figure 3**(f)**. Comparison of Figure 3**(d)** and **(f)** shows this refinement is critical for making our height maps useful.

We use $\mathcal{H}_{masked}$ and the smoothed occupancy map ($\mathcal{O}_{sm}$) together as the basis for the person tracking and activity recognition algorithms described in the next sections. Plan-view occupancy maps similar to ours have been used in person tracking methods developed by other researchers [3, 6, 10, 15], but height maps have not. Height maps preserve about as much 3D shape information as is possible in a 2D image, and therefore seem better suited than occupancy maps, which discard virtually all object shape information in the vertical dimension, 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, person representations in height maps tend to be more robust to partial occlusions than corresponding occupancy representations, provided that the person's upper body remains partially visible, as is common for high-mounted surveillance cameras.

3. Template-Based Person Tracking

Many methods for multi-target tracking have been proposed, including techniques based on Kalman filtering [2], joint probabilistic data association filters (JPDAFs) [20], and particle filtering [17]. In prior methods for tracking in plan-view maps, other researchers have used only occupancy data, and the person models are no more complex than Gaussians [3, 6, 10, 15]. We have recently demonstrated the benefits of plan-view height data, as well as "template" person models that are constructed directly from, and matched directly to, the plan-view image data [14]. These models are designed to take advantage of the details in the plan-view maps, in order to better track through partial occlusions and close interactions. In this paper, we replace the Kalman filter-based tracking of [14] with a new probabilistic technique that shares some of its foundation with particle filtering, but which is adapted for real-time operation in template-based tracking.

3.1. Adaptive Templates

We represent the configuration of the m_t people tracked at time t with the state $X^t = \left(\vec{x}_1^t, \ldots, \vec{x}_{m_t}^t\right)$, where $\vec{x}_i^t$ is the state associated with the ith tracked person. For simplicity, t superscripts are omitted hereafter except where required for clarity. Each individual person state $\vec{x}_i$ consists of the three-tuple $\left\langle \vec{p}_i, \vec{v}_i, \vec{S}_i \right\rangle$, 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 $\vec{S}_i$ represents the body configuration, or pose, of the person. While $\vec{S}_i$ might be parameterized in terms of joint angles or other pose descriptions, the estimation of such descriptions is likely expensive and highly error-prone. We find, instead, that simple templates of plan-view height and occupancy map data provide easily computed but powerful pose descriptors. These templates are small patches of the plan-view images extracted at the estimated person positions. The person state vector may now be rewritten as $\vec{x}_i = \langle \vec{p}_i, \vec{v}_i, \mathcal{T}_{H,i}, \mathcal{T}_{O,i} \rangle$, where $\mathcal{T}_{H,i}$ and $\mathcal{T}_{O,i}$ are the ith person's height and occupancy templates, respectively. Example templates created from plan-view image data are shown in Figure 4.

Because our plan-view representations of people are largely invariant to floor location relative to the camera, and because the spatial extents of most people, when viewed from overhead, fall within a limited range most of the time, we are able to obtain good tracking performance with a template

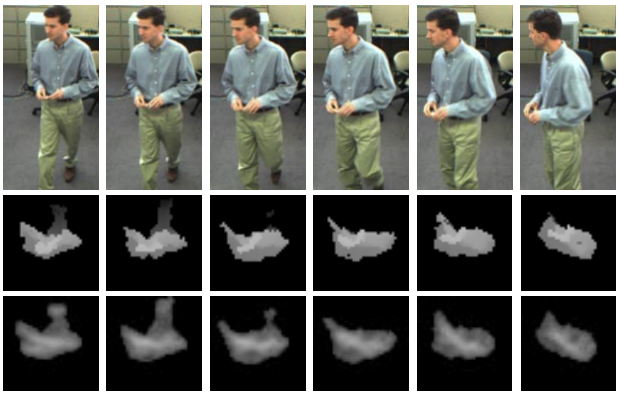

Figure 4. Evolution of templates for a tracked person over 0.9 seconds (every other frame shown). Top: Color input (depth not shown); Middle: Extracted height templates $\mathcal{T}_H$; Bottom: Extracted occupancy templates $\mathcal{T}_O$.

size that remains constant across all people and all time. We employ templates whose width and height correspond to a physical size of $2 * W_{avg}$, which is twice an estimate of the average torso width of people. We use $W_{avg} \approx 40$cm. To prevent the "slipping" of templates off tracked targets and onto background elements, we apply a "re-centering" post-tracking step, which shifts the estimated position $\vec{p}_i$ for the ith person to the center of mass of the occupancy map data in the "person-sized" square of width $2W_{avg}$ centered on $\vec{p}_i$. Both the re-centering and the constant template size would be less feasible in a "camera-view"-based tracking scheme, because a person's size in such images varies greatly with his body pose, state of occlusion, and distance from the camera.

3.2. Probabilistic Template Tracking

Let Z^t denote the plan-view map statistics for a new frame of input color-with-depth, with Z_H^t and Z_O^t denoting the height and occupancy map portions of this. At each time step t, we attempt to select the multi-person configuration X^t with maximum likelihood $P\left(X^t|Z^t\right)$, given the current measurements Z^t. From Bayes rule, and taking $P\left(Z^t\right)$ to be constant, this is equivalent to maximizing $P\left(Z^t|X^t\right)P\left(X^t\right)$.

The prior distribution $P\left(X^t\right)$ over multi-person configurations is determined in part from dynamical prediction applied to the estimated configuration X^{t-1} from the previous time step. At high system frame rates, it is reasonable to generate a predicted configuration $\widetilde{X}^t$ under the approximation of no change in velocity, body pose, or state of occlusion in each person in the time Δt since the previous estimate X^{t-1}:

$$\begin{aligned} \widetilde{\vec{p}}_i^{\,t} &= \vec{p}_i^{\,t-1} + \vec{v}_i^{\,t-1} * \Delta t \\ \widetilde{\mathcal{T}}_{H,i}^t &= \mathcal{T}_{H,i}^{t-1}; \quad \widetilde{\mathcal{T}}_{O,i}^t = \mathcal{T}_{O,i}^{t-1} \end{aligned} \tag{1}$$

Using the above person state predictions, we assume independence in the positions of the people, and construct a prior on the positional portion of X^t as a product of Gaussians $\eta\left(\widetilde{\vec{p}}_i^{\,t}, \frac{\alpha}{2}\Delta t^2, \vec{p}_i^{\,t}\right)$ centered at each of the predicted person locations $\widetilde{\vec{p}}_i^{\,t}$, and 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. Each predictive Gaussian is evaluated at the corresponding hypothesized person location $\vec{p}_i^{\,t}$. For simplicity, we assume uniform priors on the velocity and body pose portions of X^t, and omit them from $P\left(X^t\right)$.

The positions of people are not completely independent, however, since no two people typically occupy the same plan-view location. We therefore combine into the prior $P\left(X^t\right)$ an "exclusion" term that discourages estimation of different people from being at very similar locations. The full expression for our $P\left(X^t\right)$, with the exclusion term appearing at the right and normalizing factors omitted, is:

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

The expression $\eta\left(\vec{p}_i^{\,t}, W_{avg}, \vec{p}_j^{\,t}\right)$ denotes the Gaussian with "person-sized" variance W_{avg} centered at the ith person's hypothesized location but evaluated at the jth person's hypothesized location. As hypothesized locations draw nearer predicted ones, the value of the exclusion term decreases, thereby decreasing the overall configuration probability.

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 height and occupancy data. Omitting the t superscripts, we have:

$$P\left(Z|X\right) = \prod_i p\left(Z|\vec{x}_i\right) = \prod_i P\left(Z_H|\vec{p}_i, \widetilde{\mathcal{T}}_{H,i}\right) P\left(Z_O|\vec{p}_i, \widetilde{\mathcal{T}}_{O,i}\right) \tag{3}$$

$P\left(Z_H|\vec{p}_i, \widetilde{\mathcal{T}}_{H,i}\right)$ denotes the likelihood of the plan-view height map measurements given the hypothesis that person i, with predicted body pose represented by template $\widetilde{\mathcal{T}}_{H,i}$, is located at $\vec{p}_i$. $P\left(Z_O|\vec{p}_i, \widetilde{\mathcal{T}}_{O,i}\right)$ is defined analogously. These are evaluated by first computing the "sum of absolute differences" (SADs) of $\widetilde{\mathcal{T}}_{H,i}$ and $\widetilde{\mathcal{T}}_{O,i}$ with the plan-view data at $\vec{p}_i$, and then transforming the SADs to likelihoods via sigmoidal functions fitted to training data.

We combine equations (2) and (3) to obtain an expression for the posterior person configuration probability $P\left(Z^t|X^t\right)P\left(X^t\right)$, and perform tracking by maximizing this via an efficient two-stage process. First, $P\left(Z_H|\vec{p}_i, \widetilde{\mathcal{T}}_{H,i}\right)$ and $P\left(Z_O|\vec{p}_i, \widetilde{\mathcal{T}}_{O,i}\right)$ are evaluated exhaustively within regions $\mathcal{R}_i$ centered at predicted person locations $\widetilde{\vec{p}}_i$ and large enough to account for reasonable maximum prediction errors and inter-frame person acceleration. Second, new positions of the tracked people are determined sequentially for each of several (we use up to 10) randomly selected orderings of tracked people. For a given ordering, the position of the ith person is selected through maximization of the "partial configuration posterior probability" $\left(P\left(Z_i|X_i\right)P\left(X_i\right)\right)$. The partial posterior is simply the combination of equations (2) and (3), but evaluated only for people $j \leq i$ in the current ordering. Partial posteriors for person i are computed for all

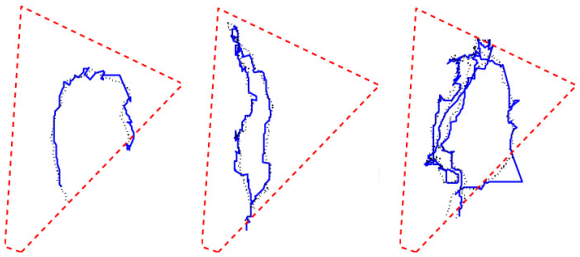

Figure 5. Comparisons of individual result tracks with estimated ground truth for three simultaneously tracked people. Results in solid lines, ground truth in dotted lines. Plan-view position of stereo pair is near lower-left corner of the images, with field of view shown in dashed lines.

locations within $\mathcal{R}_i$, while assuming the locations for people $j < i$ to be fixed. The location of the maximum value of the posterior is selected as $\vec{p}_i$ if this value is above a threshold; otherwise, we use $\vec{p}_i = \widetilde{\vec{p}}_i$. Once locations for all people in the ordering have been selected in sequence in this way, the partial posterior for the last person is in fact the full measurement probability $P\left(Z^t|X^t\right)P\left(X^t\right)$. The configuration associated with the ordering having highest posterior probability is selected as the new estimate X^t.

After estimation of the new person locations has completed, person velocities are updated using a recursive smoothing filter on position, and new templates $\mathcal{T}_H$ and $\mathcal{T}_O$ are extracted for each person as described in Section 3.1. New people are found by first removing plan-view data at estimated locations of tracked people, and then searching for other plan-view 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. Simple secondary tracking techniques, described in [14], attempt to link tracks of "lost" people (extended tracking failures) to those of "new" people detected near where lost people were last seen or are predicted to be.

3.3. Tracking Performance

An implementation of our system runs at 15 Hz on a dual 2.2GHz PC, with software computation of color-with-depth video provided by a Point Grey Triclops stereo module [19]. System computation is dominated by the Triclops software, so that the method can be expected to run equally fast on much more modest computers when hardware-assisted stereo (e.g. [4, 24, 26]) is employed. We quantitatively evaluated our method on 10 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 3**(a)**. The sequences contain over 100 challenges of types including occlusion events, close interactions, and large non-person foreground distractors (e.g. rolling chairs), but our method made only 6 "significant" errors (e.g. losing track of a person, swapping identities of two people, etc.). Related plan-view tracking methods that omit use of either height or occupancy statistics, or that use simpler person models such as Gaussians, were all found to produce at least 4 times as many errors. Our method's errors occurred primarily in cases of severe occlusion for extended periods, and/or at distant locations from the camera where depth noise produces greater instability in the plan-view templates. Incorporation of longer-term person appearance models, greater use of color data, and explicit handling of occlusions would likely eliminate most of these errors. Some example tracks are shown in Figure 5. These and other tracks were found to have a point-wise mean positional error from ground truth of 16cm.

4. Pose and Activity Recognition

Although much research has been done on the tracking of 3D articulated body motion of people in video, including depth video [9, 16], far less has dealt with the issue of initializing such trackers, by extracting 3D body pose from image data. An excellent discussion of much of this latter area of work may be found in [21]. To summarize, most prior approaches either are not fully automatic (for example, requiring manual selection of joint locations or correspondences across multiple views), are specialized for use with a limited set of poses, or demand highly complex computations to produce a single pose estimate. Many of these also rely on having multiple, widely-spaced views, on observing the body in motion, or on detailed knowledge of and constraints on articulated human body configuration.

In this work, we attempt to strike a balance of general, reliable estimation of rough body pose without excessive computation, manual guidance, widely spaced views, or *a priori* models and constraints. We formulate body pose determination as a classification task, and attempt to learn mappings between plan-view templates and a set of labeled body poses, such as "sitting", "facing left", and so on. The resulting body pose estimates are less detailed than those produced by most of the methods described above, but instead produce "high-level" pose descriptions that appear to be useful for a broad range of real-time activity recognition tasks.

4.1. SVM-Based Recognition

For recognition, we use plan-view height templates, rather than occupancy, since the former contain more 3D shape information about the tracked objects. The re-centering technique employed by the person tracking (Section 3.1) keeps the height data well-aligned with the template centers across all people and all time, thereby greatly improving our chances at reliable classification. We normalize the height values in each template by rescaling them according to the templates's 90th-percentile non-zero height value. We also normalize the spatial extent of the data within the height templates, by resampling it so that 80% of the corresponding occupancy data lies within the central W_{avg}-sized region of the template, while the template size (in pixels) and center remains unchanged. All normalization factors are kept with the normalized template data for use in classification.

We would like to recognize many different types of body pose and activities via a single flexible, efficient framework. Support vector machines (SVMs) [8] can learn highly ac-

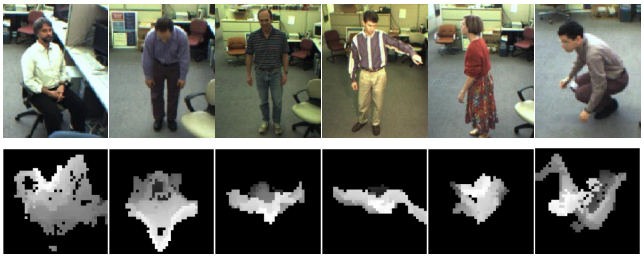

Figure 6. Example color images (depth not shown) and the extracted, normalized height templates. Left to right, poses are a) sitting b) bending over c) upright, facing camera d) arm reaching e) upright, facing left f) crouching

curate and complex decision boundaries between multiple classes of labeled points in high-dimensional spaces. If we regard the normalized height template data vectors as points in "pose space", each of which may be associated with one or more high-level pose labels such as "standing", "facing camera", and so forth, then SVMs may be trained to name the poses associated with our plan-view template data.

Using our person tracking system, we collected 1680 normalized, 39x39-pixel height templates of 4 people (3 men, 1 woman) participating in diverse activities, wearing a variety of natural clothing, under several different lighting conditions. Each data vector was hand-labeled into one or more body pose categories through observation of the original color video. We consider three pose classification tasks:

Task 1: For people standing upright in any orientation, detect reaching of an arm in any direction, as opposed to having arms in a relaxed position such at one's sides, in pockets, or folded across the chest.

Task 2: Distinguish between standing upright, sitting, bending over, crouching, and arm reaching, without regard to overall body orientation (i.e. the direction the person faces).

Task 3: Determine the direction an upright person's body is facing, choosing between the 8 categories of 0° (toward the camera), 45°, 90° (left), 135°, 180° (away from camera), 225°, 270° (right), and 315°.

The desired invariance to body orientation makes the first two tasks challenging, while for the third we might expect difficulty in distinguishing between opposite orientations (e.g. 0° and 180°). Also, although walking behavior produces distinctive plan-view map patterns, we omit detection of this here because it may be done trivially from our track data.

The "libSVM" software package [5], with a radial basis function kernel, was used to train one SVM for each of the three tasks. Next, 844 vectors of test data were collected and hand-labeled for two people not previously seen by the system (2 men, one taller and one shorter than all of the original 4 people). This data was presented to the SVMs to obtain the results shown in the rightmost column of Table 1. The most frequent errors were failure to detect arm-reaching for task 2 (typically due to poor depth data on outstretched arms), and confusions between opposite body orientations for task 3. For all tasks, however, classification was correct on more than 5 out of every 6 attempts. Furthermore, using our trained SVMs, classification of our test vectors may be done hundreds of times per second on even modest (500MHz) PCs, so that pose recognition occurs comfortably in real-time when integrated with the person tracking system described above.

4.2. Plan-View "Eigenposes"

Because our height templates typically contain 225-2500 pixels, their dimensionality is often too great for practical or efficient use with some types of classifiers. Even for our SVMs, which classify in real-time, the training on such large data vectors may be time-consuming. Applications that seek to train classifiers at run-time, perhaps for building person-specific classifiers for some poses and activities, may thus have difficulty running in real-time. A more compact data representation might also enable more insightful analysis, and would alleviate storage and transmission bandwidth concerns that are important in many systems.

We therefore seek to re-describe the normalized plan-view height templates in terms of a linear combination of basis images, where the number of coefficients in the combination is much less than the basis image dimensionality. We determine a suitable choice of basis images, namely one that captures the dominant components of variation in the plan-view height templates, by applying PCA to a set of training templates. The data to be classified is then projected onto some number of the principal components associated with the highest eigenvalues, to obtain a low-dimensional, PCA-coefficient representation. This is precisely the technique of the classic "eigenfaces" work of Turk and Pentland; the mathematics may be found in [23].

We applied PCA to the 1680 normalized height templates collected in Section 4.1. The resulting "mean pose" template, a plot of the highest 100 eigenvalues, and the most significant plan-view "eigenposes" are shown in Figure 7. The mean looks much like a "crescent moon", suggesting that this may be a good generic model for people in plan-view height maps. The first eigenpose shows the dominant variation to be the extent to which height map data protrudes from the body torso. Other eigenposes capture more complex variations.

The coefficients of projection of the normalized height templates onto some small number of eigenposes, together with the template normalizing factors, can be used to form new data vectors for pose classification. To assess the dependence of pose recognition on the number of "eigenposes" used to represent the data, we repeated the experiments of Section 4.1 using the top 10, 30, and 100 principal components of our pose basis. The results are shown in the first three columns of Table 1, and may be compared there with the results when PCA is not used. Performance improved for all three tasks as the number of principal components is in-

Table 1: Body Pose Recognition Performance
Numbers indicate percentage of correct classification on test data.

Task (and test set size)	*Size of Eigenpose Basis*			
	10	30	100	no PCA
1: Standing vs. Reaching (543)	69.1	77.7	85.1	89.3
2: 5-way Pose Discrimin. (844)	70.9	78.7	82.6	85.4
3: Body Orientation (348)	86.2	91.4	93.1	91.4

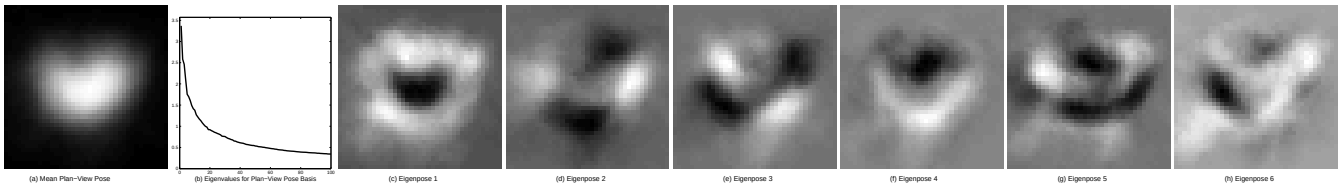

Figure 7. Plan-view ''eigenpose'' basis. (a) Mean pose, (b) Magnitude of eigenvales 1-100, (c)-(h) First six eigenposes.

creased, but even for just 10 components, recognition from a single frame of data is good enough that a simple scheme for integrating results over time could be expected to yield good pose and activity detection. Results for the 100-component basis are approximately as good as for when the normalized height templates are used directly, suggesting that 100 degrees of freedom are adequate for capturing important template differences with respect to our classification tasks.

Eigenpose-based classification still runs easily in real-time when integrated with the person tracking method of Section 3, with the reduced classification time (due to smaller input vectors) trading off against the increased time required to project height templates onto the eigenpose basis. With the smaller data vectors, we now have greater flexibility in adapting our classifiers and building new ones in real-time, in response to new data collected by the running person tracker. We plan to exploit this capability in future applications.

4.3. Activity Recognition

While body pose is best regarded as instantaneous, human "activities" are usually considered to occur over some non-instantaneous time. Let Z_H^t denote the plan-view height map measurements for the current time step t, and let $\mathbf{Z}_H^t$ denote the history of such measurements from time 0 to time t. We want to assess the likelihood that some activity A_i^t, where i indexes over a set of possible activities, is occurring at time t, given the measurement history $\mathbf{Z}_H^t$. If the set of likelihoods for all the activities is normalized to sum to 1 at each time step, then each activity likelihood may be expressed recursively in terms of a direct dependence on only the current measurements and the set of activity likelihoods from the previous time step:

$$\begin{aligned} P\left(A_i^t|\mathbf{Z}_H^t\right) &= \sum_j P\left(A_i^t|Z_H^t, A_j^{t-1}\right) P\left(A_j^{t-1}|\mathbf{Z}_H^{t-1}\right) \\ &\approx P\left(A_i^t|Z_H^t\right) P\left(A_i^t|A_i^{t-1}\right) P\left(A_i^{t-1}|\mathbf{Z}_H^{t-1}\right) + \\ &\quad \sum_{i\neq j} P\left(A_i^t|Z_H^t\right) P\left(A_i^t|A_j^{t-1}\right) P\left(A_j^{t-1}|\mathbf{Z}_H^{t-1}\right) \end{aligned} \quad (4)$$

In the last line above, we assume conditional independence of current activity probabilities on previous activity probabilities and the current measurements. Although we expect some temporal continuity in activities, we do not expect a single activity to continue over all time. Hence, we assume a small "innovation" probability β that the observed activity will change from one time step to the next. We replace $P\left(A_i^t|A_j^{t-1}\right)$ and $P\left(A_i^t|A_i^{t-1}\right)$ in equation (4) with β and $(1-\beta)$, respectively, and simplify to obtain:

$$P\left(A_i^t|\mathbf{Z}_H^t\right) = P\left(A_i^t|Z_H^t\right)\left[\beta + (1-2\beta)\, P\left(A_i^{t-1}|\mathbf{Z}_H^{t-1}\right)\right] \quad (5)$$

A given type of activity A_i is associated with a set $[\mathcal{S}_i]$ of one or more instantaneous body poses that typically occur while A_i is in progress. For instance, "sitting", "facing toward the camera", and "reaching" might characterize the activity of typing on a computer keyboard, if the camera is mounted on the computer monitor. We would like all height map templates that suggest any of the poses in $[\mathcal{S}_i]$ to contribute to the likelihood that A_i is occurring. We accomplish this by first converting the raw (unthresholded) SVM height-template classifier outputs to instantaneous pose likelihoods $P\left(S_i|Z_H\right)$. We achieve this via the method of [25], which is based on solving linear systems created from pairwise comparisons of outputs of classifiers on training data. Next, we express the activity likelihoods $P\left(A_i|Z_H\right)$ of equation (5) in terms of these pose likelihoods. In this paper, we consider only relatively simple activities whose associated set of body poses $[\mathcal{S}_i]$ may be accurately grouped as a single class by some SVM. In such cases, we may simply replace all $P\left(A_i|Z_H\right)$ in equation (5) with $P\left(S_i|Z_H\right)$.

For 6 minutes of test sequences, we attempted to detect activities corresponding to each of the pose classes of tasks 2 and 3 of Section 4.1. More precisely, classifiers for both of these tasks were applied to all height templates extracted from the sequences, with the resulting unthresholded SVM outputs being converted to pose likelihoods by the method of [25] and then substituted at each time step for the corresponding activity likelihoods in equation (5). We assume a uniform prior on activities at time $t = 0$. We detect an occurrence of an activity each time its probability exceeds a threshold θ_{act} for a duration longer than δ_{min}. Hysteresis is applied to prevent re-detection of the same activity within δ_{min} of its last detection. As a result of this process, each test sequence frame may be labeled as containing up to two different activities, where these labels have the same names as the pose classes of Section 4.1 but refer to the "activity" of people staying in the corresponding pose for an extended time period. The causal nature of this activity detection scheme, together with the fast pose classification times, allows real-time execution at 15Hz in our implementation.

Figure 8 shows activity probabilities over time for selected parts of the test sequences. These results were obtained with pose classifiers trained on normalized height templates (without PCA projection) for people not appearing in the test sequences. Occurrences of distinct activities over periods of time are visible as broad probability peaks in the graphs. In the middle of Figure 8**(a)**, some confusion between the activities of facing opposite directions is observable, due to the difficulty of the underlying pose discriminations. Brief,

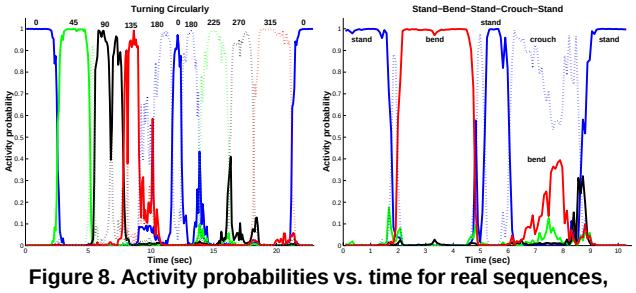

Figure 8. Activity probabilities vs. time for real sequences, with $\beta = 0.1$, and some peaks labeled. (a) Person turning in place $360°$. 8 orientation probs plotted; opposites (e.g. $0°$ and $180°$) have same color, but one solid and one dashed; (b) Person sequentially standing (blue), bending over (red), standing, crouching (dash-blue), standing.

non-troublesome confusions between crouching and standing, and crouching and bending over, appear in Figure 8**(b)**.

The locations of peaks above the θ_{act} threshold were compared with a hand-labeling of 68 activities observed in the test sequences. Using $\theta_{act} = 0.9$ and $\delta_{min} = 0.3$sec, an overall detection rate of 83%, with 7 false positives, was obtained. Almost all errors were due either to mis-classification of arm-reaching due to poor depth data, or confusion between oppositely directed body orientations. We expect this success rate to decline for sequences containing more partial occlusions and inter-person interactions, and with more activities occurring at greater distance from the camera. "Events" of one activity changing to another (e.g. from "standing upright" to "sitting") are clearly visible in the sequence of peaks of different activities in Figure 8. Event detection and modeling may be done more formally with Hidden Markov Models or other techniques.

4.4. Related Work

The fast, general classification-based scheme for body pose and activity recognition presented here is related to several recently published research efforts. Cohen and Li [7] use multiple widely-spaced cameras to compute 3D triangular visual hull surfaces from a set of silhouettes, and then employ SVMs to classify human postures based on 3D shape descriptors derived from the surfaces. Haritaoglu et. al. [11] construct 3D silhouettes from stereo disparity data, and then detect body postures and parts by comparing shape histograms derived from the silhouettes with a hierarchy of exemplars. Shakhnarovich et. al. [22] present a novel, efficient method for indexing data examples with respect to a particular estimation task, and demonstrate its application to articulated body pose recovery for new images. A related but slower technique is applied to hand pose estimation in [1]. Mori and Malik [18] apply "shape context" representations of contours to body pose estimation from silhouettes.

5. Conclusions and Future Work

We have described a method for integrating fast, accurate human body pose and activity recognition with real-time, robust multi-person tracking. Plan-view template person models lie at the core of both methods, and take advantage of the detailed 3D information available in 2D plan-view statistical projections of real-time depth data. Only a single compact stereo camera unit is needed for input, and no detailed knowledge of human body joints and pose constraints is required. All of the methods described in this paper are easily extended to systems of multiple stereo cameras, through fusion of plan-view maps produced by the individual cameras. Performance could likely be improved significantly through use of other types of plan-view statistics, more sophisticated tracking methodology (e.g. with multiple hypotheses), greater use of color and motion information, and explicit representation and analysis of dynamics in plan-view templates.

References

[1] V. Athitsos, S. Sclaroff. "Estimating 3D hand pose from a cluttered image." In *CVPR'03*.

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

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

[4] Canesta Inc. http://www.canesta.com

[5] C.-C. Chang, C.-J. Lin. "LibSVM: a library for support vector machines." 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[6] N. Checka, K. Wilson, V. Rangarajan, T. Darrell. "A probabilistic framework for multi-modal multi-person tracking." In *Wkshp. on Multi-Object Tracking*, 2003.

[7] I. Cohen, H. Li. "Inference of human postures by classification of 3D human body shape." In *Wkshp. Analysis and Modeling of Faces and Gestures*, 2003.

[8] C. Cortes, V. Vapnik. "Support-vector network." *Mach. Learning*, 20:273-297, 1995.

[9] M. Covell, A. Rahimi, M. Harville, T. Darrell. "Articulated-pose estimation using brightness- and depth-constancy constraints." *CVPR'00*.

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

[11] I. Haritaoglu, D. Beymer, M. Flickner. "$Ghost^{3D}$: detecting body posture and parts using stereo." In *Wkshp. Motion+Video Comp.*, 2002.

[12] R. Gvili, A. Kaplan, E. Ofek, G. Yahav. "Depth keying." In *SPIE Elec. Imaging*, 2003.

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

[14] M. Harville. "Stereo person tracking with adaptive plan-view templates of height and occupancy statistics". *J. Image and Vision Comp. (22)*, No. 2, pp. 127-142, 2004.

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

[16] N. Jojic, M. Turk, T. Huang. "Tracking self-occluding articulated objects in dense disparity maps." In *ICCV'99*.

[17] J. MacCormick, A. Blake. "A probabilistic exclusion principle for tracking multiple objects". In *ICCV'99*.

[18] G. Mori, J. Malik. "Estimating human body configuration using shape context matching." In *ECCV'02*.

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

[20] C. Rasmussen, G. Hager. "Joint probablistic techniques for tracking multi-part objects." In *CVPR'98*.

[21] R. Rosales, M. Siddiqui, J. Alon, S. Sclaroff. "Estimating 3D body pose using uncalibrated cameras." In *CVPR'01*.

[22] G. Shakhnarovich, P. Viola, T. Darrell. "Fast pose estimation with parameter-sensitive hashing." In *ICCV'03*.

[23] M. Turk, A. Pentland. "Eigenfaces for recognition." *J. Cog. Neuro.*, Vol. 3, No. 1, pp. 71-86, 1991.

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

[25] T.-F. Wu, C.-J. Lin, R. Weng. "Probability estimates for multi-class classification by pairwise coupling." In *NIPS'03*.

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