

## INTEGRATING PLAN-VIEW TRACKING AND COLOR-BASED PERSON MODELS FOR MULTIPLE PEOPLE TRACKING

*Luca Iocchi, Robert C. Bolles*

Artificial Intelligence Center  
SRI International, Menlo Park, CA, USA  
E-mail {iocchi,bolles}@ai.sri.com

### ABSTRACT

Tracking multiple people in a dynamic environment is important in many applications. Recent research in this area has focused either on geometric analysis or appearance models. In this paper we distinguish four types of tracking problems, and then describe an approach for combining geometric analysis and appearance-based tracking to “hold on to” people in three of these situations.

### 1. INTRODUCTION

Tracking multiple people is relevant to many applications, such as security, environmental surveillance, and marketing analysis. The ideal behavior of a people tracking system is to provide a set of tracks (possibly in an environment reference system) that are in a one-to-one correspondence with the people appearing in the field of view of the sensor, i.e. no tracks associated with different people and no multiple tracks associated with the same person. Factors that make this problem difficult include occlusion, complex light sources, and large size changes in the field of view.

We have identified the following four types of tracking problems: 1) *continuous tracking*, people are always completely visible to the sensor; their appearance can change, due to such things as changes in their pose and illumination; 2) *tracking with partial occlusions*, people are partially occluded for portions of their paths, but the system can identify pieces of them; 3) *tracking with short-term occlusions*, people are completely occluded for short periods of time; the system loses one or more of them completely, but it is able to use their positions and velocities to predict where they will reappear. 4) *tracking with extended occlusions*, people are occluded or leave the field of view of the sensor for extended periods of time; in this case, the tracking program has to rely almost exclusively on appearance models to recognize the people.

Research on tracking people has focused on two approaches. The first one uses stereo sensors (or known room geometry) to estimate the 3-D coordinates of the tracked people, maps these positions onto a floor plan, and then applies a Kalman filter to track them in the 2-D plan view

(e.g., [1, 2]). We refer to this approach as *plan-view* tracking because it concentrates on the 2-D floor plan. The second approach constructs an appearance model for each person, and then tracks them over time, adjusting the model to take into account changing lighting conditions (for example, see [3, 4, 5]). Both of these techniques have trouble when two or more people get close to each other, because the people merge into a single entity and then they move away from each other. The geometric properties and constraints available to the plan-view trackers are often insufficient for identifying people when they separate. The appearance-based techniques often generate contaminated models of the people when they get close together, because they cannot tell which portions of the image belong to which person.

Integration of plan-view tracking and appearance models has been considered in [1, 6, 7]. However, these works are not focused on the data association problem that arises when the system has to consider multiple observations and multiple tracks. They use appearance models to discriminate when people are close, thus implementing a kind of “greedy” algorithm with respect to the one proposed in this paper. Moreover, the error analysis has only been performed in terms of false positive and false negative rates, without measuring the ability of the system to assign one and only one track to every person in the scene (except for [5], where this is partially considered).

In this paper we have integrated plan-view tracking and appearance models in a novel way, by focusing on the problem of data association arising when tracking multiple people, and we provide a solution that is suitable for tracking people in situations of partial or short-term occlusions. In addition, we have performed extensive experimental evaluation of the approach introducing specific metrics and showing that it significantly increases the probability of correctly identifying people when they leave a compact group.

### 2. APPROACH

The approach is based on a calibrated stereo vision system. The main processing components are: 1) dynamic background modeling (including intensity, range, and edge in-

formation), 2) image-based foreground segmentation (integrating intensity and depth information), 3) projection of foreground points into the plan-view, 4) refined segmentation (integrating image segmentation and plan-view segmentation), 5) color based appearance model acquisition, and 6) people tracking obtained by using a Kalman Filter in an integrated space considering both position and velocity of people in the environment and their appearance. In the rest of this section we briefly describe plan-view computation, color-based person models, and integrated tracking.

## 2.1. Plan-View

In order to compute the position of people in the world, we compute a *plan view* [2] of the scene, by projecting all foreground points into the plan view reference system. This is achieved by using the stereo calibration information to map disparities into the sensor’s 3-D coordinate system and then the external calibration information to map these data in a world reference system.

For plan view segmentation, we compute a *height map*, that is a discrete map relative to the ground plane in the scene, where each cell is filled with the maximum height of all the 3-D points whose projection lies in that cell. The *height map* is then searched to detect connected components (*world blobs*). Since we are interested in person detection, *world blobs* are filtered on the basis of their size in the plan view and their height, thus removing blobs with sizes and heights inconsistent with people.

Plan View Segmentation is able to correctly deal with partial occlusions that are not detected by foreground analysis. By considering the association between pixels in the *image blobs* (detected through foreground segmentation) and *world blobs*, we are able to determine image masks corresponding to each person, which we call *person blobs*. This process allows for refining foreground segmentation in situations of partial occlusions and for correctly building person appearance models.

## 2.2. Person Models

Several techniques based on appearance models have been proposed (e.g., [3, 4, 5]). They work well when there is little or no occlusion of one person by another. The best viewpoint for these systems is typically a frontal view.

In order to effectively integrate plan-view tracking and appearance matching, one must consider: 1) computational constraints, since tracking needs to be performed in real-time and time for building and comparing models is further limited by other processing modules; 2) placement of the camera, for example, a frontal position that is widely used for some techniques is not an ideal choice for plan-view tracking, while a camera placed high on the ceiling pointing down with an angle of approximately 30 degrees with

respect to the horizon is a better choice for both plan-view tracking and person model acquisition; 3) occlusions, due to either other people or furniture in the scene.

We represent each person as a structure consisting of a head, torso, and legs. We represent each body part by a set of one-dimensional color distributions, because they capture the key color variations of people’s clothes, they are easy to construct, and two models can be compared quickly.

We denote the color distribution in the color space  $C$  for a track or an observation  $O$  by  $D_C(O)$ . To increase the robustness of color matching, we define a person model as a set of color distributions  $D_C^{i,\Delta t}(O)$  where  $i$  denotes a specific part of the detected person, and  $\Delta t$  is the time interval (5 frames in our implementation) over which the color distribution was computed.

The part of the person  $i$  is a combination of *part of the body* (hair, torso, legs) and *direction of movement* (any, front, back, left, right). The information about the part of the body is obtained in two steps: 1) we use 3D information from stereo to determine the height above the ground of the top and bottom parts of the image blob and we compute a virtual bounding box of the person from his/her head to his/her feet that takes into account vertical occlusions (i.e. the virtual bounding box “estimates” the position of the feet of a person if they are not visible); 2) then we use a generic model of a person to identify the head, torso, and legs of each tracked person. Notice that the first step helps to determine which parts of a person are visible. Information about the head is further processed with a skin-color filter that identifies and removes regions of skin so that the color distribution model focuses on the hair. Direction of movement is computed from plan-view tracking information. A special mode *any* integrates the measurements from all directions, using the assumption that people’s clothes appear approximately the same from any direction.

Finally, for efficiency reasons, we have used two one-dimensional color spaces formed by 1) the grayscale component  $Y$ ; 2) the hue component  $H$ . The  $H$  distribution is limited to the pixels that are not too dark or too bright and with sufficient color saturation. Therefore, the complete person model is

$$D_C^{\Delta t}(O) = \{(D_Y^{i,\Delta t}(O), D_H^{i,\Delta t}(O)) \mid i \in \mathcal{I}\}$$

$$\mathcal{I} = \{hair, torso, legs\} \times \{any, front, back, left, right\}.$$

The person (model) matching procedure (that is integrated in the tracking process) uses the body part information to limit the comparisons to visible parts and the direction of motion information to select the most similar models available. For example, the *front torso* of a model is compared with a *front torso* of another model if it is available, otherwise with the *back torso* if it is available, etc.

### 2.3. Integrated Tracking

Tracking is modeled in a probabilistic framework based on a mono-modal probability distribution, like in [1, 6, 7]. It is performed by maintaining a set of *people tracks*  $x_t$ , updated with the measurements of *person blobs* and *appearance models*  $z_t$ . As a difference with previous approaches, the state of each tracked object has two components: the spatial parameters, and the appearance.

The probability distribution  $p(x_t)$  is represented as a collection of Gaussians  $P_t = \{N(\mu_{i,t}, \sigma_{i,t}), w_{i,t} \mid i = 1..n\}$ , each one modeling the information about a single person.  $N(\mu_{i,t}, \sigma_{i,t})$  is a Gaussian in a multi-dimensional space representing position, velocity, and appearance of the  $i^{\text{th}}$  person that is tracked at time  $t$  and  $w_{i,t}$  a weighting factor. Similarly, observations in  $z_t$  are represented as a set of Gaussians  $Z_t = \{N(\mu'_{j,t}, \sigma'_{j,t}) \mid j = 1..m\}$  denoting position and appearance of detected people in the current frame.

The probability distribution  $p(x_t|z_t)$  is computed by a set of Kalman Filters. People position is predicted with a constant velocity model, while their appearance is updated with a constant model. This model is adequate for many normal situations in which people walk in an environment. It provides a clean way to smooth the trajectories and to hold onto a person that is partially occluded for a few frames.

For this representation, data association is an important issue to deal with. In general, at every step, the tracker must make an association between  $m$  observations (*person blobs*) and  $n$  tracked people (*people tracks*). Association is solved by computing a distance  $d_{i,j}$  between the  $i^{\text{th}}$  track  $N(\mu_{i,t|t-1}, \sigma_{i,t|t-1})$  and the  $j^{\text{th}}$  observation  $N(\mu'_{j,t}, \sigma'_{j,t})$ . Here  $N(\mu_{i,t|t-1}, \sigma_{i,t|t-1})$  is the predicted estimate of the status of the  $i^{\text{th}}$  person and the distance currently implemented is a combination of the Mahalanobis distance of the Gaussians representing position and velocity of the person and a matching procedure for the appearance models.

An association between the predicted state of the system  $P_{t|t-1}$  and the current observations  $Z_t$  is denoted with a function  $f$ , that associates each track  $i$  to an observation  $j$ , with  $i = 1..n$ ,  $j = 1..m$ , and  $f(i_1) \neq f(i_2)$ ,  $\forall i_1 \neq i_2$ . The special value  $\perp$  is used for denoting that a track is not associated to any observation (i.e.  $f(i) = \perp$ ). Let  $\mathcal{F}$  be the set of all possible associations of the current tracked people with current observations. Data association is then computed by solving the following minimization problem

$$\operatorname{argmin}_{f \in \mathcal{F}} \sum_i d_{i,f(i)}$$

where a fixed maximum value is used for  $d_{i,f(i)}$  when  $f(i) = \perp$ . Although this is a combinatorial problem, the sizes of the sets  $P_t$  and  $Z_t$  are limited (usually not greater than 4), so  $|\mathcal{F}|$  is small and this problem can be effectively solved.

The association  $f^*$ , that is the solution of this problem, is chosen for updating  $p(x_t)$ , i.e. for computing the new

status of the system  $P_t$ . During the update step the weights  $w_{i,t}$  are computed from  $w_{i,t-1}$  and  $d_{i,f^*(i)}$ , and if a weight goes below a given threshold, the person is considered *lost*. Moreover, for observations in  $Z_t$  that are not associated to any person by  $f^*$  a *new* Gaussian is entered in  $P_t$ .

The main difference with previous approaches [1, 6, 7] is that we integrate both plan-view and appearance information in the status of the system, and by solving the above optimization problem we find the best matching between observations and tracker status by considering in an integrated way the information about the position of the people in the environment and their appearance.

### 3. EXPERIMENTAL EVALUATION

Evaluation of a people tracking system is very important for determining the effectiveness of the implemented method, however only a few papers report extensive results and adequate error metrics. Systems are usually evaluated only by measuring *false positives* or false alarms (i.e. detecting non-people), and *false negatives* or missed detections [1, 8]. While these are informative measures, it is also important to consider how well the tracker keeps track of specific people. Thus two additional metrics are important: *track split*, when the track for a person is split into two or more pieces even though the person stays in the field of view of the sensor (this is called a false positive in [5]); *track switch*, when a track slips off a person and is incorrectly assigned to another person. Most of these errors are associated with occlusions, although they may happen when people are not occluded.

In this paper we describe the results of an extensive experimental evaluation that has been performed using four stereo sensors in an office environment during normal activities. These sensors monitor a corridor (from two different points of view), a document preparation room and a waiting room, observing a wide variety of behaviors, including fast walking in the corridors, stopping at a soda machine, and long visits to the document preparation area.

We focus on situations in which two or more people are close to each other, which generally causes the most problems. To this end, we have recorded video streams containing at least 2 people in the scene (this has been done automatically by the tracker system itself). We have run two versions of the tracking system on these sequences. The first one only uses plan-view tracking. The second one uses the integrated tracking including the color distribution models.

The output of the tracker is a set of *track strips* (see Figure 1), each containing 4 snapshots: 1) the first frame in which the person is noticed (white bounding box); 2) the frame in which an ID is assigned to the track (colored bounding box); 3) the last tracked frame and the projection of the trajectory followed (the track has the same color as the bounding box in the second snapshot); 4) ten frames af-



**Fig. 1.** An example of track strip.

ter the last track (to check for track splitting and false negatives). We visually inspected *track strips* to determine the kinds of errors reported above. In our analysis, we focused on *track split* and *track switch* errors, because *false positives* and *false negatives* are not related to the integration of plan-view tracking and color models and significantly fewer.

The following table shows the results of our error analysis over a number of video streams taken by different cameras. The table shows the total number of video frames analyzed, plus the number of frames containing two or more “close” people (in parenthesis), the number of tracks and the number of tracks involving close people (in parenthesis). The second row presents the number of splitting and switching errors. The first value is the number of errors made by the plan-view tracking system and the second is the number of errors made by the integrated system.

Frames	<b>190412 (3573)</b>	Tracks	<b>627 (80)</b>
Track split	<b>4/5</b>	Track switch	<b>35/12</b>

For video streams with no intersecting tracks or with frames containing people that are relatively far away from each other, we have no errors, since plan-view tracking is able to deal with such situations. All the errors occurred within the 3573 frames (out of 190412), in which two people were close. This value determines approximately the number of times in which a “difficult” data association problem has been solved by the tracker. The integrated tracker reduces the number of errors from 39 to 17, which is a significant improvement.

From the analysis of the detected errors, we have identified some typical erroneous situations. For example, in situations where two people are close each other (more precisely when one occludes the other for more than a half of his/her body) and when they are either too close or too far from the sensor for several frames, the system fails to produce a good segmentation and the person model is affected by errors due to the presence of pixels belonging to more than one person. Consequently, when people move away the tracking system has erroneous information, which leads to an incorrect assignment of tracks to people.

Computational performance of a tracking system is important because a low frame rate produces larger gaps to

be bridged by the tracker. The overall tracking system can process high-resolution 640x480 images in 50 to 120 ms (depending on the number of tracked people in the scene) on a Pentium 4 2.80GHz CPU, thus allowing for real-time implementation and good performance in tracking.

#### 4. REFERENCES

- [1] D. Beymer and K. Konolige, “Real-time tracking of multiple people using stereo,” in *Proc. of IEEE Frame Rate Workshop*, 1999.
- [2] T. Darrell, D. Demirdjian, N. Checka, and P. F. Felzenszwalb, “Plan-view trajectory estimation with dense stereo background models,” in *Proc. of 8th Int. Conf. on Computer Vision (ICCV’01)*, 2001, pp. 628–635.
- [3] I. Haritaoglu, D. Harwood, and L. S. Davis, “An appearance-based body model for multiple people tracking,” in *Proc. of 15th Int. Conf. on Pattern Recognition (ICPR’00)*, 2000.
- [4] A. W. Senior, “Tracking with probabilistic appearance models,” in *Proc. of ECCV workshop on Performance Evaluation of Tracking and Surveillance Systems (PETS’02)*, 2002, pp. 48–55.
- [5] R. Cucchiara, C. Grana, G. Tardini, and R. Vezzani, “Probabilistic people tracking for occlusion handling,” in *Proc. of 17th Int. Conf. on Pattern Recognition (ICPR’04)*, 2004.
- [6] J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. Hale, and S. Shafer, “Multi-camera multi-person tracking for easyliving,” in *Proc. of Int. Workshop on Visual Surveillance*, 2000.
- [7] A. Mittal and L. S. Davis, “M2Tracker: A multi-view approach to segmenting and tracking people in a cluttered scene using region-based stereo,” in *Proc. of the 7th European Conf. on Computer Vision (ECCV)*, 2002.
- [8] M. T. Yang, Y. C. Shih, and S. C. Wang, “People tracking by integrating multiple features,” in *Proc. of 17th Int. Conf. on Pattern Recognition (ICPR’04)*, 2004.