Multi-camera Tracking and Segmentation of Occluded People on Ground Plane Using Search-Guided Particle Filtering

Kyungnam Kim[1,2] and Larry S. Davis[1]

[1] Computer Vision Lab, University of Maryland, College Park, MD 20742
{knkim, lsd}@umiacs.umd.edu
http://www.umiacs.umd.edu/~knkim
[2] IPIX Corporation, Sunset Hills Rd. Suite 410, Reston, VA, 20190

Abstract. A multi-view multi-hypothesis approach to segmenting and tracking multiple (possibly occluded) persons on a ground plane is proposed. During tracking, several iterations of segmentation are performed using information from human appearance models and ground plane homography. To more precisely locate the ground location of a person, all center vertical axes of the person across views are mapped to the top-view plane and their intersection point on the ground is estimated. To tackle the explosive state space due to multiple targets and views, iterative segmentation-searching is incorporated into a particle filtering framework. By searching for people's ground point locations from segmentations, a set of a few good particles can be identified, resulting in low computational cost. In addition, even if all the particles are away from the true ground point, some of them move towards the true one through the iterated process as long as they are located nearby. We demonstrate the performance of the approach on several video sequences.

1 Introduction

Tracking and segmenting people in cluttered or complex situations is a challenging visual surveillance problem since the high density of objects results in occlusion. Elgammal and Davis [20] presented a general framework which uses maximum likelihood estimation and occlusion reasoning to obtain the best arrangement for people. To handle more people in a crowded scene, Zhao and Nevatia [9] described a model-based segmentation approach to segment individual humans in a high-density scene using a Markov chain Monte Carlo method.

When a single camera is not sufficient to detect and track objects due to limited visibility or occlusion, multiple cameras can be employed. There are a number of papers which address detection and tracking using overlapping or non-overlapping multiple views, for example, [6, 7, 19]. M$_2$Tracker [19], which is similar to our work, used a region-based stereo algorithm to find 3D points inside an object, and Bayesian pixel classification with occlusion analysis to segment people occluded in different levels of crowd density. Unlike M$_2$Tracker's requirement of having calibrated stereo pairs of cameras, we do not require strong

A. Leonardis, H. Bischof, and A. Pinz (Eds.): ECCV 2006, Part III, LNCS 3953, pp. 98–109, 2006.

calibration, but only a ground plane homography. For outdoor cameras, it is practically very challenging to accurately calibrate them, so 3D points at a large distance from a camera are difficult to measure accurately.

Our goal is to 'segment' and 'track' people on a ground plane viewed from multiple overlapping views. To make tracking robust, multiple hypothesis trackers, such as *particle filter* [12], are widely used [17, 16]. However, as the numbers of targets and views increase, the state space of combination of targets' states increases exponentially. Additionally, the observation processes for visual tracking are typically computationally expensive. Previous research has tried to solve this state space explosion issue as in [13, 1, 14, 8, 15]. We also designed our tracker to solve this issue. Each hypothesis is refined by iterative mean-shift-like multi-view segmentation to maintain mostly "good" samples, resulting in lower computational cost.

This paper is organized as follows. Sec.2 presents a human appearance model. A framework for segmenting and tracking occluded people moving on a ground plane is presented in Sec.3. In Sec.4, the multi-view tracker is extended to a multi-hypothesis framework using particle filtering. We demonstrate the experimental results of the proposed approach on video sequences in Sec.5. Conclusion and discussion are given in the final section.

2 Human Appearance Model

First, we describe an appearance color model as a function of height that assumes that people are standing upright and are dressed, generally, so that consistently colored or textured color regions are aligned vertically. Each body part has its own color model represented by a color distribution. To allow multimodal densities inside each part, we use kernel density estimation.

Let $M = \{\mathbf{c}_i\}_{i=1...N_M}$ be a set of pixels from a body part with colors $\mathbf{c}_i$. Using Gaussian kernels and an independence assumption between d color channels, the probability that an input pixel $\mathbf{c} = \{c_1, ..., c_d\}$ is from the model M is estimated as

$$p_M(\mathbf{c}) = \frac{1}{N_M} \sum_{i=1}^{N_M} \prod_{j=1}^{d} \frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{1}{2}\left(\frac{c_j - c_{i,j}}{\sigma_j}\right)^2} \tag{1}$$

In order to handle illumination changes, we use normalized color $(r = \frac{R}{R+G+B}, g = \frac{G}{R+G+B}, s = \frac{R+G+B}{3})$ or Hue-Saturation-Value (HSV) color space with a wider kernel for 's' and 'V' to cope with the higher variability of these lightness variables. We used both the normalized color and HSV spaces in our experiments and observed similar performances.

Viewpoint-independent models can be obtained by viewing people from different perspectives using multiple cameras. A related calibration issue was addressed in [2, 5] since each camera output of the same scene point taken at the same time or different time may vary slightly depending on camera types and parameters. We used the same type of cameras and observed there is almost no difference between camera outputs except for different illumination levels (due to shadow and orientation effects) depending on the side of person's body. This level of variability is accounted for by our color model.

3 Multi-camera Multi-person Segmentation and Tracking

3.1 Foreground Segmentation

Given image sequences from multiple overlapping views including people to track, we start by performing detection using background subtraction to obtain the foreground maps in each view. The codebook-based background subtraction algorithm [18] is used. Its shadow removal capability increases the performance of segmentation and tracking.

Each foreground pixel in each view is labelled as the best matching person (i.e., the most likely class) by Bayesian pixel classification as in [19]. The posterior probability that an observed pixel $\mathbf{x}$ (containing both color $\mathbf{c}$ and image position (x, y) information) comes from person k is given by

$$P(k|\mathbf{x}) = \frac{P(k)P(\mathbf{x}|k)}{P(\mathbf{x})} \tag{2}$$

We use the color model in Eq.1 for the conditional probability $P(\mathbf{x}|k)$. The color model of the person's body part to be evaluated is determined by the information of $\mathbf{x}$'s position as well as the person's ground point and full-body height in the camera view (See Fig.1(a)). The ground point and height are determined initially by the method defined subsequently in Sec.3.2.

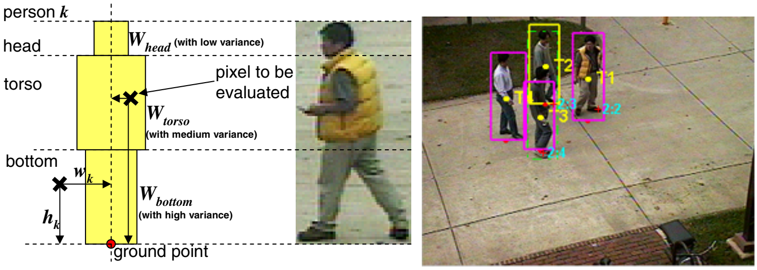
The prior reflects the probability that person k occupies pixel $\mathbf{x}$. Given the ground point and full-body height of the person, we can measure $\mathbf{x}$'s height from the ground and its distance to the person's center vertical axis. The occupancy probability is then defined by

$$O_k(h_k(\mathbf{x}), w_k(\mathbf{x})) = P[w_k(\mathbf{x}) < W(h_k(\mathbf{x}))] = 1 - \mathrm{cdf}_{W(h_k(\mathbf{x}))}(w_k(\mathbf{x})) \tag{3}$$

where $h_k(\mathbf{x})$ and $w_k(\mathbf{x})$ are the height and width of $\mathbf{x}$ relative to the person k. h_k and w_k are measured relative to the full height of the person. $W(h_k(\mathbf{x}))$ is the person's height-dependent width and $\mathrm{cdf}_W(.)$ is the cumulative density function for W. If $\mathbf{x}$ is located at distance $W(h_k(\mathbf{x}))$ from the person's center at a distance W, the occupancy probability is designed so that it will be exactly 0.5 (while it increases or decreases as $\mathbf{x}$ move towards or move away from the center).

The prior must also incorporate possible occlusion. Suppose that some person l has a lower ground point than a person k in some view. Then the probability that l occludes k depends on their relative positions and l's (probabilistic) width. Hence, the prior probability $P(k)$ that a pixel $\mathbf{x}$ is the image of person k, based on this occlusion model, is

$$P(k) = O_k(h_k, w_k) \prod_{g_y(k) < g_y(l)} (1 - O_l(h_l, w_l)) \tag{4}$$


person k
head
torso
bottom
W_{head} (with low variance)
pixel to be evaluated
W_{torso} (with medium variance)
w_k
h_k
W_{bottom} (with high variance)
ground point
2
1
1
3
2:3
2:2
2:4

(a) (b)

Fig. 1. (a) Illustration of appearance model, (b) Bounding box detection

where $g_y(k)$ is the y-location of the ground point of k and $\mathbf{x}$ is omitted for simplicity (i.e., $h_k = h_k(\mathbf{x})$ and $w_k = w_k(\mathbf{x})$). The best class k^* is determined by maximum a posteriori (MAP) estimation: $k^* = \arg\max_k P(k)P(\mathbf{x}|k)$. Finally, the foreground maps are segmented into the best matching persons based on their appearance models and occlusion information.

3.2 Model Initialization and Update

Full automatic tracking is enabled by initializing the human appearance model when a person is detected in a view by searching for isolated foreground blobs (See Fig.1(b)). In order to get a bounding box of a person from the foreground map, we used the object detection technique in [3]. The bounding boxes in the figure were created when the blobs are isolated before. For the case when a person does not constitute an isolated blob, a manual selection is employed.

The full-body height of a person is initialized upon model creation and is updated during segmentation. In some cases, fixing the average height scaled by the y-location of the ground point provides a robust height measurement when the segmentation is unreliable. When the unclassified pixels (those having a probability in Eq.1 lower than a given threshold) constitute a connected component of non-negligible size, a new appearance model should be created.

3.3 Multi-view Integration

Ground Plane Homography. The segmented blobs across views are integrated to obtain the ground plane locations of people. The correspondence of a human across multiple cameras is established by the geometric constraints of planar homographies. For N_V camera views, $N_V(N_V - 1)$ homography matrices can possibly be calculated for correspondence; but in order to reduce the computational complexity we instead reconstruct the top-view of the ground plane on which the hypotheses of peoples' locations are generated.

Integration by Vertical Axes. Given the pixel classification results from Sec.3.1, a ground point of a person could be simply obtained by detecting the lowest point of the person's blob. However those ground points are not reliable due to the errors from background subtraction and segmentation.

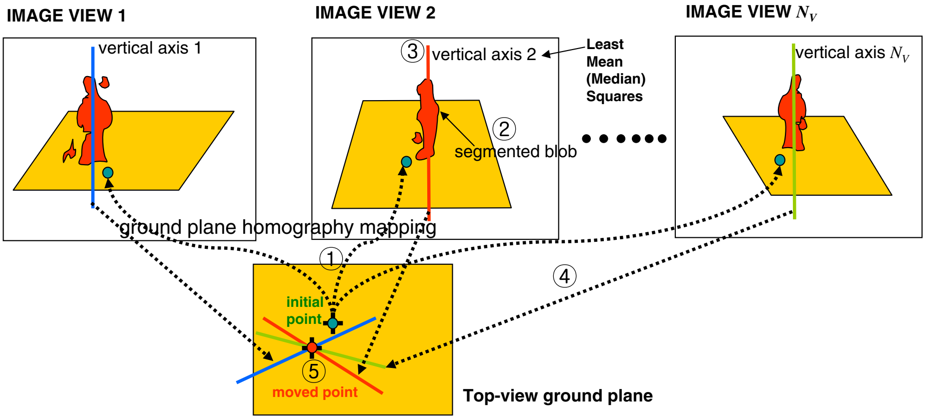


Fig. 2. All vertical axes of a person across views intersect at (or are very close to) a single point when mapped to the top-view

We, instead, develop a localization algorithm that employs the center vertical axis of a human body, which can be estimated more robustly even with poor background subtraction [11]. Ideally, a person's body pixels are arranged more of less symmetrically about a person's central vertical axis. An estimate of this axis can be obtained by Least Mean Squares of the perpendicular distance between the body pixel and the axis as in ③ in Fig.2. Alternatively, the Least *Median* Squares could be used since it is more robust to outliers.

The homographic images of all the vertical axes of a person across different views intersect at (or are very close to) a single point (the location of that person on the ground) when mapped to the top-view (See [11]). In fact, even when the ground point of a person from some view is occluded, the top-view ground point integrated from all the views is obtainable if the vertical axis is estimated correctly. This intersection point can be calculated by minimizing the perpendicular distances to the axes. Fig.2 depicts an example of reliable detection of the ground point from the segmented blobs of a person. The N_v vertical axes are mapped to the top-view and transferred back to each image view. Let each axis L_i be parameterized by two points $\{(x_{i,1}, y_{i,1}), (x_{i,2}, y_{i,2})\}_{i=1 \ldots N_V}$. When mapped to the top-view by homography as in ④ in Fig.2, we obtain $\{(\hat{x}_{i,1}, \hat{y}_{i,1}), (\hat{x}_{i,2}, \hat{y}_{i,2})\}_{i=1 \ldots N_V}$. The distance of a ground point (x, y) to the axis is written as $d\left((x, y), L_i\right) = \frac{|a_i x + b_i y + c_i|}{\sqrt{a_i^2 + b_i^2}}$where $a_i = \hat{y}_{i,1} - \hat{y}_{i,2}$, $b_i = \hat{x}_{i,2} - \hat{x}_{i,1}$, and $c_i = \hat{x}_{i,1}\hat{y}_{i,2} - \hat{x}_{i,2}\hat{y}_{i,1}$. The solution is the point that minimizes a weighted sum of square distances:

$$(x^*, y^*) = \arg\min_{(x,y)} \sum_{i=1}^{N_V} w_i^2 d^2((x, y), L_i) \tag{5}$$

The weight w_i is determined by the segmentation quality (confidence level) of the body blob of L_i (We used the pixel classification score in Eq.2).

If a person is occluded severely by others in a view (i.e., the axis information is unreliable), the corresponding body axis from that view will not contribute heavily to the calculation in Eq.5. When only one axis is found reliably, then the lowest body point along the axis is chosen.

To obtain a better ground point and segmentation result, we can iterate the segmentation and ground-point integration process until the ground point converges to a fixed location within a certain bound ϵ. That is, given a set of initial ground-point hypotheses of people as in ① in Fig.2, segmentation in Sec.3.1 is performed (②), and then newly moved ground points are obtained based on multi-view integration (④ and ⑤). These new ground points are an input to the next iteration. 2-3 iterations gave satisfactory results for our data sets.

There are several advantages of our approach. Even though a person's ground point is invisible or there are segmentation and background subtraction errors, the robust final ground point is obtainable once at least two vertical axes are correctly detected. When total occlusion occurs from one view, robust tracking is possible using the other views' information if available; visibility of a person can be maximized if cameras are placed at proper angles. Since the good views for each tracked person are changing over time, our algorithm maximizes the effective usage of all available information across views. By iterating the multi-view integration process, a ground point moves to the optimal position that explains the segmentation results of all views. This nice property is used, in the next section, for a small number of hypotheses to explore in a large state space that incorporates multiple persons and multiple views.

4 Extension to Multi-hypothesis Tracker

Next, we extend our single-hypothesis tracker to one with multiple hypotheses. A single hypothesis tracker, while computationally efficient, can be easily distracted by occlusion or nearby similarly colored objects.

As the number of targets and views increase, the state space of combination of targets' states increases exponentially. Additionally, the observation processes for visual tracking are typically very expensive. We would, therefore, choose to employ techniques that require small numbers of particles.

The iterative segmentation-searching presented in Sec.3 is naturally incorporated with a particle filtering framework. There are two advantages - (1) By searching for a person's ground point from a segmentation, a set of a few good particles can be identified, resulting in low computational costs, (2) Even if all the particles are away from the true ground point, some of them will move towards the true one as long as they are initially located nearby. This does not happen generally with particle filters, which need to wait until the target "comes to" the particles.

Our final algorithm of segmentation and tracking is presented with a particle filter overview and our state space, dynamics, and observation model.

4.1 Overview of Particle Filter, State Space, and Dynamics

The key idea of particle filtering is to approximate a probability distribution by a weighted sample set $S = \{(\mathbf{s}^{(n)}, \pi^{(n)}) | n = 1...N\}$. Each sample, $\mathbf{s}$, represents one hypothetical state of the object, with a corresponding discrete sampling probability π, where $\sum_{n=1}^{N} \pi^{(n)} = 1$. Each element of the set is then weighted in terms of the observations and N samples are drawn with replacement, by choosing a particular sample with probability $\pi_t^{(n)} = P(\mathbf{z}_t | \mathbf{x}_t = \mathbf{s}_t^{(n)})$.

In our particle filtering framework, each sample of the distribution is simply given as $s = (x, y)$ where x, y specify the ground location of the object in the *top-view*. For multi-person tracking, a state $\mathbf{s}_t = (\mathbf{s}_{1,t}, ..., \mathbf{s}_{N_p,t})$ is defined as a combination of N_p single-person states. Our state transition dynamic model is a random walk where a new predicted single-person state is acquired by adding a zero mean Gaussian with a covariance $\mathbf{\Sigma}$ to the previous state. Alternatively, the velocity $\dot{x}, \dot{y}$ or the size variable *height* and *width* can be added to the state space and then a more complex dynamic model can be applied if relevant.

4.2 Observation

Each person is associated with a reference color model $\mathbf{q}^\star$ which is obtained by histogram techniques [16]. The histograms are produced using a function $b(\mathbf{c}_i) \in \{1, ..., N_b\}$ that assigns the color vector $\mathbf{c}_i$ to its corresponding bin. We used the color model defined in Sec.2 to construct the histogram of the reference model in the normalized color or HSV space using N_b (e.g., $10 \times 10 \times 5$) bins to make the observation less sensitive to lighting conditions.

The histogram $\mathbf{q}(C) = \{q(u; C)\}_{u=1...N_b}$ of the color distribution of the sample set C is given by

$$q(u; C) = \eta \sum_{i=1}^{N_C} \delta[b(\mathbf{c}_i) - u] \tag{6}$$

where u is the bin index, δ is the Kronecker delta function, and η is a normalizing constant ensuring $\sum_{u=1}^{N_b} q(u; C) = 1$. This model associates a probability to each of the N_b color bins.

If we denote $\mathbf{q}^\star$ as the reference color model and $\mathbf{q}$ as a candidate color model, $\mathbf{q}^\star$ is obtained from the stored samples of person k's appearance model as mentioned before while $\mathbf{q}$ is specified by a particle $\mathbf{s}_{k,t} = (x, y)$. The sample set C in Eq.6 is replaced with the sample set specified by $\mathbf{s}_{k,t}$. The top-view point (x, y) is transformed to an image ground point for a certain camera view v, $H_v(\mathbf{s}_{k,t})$, where H_v is a homography mapping the top-view to the view v. Based on the ground point, a region to be compared with the reference model is determined. The pixel values inside the region are drawn to construct $\mathbf{q}$. Note that the region can be constrained from the prior probability in Eq.4, including the occupancy and occlusion information (i.e., by picking pixels such that $P(k) > Threshold$, typically 0.5). In addition, as done in pixel classification, the color histograms are separately defined for each body part to incorporate the spatial layout of the color distribution. Therefore, we apply the likelihood as the sum of the histograms associated with each body part.

Then, we need to measure the data likelihood between $\mathbf{q}^\star$ and $\mathbf{q}$. The Bhattacharyya similarity coefficient is used to define a distance d on color histograms: $d[\mathbf{q}\star, \mathbf{q}(\mathbf{s})] = \left[1 - \sum_{u=1}^{N_b} \sqrt{q \star (u) q(u; \mathbf{s})}\right]^{\frac{1}{2}}$. Thus, the likelihood $(\pi_{v,k,t})$ of person k consisting of N_r body parts at view v, the actual view-integrated likelihood $(\pi_{k,t})$ of a person $\mathbf{s}_{k,t}$, and the final weight of the particle $(\pi_{k,t})$ of a concatenation of N_p person states are respectively given by:

$$\pi_{v,k,t} \propto e^{\sum_{r=1}^{N_r} -\lambda d^2[\mathbf{q}_r^\star, \mathbf{q}_r(H_v(\mathbf{s}_{k,t}))]}, \quad \pi_{k,t} = \mathit{\Pi}_{v=1}^{N_V} \pi_{v,k,t}, \quad \pi_t = \mathit{\Pi}_{k=1}^{N_p} \pi_{k,t} \qquad (7)$$

where λ is a constant which can be experimentally determined.

4.3 The Final Algorithm

The algorithm below combines the particle filtering framework described before and the iterated segmentation-and-search in Sec.3 into a final multi-view multi-target multi-hypothesis tracking algorithm. Iteration of segmentation and multi-view integration moves a predicted particle to an a better position on which all the segmentation results of the person agree. The transformed particle is re-sampled for processing of the next frames.

Algorithm for Multi-view Multi-target Multi-hypothesis tracking

```
I.   From the "old" sample set $S_{t-1} = \{\mathbf{s}_{t-1}^{(n)}, \pi_{t-1}^{(n)}\}_{n=1,...,N}$ at time $t-1$,
     construct the new samples as follows:
II.  Prediction: for $n = 1, ..., N$, draw $\tilde{\mathbf{s}}_t^{(n)}$ from the dynamics. Iterate Step
     III to IV for each particle $\tilde{\mathbf{s}}_t^{(n)}$.
III. Segmentation & Search
     $\tilde{\mathbf{s}}_t = \{\tilde{\mathbf{s}}_{k,t}\}_{k=1...N_p}$ contains all persons' states. The superscript $(n)$ is
     omitted through the Observation step.
     i.   for $v \leftarrow 1$ to $N_V$ do
              (a) For each person $k$, $(k = 1...N_p)$, transform the top-view point
                  $\tilde{\mathbf{s}}_{k,t}$ into the ground point in view $v$ by homography, $H_v(\tilde{\mathbf{s}}_{k,t})$
              (b) perform segmentation on the foreground map in view $v$ with
                  the occlusion information according to Sec2.
          end for
     ii.  For each person $k$, obtain the center vertical axes of the person across
          views, then integrate them on the top-view to obtain a newly moved
          point $\tilde{\mathbf{s}}_{k,t}^*$ as in Sec3.
     iii. For all persons, if $\|\tilde{\mathbf{s}}_{k,t} - \tilde{\mathbf{s}}_{k,t}^*\| < \varepsilon$, then go to the next step. Otherwise,
          set $\tilde{\mathbf{s}}_{k,t} \leftarrow \tilde{\mathbf{s}}_{k,t}^*$ and go to Step III-i.
```

```
IV. Observation
    i. for $v \leftarrow 1$ to $N_V$ do
           For each person $k$, estimate the likelihood $\pi_{v,k,t}$ in view $v$ according to Eq.7. $\tilde{\mathbf{s}}_{k,t}$ needs to be transferred to view $v$ by mapping through $H_v$ for evaluation. Note that $\mathbf{q}_r(H_v(\tilde{\mathbf{s}}_{k,t}))$ is constructed only from the non-occluded body region.
       end for
    ii. For each person $k$, obtain the person likelihood $\pi_{k,t}$ by Eq.7.
    iii. Set $\pi_t \leftarrow \Pi_{k=1}^{N_p} \pi_{k,t}$ as the final weight for the multi-person state $\tilde{\mathbf{s}}_t$.
V. Selection: Normalize $\{\pi_t^{(n)}\}_i$ so that $\sum_{n=1}^{N} \pi_t^{(n)} = 1$.
   For $i = n...N$, sample index $a(n)$ from discrete probability $\{\pi_t^{(n)}\}_i$ over $\{1...N\}$, and set $\mathbf{s}_t^{(n)} \leftarrow \tilde{\mathbf{s}}_t^{a(n)}$.
VI. Estimation: the mean top-view position of person $k$ is $\sum_{n=1}^{N} \pi_t^{(n)} \mathbf{s}_{k,t}^{(n)}$.
```

5 Experiments

We now present experimental results obtained on outdoor and indoor multi-view sequences to illustrate the performance of our algorithm.

The results on the indoor sequences are depicted in Fig.3. The bottom-most row shows how the persons' vertical axes are intersecting on the top-view to obtain their ground points. Small orange box markers are overlaid on the images of frame 198 for determination of the camera orientations. Note that, in the figures of 'vertical axes', the axis of a severely occluded person does not contribute to localization of the ground point. When occlusion occurs, the ground points being tracked are displaced a little from their correct positions but are restored to the correct positions quickly. Only 5 particles (one particle is a combination of 4 single-person states) was used for robust tracking. Those indoor cameras could be easily placed properly in order to maximize the effectiveness of our multi-view integration and the visibility of the people.

Fig.4(a) depicts the graph of the total distance error of people's tracked ground points to the ground truth points. It shows the advantage of multiple views for tracking of people under severe occlusion.

Fig.4(b) visualizes the homographic top-view images of possible vertical axes. A vertical axis in each indoor image view can range from 1 to each maximum image width. 7 transformed vertical axes for each view are depicted for visualization. It helps to understand how the vertical axis location obtained from segmentation affects ground point (intersection) errors on the top-view. When angular separation is close to 180 degrees (although visibility is maximized), the intersection point of two vertical axes transformed to top-view may not be reliable because a small amount of angular perturbation make the intersection point move dramatically.

The outdoor sequences (3 views, 4 persons) are challenging in that three people are wearing similarly-colored clothes and the illumination conditions change

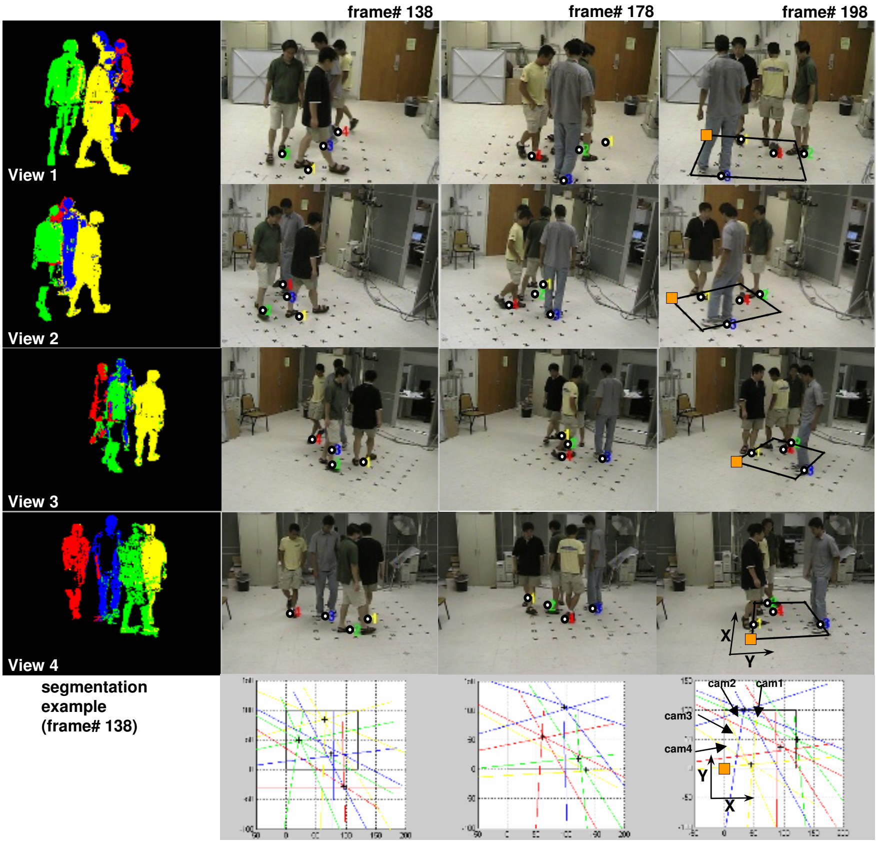
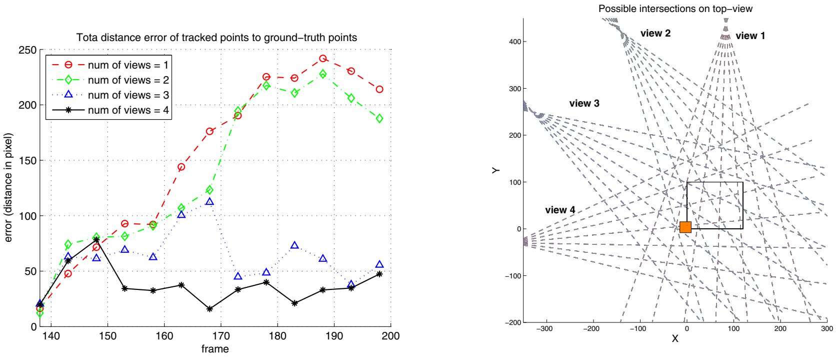


Fig. 3. The tracking results of 4-view indoor sequences from Frame 138 to 198 are shown with the segmentation result of Frame 138



(a) Total distance error of persons' tracked ground points to the ground truth points

(b) Homographic images all different vertical axes

Fig. 4. Graphs for indoor 4 camera views

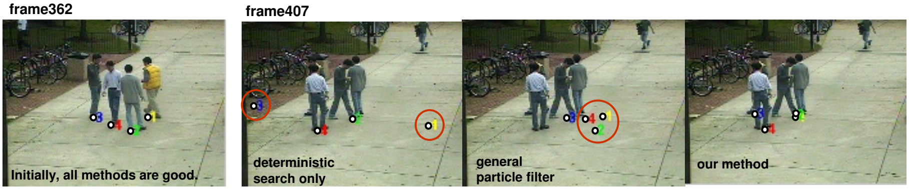


Fig. 5. Comparison on three methods: While the deterministic search with a single hypothesis (persons 2 and 4 are good, cannot recover lost tracks) and the general particle filter (only person 3 is good, insufficient observations during occlusion) fail in tracking all the persons correctly, our proposed method succeeds with a minor error. The view 2 was only shown here. The proposed system tracks the ground positions of people afterwards over nearly 1000 frames.

over time, making segmentation difficult. In order to demonstrate the advantage of our approach, single hypothesis (deterministic search only) tracker, general particle filter, and particle filter with deterministic search by segmentation (our proposed method) are compared in Fig.5. The number of particles used is 15.

6 Conclusion and Discussion

A framework to segment and track people on a ground plane is presented. Human appearance models are used to segment foreground pixels obtained from background subtraction. We developed a method to effectively integrate segmented blobs across views on a top-view reconstruction, with a help of ground plane homography. The multi-view tracker is extended to a multi-hypothesis framework using particle filtering.

We have illustrated results on challenging videos to show the usefulness of the proposed approach. Segmentation of people is expedited by processing subsampled foreground pixels and robust tracking is achieved without loss of accuracy; it was actually confirmed by the experiments with sub-sampling by factors from 2 to 70.

In order to make our system more general, several improvements could be considered, such as handling different observed appearances of an object across views [2], extending the method to tracking in environments which are not planar, or including automatic homography mapping [10].

Acknowledgements

We thank the U.S. Government for supporting the research described in this paper. We are also grateful to Bohyung Han for very helpful suggestions about the particle filtering framework.

References

1. Zhuowen Tu, Song-Chun Zhu, “Image segmentation by data-driven Markov chain Monte Carlo,” *IEEE Transactions on PAMI*, Volume 24, Issue 5, May 2002.
2. Omar Javed, Khurram Shafique and Mubarak Shah, “Appearance Modeling for Tracking in Multiple Non-overlapping Cameras,” *IEEE CVPR 2005.*
3. A. W. Senior, “Tracking with Probabilistic Appearance Models,” in *ECCV workshop on Performance Evaluation of Tracking and Surveillance Systems*, 2002, pp 48-55.
4. I. Haritaoglu, D. Harwood, L.S. Davis, “W^4: real-time surveillance of people and their activities” *IEEE Transactions on PAMI*, Volume: 22, Issue: 8, Aug 2000.
5. Chang, T.H., Gong, S., Ong, E.J., “Tracking Multiple People Under Occlusion Using Multiple Cameras,” BMVC 2000.
6. J. Kang, I. Cohen, G. Medioni, “Multi-Views Tracking Within and Across Uncalibrated Camera Streams”, *ACM SIGMM 2003 Workshop on Video Surveillance*, 2003.
7. Javed O, Rasheed Z, Shafique K and Shah M, “Tracking Across Multiple Cameras With Disjoint Views,” *The Ninth IEEE ICCV*, 2003.
8. D. Comaniciu, P. Meer, “Mean Shift: A Robust Approach toward Feature Space Analysis,” *IEEE Trans. on PAMI*, Vol. 24, No. 5, 603-619, 2002.
9. T. Zhao, R. Nevatia, “Bayesian human segmentation in crowded situations,” *CVPR*, June 2003.
10. Chris Stauffer, Kinh Tieu. “Automated multi-camera planar tracking correspondence modeling,” *CVPR*, vol. 01, no. 1, p. 259, 2003.
11. Min Hu, Jianguang Lou, Weiming Hu, Tieniu Tan, “Multicamera correspondence based on principal axis of human body,” *IEEE ICIP*, 2004.
12. S. Arulampalam, S. Maskell, N. J. Gordon, and T. Clapp, “A Tutorial on Particle Filters for On-line Non-linear/Non-Gaussian Bayesian Tracking”, *IEEE Transactions of Signal Processing*, Vol. 50(2), pages 174-188, February 2002.
13. J. Deutscher, A. Blake and Ian Reid, “Articulated Body Motion Capture by Annealed Particle Filtering,” *CVPR* 2000.
14. J. Sullivan and J. Rittscher, “Guiding random particles by. deterministic search,” *ICCV*, 2001.
15. C. Shan, Y. Wei, T. Tan, F. Ojardias, “Real Time Hand Tracking by Combining Particle Filtering and Mean Shift,” *IEEE International Conference on Automatic Face and Gesture Recognition*, 2004.
16. P. Perez, C. Hue, J. Vermaak, and M. Gangnet, “Color-based probabilistic tracking,” *ECCV* 2002.
17. Michael Isard and Andrew Blake, “CONDENSATION – conditional density propagation for visual tracking,” *Int. J. Computer Vision*, 29, 1, 5–28, 1998.
18. K. Kim, T.H. Chalidabhongse, D. Harwood, L. Davis, “Real-time foreground-background segmentation using codebook model,”, *Real-Time Imaging*, June 2005.
19. Anurag Mittal and Larry S. Davis, “M_2Tracker: A Multi-View Approach to Segmenting and Tracking People in a Cluttered Scene,” *IJCV*, Vol. 51 (3), 2003.
20. A. Elgammal and L. S. Davis, “Probabilistic Framework for Segmenting People Under Occlusion”, *ICCV*, Vancouver, Canada July 9-12, 2001.