ROBUST MULTI-CAMERA 3D PEOPLE TRACKING WITH PARTIAL OCCLUSION HANDLING

Huan Jin[1,3] *and Gang Qian*[2,3]

[1]Dept. of Computer Science and Engineering and
[2]Dept. of Electrical Engineering and
[3]Arts, Media and Engineering Program,
Arizona State University, Tempe, AZ 85287, USA
{Huan.Jin,Gang.Qian}@asu.edu

ABSTRACT

This paper presents an approach to robust 3D people tracking using multiple synchronized and calibrated cameras. The goal is to improve people tracking accuracy when the subjects being tracked partially occlude each other in some of the camera views. To achieve this goal, Monte Carlo fine-tuning is deployed to rectify 3D people locations obtained from partially occluded image observations. In our approach, Gaussian mixture models and axis-parallel ellipsoids are used to represent the appearance and the 3D body structures of the subjects, respectively. Related parameters are learned off-line. Experimental results obtained using real videos illustrate that the proposed approach is capable of accurate and robust 3D people tracking under partial or complete occlusions.

Index Terms— 3D people tracking, Kalman filter, partial occlusion, Monte Carlo fine-tuning, triangulation

1. INTRODUCTION

Robust 3D people tracking is a challenging task for computer vision. Accurate 3D people location can greatly improve the performance of high-level tasks such as trajectory reasoning, activity inference and event understanding.

The major challenge for robust people tracking is occlusion. To overcome this challenge, many multi-camera tracking approaches have been presented recently, for example [1, 2, 3, 4, 5, 6, 7]. These approaches assume that people walk on a ground plane. Hence people tracking can be simplified into a 2D tracking problem on the ground plane. In addition, tracking results from multiple cameras can be connected through homography among cameras with respect to the ground plane to further handle occlusions. For instance, in [1] ground plane occupancy is first estimated from foreground images to provide a basis for further people tracking based on color and refinement using global trajectory optimization. However, the performance of these approaches will deteriorate if this assumption is violated, e.g. when the people being tracked are having various activities such as jumping or sitting down. The epipolar constraint has also been used to recover tracking in a camera view when the target has been occluded in this view [6]. Although such approaches using the epipolar constraint do not constrain activities of the subjects, they suffer in the presence of false candidates along the epipolar line. To overcome these limitations, our previous work [8] presented a simple way to compute 3D location from 2D image information with complete occlusion handling. A 2D circular search region obtained from 3D predication is used to reject outliers (image regions similar to the target) when an object is fully occluded in a camera view.

Although these approaches can provide a robust tracking under complete occlusions, the 3D people tracking accuracy will deteriorate when the subject is partially occluded. Such scenarios can easily take place when multiple subjects are present in the tracking space. In such case, the center of 2D partially occluded object is difficult to obtain correctly, which results in an inaccurate 3D tracking.

In this paper, we present a robust people localization system by 3D tracking of torso, and propose an effective Monte Carlo based method to improve tracking accuracy in the presence of partial occlusions. A Gaussian-based color model is learned to represent human appearance. Human torso is modeled as an axis-parallel ellipsoid whose parameters are learned using gradient-descent method in advance. A number of 3D samples are drawn from a Gaussian distribution based on the estimated 3D location from triangulation algorithm and 3D human torso structure. These samples will be back-projected onto all available camera views to compute matching scores indicating how well these samples match color segmentations in each view. The fine-tuned 3D location is computed by weighting all the samples. Kalman filtering is then used to predict and smooth final tracking results.

A closely related method to handle occlusions was presented in [9] for people tracking with 3D target and environment models using a particle filter. Similar to our method, a 3D ellipsoid is also adopted in their work to represent the human. Partial occlusions are solved using the matching score in particle filtering. The ellipsoid's structure parameters are incorporated in the state space. However, as the number of targets and views increases, the state space of combination of target's states increase dramatically. To make the tracking robust and efficient, we learn 3D ellipsoid structure based on the matching score in advance. The estimated 3D tracking results are efficiently computed from 2D tracking results. Monte Carlo fine-tuning is used to verify and refine these 3D tracking results. Additionally, the target's color information is not used in [9]. So, it is error-prone to maintain target identity solely based on the continuity of tracking trajectories.

There are certain assumptions in our approach. Firstly, at least three cameras are required to integrate multiple 2D tracking results. Secondly, we assume that people are conducting activities mostly with upright upper body pose. Thirdly, we assume that partial occlusions are only caused by targets being tracked. The current approach cannot effectively handle situations where occlusions were introduced by other objects/people that are not tracked by the system. Finally, the illumination variation of the space needs to be

maintained within an acceptable range.

2. HUMAN BODY MODELING

2.1. Color Modeling

In order to track subjects in varying lighting conditions, we adopt hue, saturation, value (HSV) color space because it separates out hue (color) from saturation and brightness and hue is relatively reliable to identify different subjects in uneven lighting environment. Considering that the clothes of the subjects can contain a mixture of colors, we represent the color appearance of a target using a mixture of Gaussian in the HSV space. For example, the color distribution of target j is represented a 3D Gaussian C_j with mean μ_j and covariance matrix Σ_j in 3D color space. Thus, the probability that a pixel HSV value ξ belongs to target j can be represented in (1).

$$P(\xi|C_j) = \frac{1}{(2\pi)^{\frac{3}{2}}|\Sigma_j|^{\frac{1}{2}}} exp\{-\frac{1}{2}(\xi-\mu_j)^T\Sigma_j^{-1}(\xi-\mu_j)\} \quad (1)$$

In our application, since the subjects wear shirts with a dominant color, color distribution can be approximated using a single Gaussian. A color-based target model is trained in the tracking initialization stage. We manually select subject's torso region in each camera view. The corresponding color model is statistically computed based on all the pixels' HSV values in the initial torso region, and is then associated with each person. Outlier points, which can be caused by image noise and specular highlights, have little influence upon this representation. This probabilistic based color model also makes the people tracking insensitive to small variation of ambient illumination.

2.2. Structure Modeling

In our approach, subjects' 3D location is obtained by 3D tracking of the torso. Assume that most of time people are in poses with upper body upright, such as walking, running, sitting down. An axis-parallel ellipsoid is sufficient to represent human torso. We model human torso as an upright ellipsoid with two parameters (h, r), which are ellipsoid's height and radius respectively.

To reduce the dimensionality of the tracking state vector, we learn the structure parameters (h, r) using training data. Namely, given color segmentation sequences $\mathbf{E} = \{E_i^{(c)}\}_{i=1..N,c=1..M}$, where i is image index and c is camera view index, the best ellipsoid structure is obtained in (2):

$$(h, r) = \underset{(h,r)}{\operatorname{argmax}} \prod_{i=1}^{N}\prod_{c=1}^{M} m(X_i, h, r, E^c) \quad (2)$$

where $m(X_i, h, r, E^c)$ is a matching score indicating how well the projection produced using the structure and position parameters can match the silhouette of the color-segmented torso image. We use a matching score m [9], shown in (3), based on the ellipsoid's 3D location X, structure parameters (h, r) and the human torso segmentation E^c in camera c.

$$m(X, h, r, E^c) = \frac{|F^c(X, h, r) \cap E^c|}{|F^c(X, h, r) \cup E^c|} \quad (3)$$

where $F^c(X, h, r)$ indicates the projected 2D ellipse area in camera view c. For multiple people, $F(\mathbf{X}, \mathbf{h}, \mathbf{r})$ is the union of all ellipse areas.

Given camera projection matrix $\mathbf{P}$, computed from camera extrinsic and intrinsic parameters, it's straightforward to map an ellipsoid quadric $\mathbf{Q}$ (which is specified by X_i, h, r) to an ellipse conic $\mathbf{C}$ in a 2D image plane using (4) [10]:

$$\mathbf{C} = (\mathbf{P}\mathbf{Q}^{-1}\mathbf{P}^T)^{-1} \quad (4)$$

Hence, $F(\mathbf{X}, \mathbf{h}, \mathbf{r})$ can be easily obtained.

In practice, for each person, a walking sequence including body rotation is recorded. We extract 20 frames to represent the sequence. The ground-truth of 3D location for each frame is computed using triangulation. To achieve this learning, we use gradient-descent method to find the optimal 3D structure parameters.

3. MULTI-CAMERA TRACKING

An overview of multi-camera 3D people tracking is illustrated by the diagram shown in Fig. 1. Tracking is initialized manually.

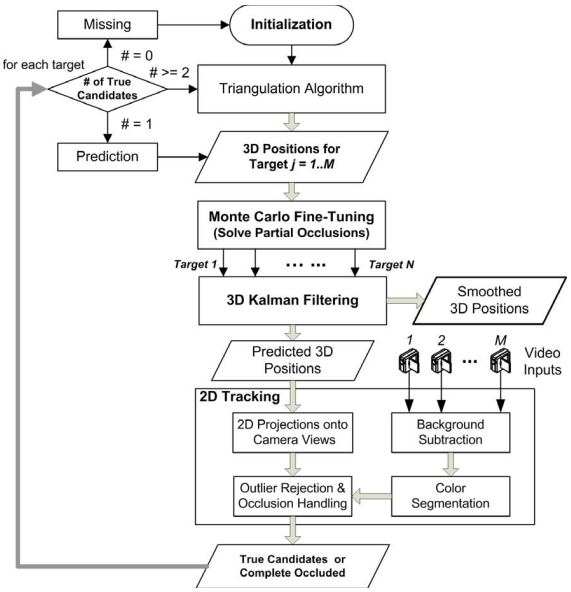


Fig. 1. Diagram of 3D People Tracking

3.1. 3D Kalman filtering

In order to obtain smoothed 3D people location and conduct reliable location prediction, Kalman filters are deployed. Each target is assigned a Kalman filter to perform tracking smoothing and prediction. The first-order motion model is adopted to represent motion dynamics. The state vector $\mathbf{X}_t$ of the Kalman filter includes the target's 3D position (x, y, z) and velocity $(\dot{x}, \dot{y}, \dot{z})$. The Kalman filter uses $\mathbf{Z}_t$, the result from the Monte Carlo fine-tuning as input and then provides smoothed location estimate for the current time t and location prediction for the next time instant $t + 1$. The smoothed location is output by the tracking system as the final 3D tracking result and the predicted subject location at $t+1$ is used to specify 2D search region in the upcoming image frame.

3.2. Robust 2D Tracking and Outlier Rejection

The goal of 2D tracking is to detect people locations within the search region and to reject outliers which have similar color to the people being tracked. We firstly do color segmentation to extract candidate blobs. Outlier rejection is then performed only within the search region which is suggested by Kalman prediction.

3.2.1. Color Segmentation

Given image sequences from multiple camera views including people to track, we firstly use background subtraction to obtain the foreground maps in each view. In our approach, we adopt the codebook-based background subtraction algorithm [11] which is able to remove shadow. Then, we do simple but effective color segmentation based on the target color model. All foreground pixels in each view are further tested on each target color model using (1). The pixel i is labeled as target j if $P(\xi_i|C_j) > T_j$, where T_j is the threshold for target j. After that, we apply morphological operation *Opening* to further remove noise and fill in small holes. Finally, *Connected Component Analysis* is employed to get all pixel-connected blobs.

3.2.2. Outlier Rejection

In order to remove outliers introduced by cluttered and varying scenes, we specify a rectangular search area to detect the true target. The search area $\mathcal{R}_t$ at time t is defined as $\mathcal{R}_t = (x_t, y_t, H_t, W_t)$, where (x_t, y_t) is the predicted 2D position obtained from the projection of 3D Kalman prediction. The search rectangle size (H_t, W_t) is adaptively determined by (5).

$$\begin{pmatrix} H_t \\ W_t \end{pmatrix} = (\alpha s_{t-1} + \beta) \begin{pmatrix} B^h_{t-1} + \delta^h \\ B^w_{t-1} + \delta^w \end{pmatrix} \quad (5)$$

where s_{t-1} is the target's 2D speed at time $t-1$, α and β are scaling factors. (B^h_{t-1}, B^w_{t-1}) are the height and width of the bounding box of the foreground target from color segmentation at time $t-1$, and (δ^h, δ^w) are constants that prevent the search rectangle vanish due to complete occlusion . Note that (B^h_{t-1}, B^w_{t-1}) are not the size of the search rectangle at $t-1$. The search rectangle is adaptive to the target's 2D speed since high speed introduces more uncertainties. β provides a lower bound on the size of the search rectangle when the target stops. In practice, α and β are manually set to be 0.5 and 1.2 respectively. (δ^h, δ^w) is set to (40,30). Any target candidate after color segmentation outside of $\mathcal{R}_t$ will be rejected as an outlier. It's still possible to have some candidates within $\mathcal{R}_t$. The one closest to (x_t, y_t) will then be selected as the 2D target location in this camera view at the current time instant.

3.3. Monte Carlo Fine-Tuning

Partial occlusion is more common than complete occlusion in multiple people tracking. Inaccurate 2D centroid localization of the color segmentation due to partial occlusion results in a drift in 3D location computation. To accurately track partially occluded people, we propose Monte Carlo based method to fine-tune the 3D observation location.

At each time instant t, we draw N Gaussian samples $\mathbf{S}_t = \{\mathbf{s}^i_t, \pi^i_t\}^N_{i=1}$ based on the mean $\tilde{\mathbf{X}}_t$ and covariance $\mathbf{\Sigma}_t$. $\tilde{\mathbf{X}}_t$ is the triangulation result and $\mathbf{\Sigma}_t$ is directly obtained from Kalman filter's previous state covariance matrix $\mathbf{P}_{t-1}$. Each sample $\mathbf{s}^i_t$ is associated with a weight π^i_t.

To compute π^i_t, in each camera view, the projection of the ellipsoid at position $\mathbf{s}^i_t$ is first obtained using (4). Then, the matching score given by (3)is evaluated for each camera view to see how well the projection can match the target foreground from color segmentation. Finally, the sample weight π^i_t is given by the product of the matching scores of all camera views:

$$\pi^i_t = \prod_{c=1}^{M} m(\mathbf{s}^i_t, \mathbf{h}, \mathbf{r}, E^c_t) \quad (6)$$

Since one ground position cannot be occupied by two persons, sample drawing needs to be controlled by this valid rule. For a sample $\mathbf{s} = \{(x_j, y_j, z_j)\}_{j=1\ldots K}$ where K is the number of people, if there exists two ellipsoids u and v such that

$$\sqrt{(x_u - x_v)^2 + (y_u - y_v)^2} < r_u + r_v \quad (7)$$

we simply set the corresponding $\pi = 0$. After computing the weight of each sample, the fine-tuned 3D location $\mathbf{Z}_t$, which serves as Kalman filter's input, is estimated in (8)

$$\mathbf{Z}_t = \frac{\sum_{i=1}^N \pi^i_t \mathbf{s}^i_t}{\sum_{i=1}^N \pi^i_t} \quad (8)$$

3.4. Handling of Complete Occlusion

In our approach, if there are at least two cameras detecting the target, we can use triangulation to compute 3D location based on visible camera views. To continue 2D tracking and outlier rejection in the camera view where the person is completely occluded, we project smoothed 3D location onto the camera image plane and assume the projected 2D location is the occluded target's position. Therefore, the person identity can be maintained when he/she reappears in the view.

When the target is detected in only one camera view, we are not able to recover the 3D location of object reliably [10]. Thus, during a short period of time, we approximate the 3D location $\tilde{X}_t$ using 3D target location $\tilde{X}_{t-1}$ and the current 2D target location $\tilde{x}_t$ provided by the view in which the object is visible [8]. Therefore, the target can be continuingly tracked over the period when it's visible in only one view. However, if this situation lasts over a long period of time, the estimation error will be accumulated and the 3D location estimate will drift away from the true value.

4. EXPERIMENTAL RESULTS

Our tracking system consists of three color CCD cameras (Dragonfly2,Point Grey research) and a PC (Pentium IV 3GHz, 1GB RAM). Image resolution is 320 × 240. The cameras are calibrated in advance. In our experiment, human structure parameters are learned in advance (Person 1: $h_1 = 0.0825$, $r_1 = 0.0473$; Person 2: $h_2 = 0.075$, $r_2 = 0.0425$, $1unit \cong 427cm$). The number of fine-tuning samples N was set to 200. To better cover the activity space, we set large FOV for each camera. So, lens distortion [10] is also handled for every frame.

In order to evaluate the tracking performance of our system, we extract two video sequences from a long video clip. Sequence 1 is used to demonstrate the capability of tracking under partial occlusions. Sequence 2 is to demonstrate the capability of tracking under complete occlusions and outlier rejection for one camera view.

In Fig. 2, two people with different upper body color are walking in the activity space causing partial occlusions (*frame 500, 518 and 543*) for some cameras. The row C1, C2 and C3 show the real camera images. Each frame is superimposed with a rectangular search region and target identity number. Note that the size of search region is adaptively changed according to color segmentation and 2D moving speed. The row P1, P2 and P3 show the estimated 2D corresponding ellipses projected from 3D ellipsoids that represent human upper bodies. In *frame 518*, two person are walking very closely. Person 1 is partially occluded in camera C1 and person 2 is also partially occluded in camera C3 at the same time. However, the correct 2D projection results in P1 and P3 show that 3D location is rectified using Monte Carlo fine-tuning, even though color segmentation results will give an inaccurate estimate of the centers of two people in 2D image planes. Fig. 3 shows two people's tracking trajectories

in XYZ directions for sequence 1. The trajectory is smooth in three dimensions. The results show that our approach can handle partial occlusion effectively and provide an accurate 3D people location.

In Fig. 4, we show three frames (*frame 589, 621 and 738*) in camera C1 from sequence 2. In *frame 589*, person 2 is completely occluded by person 1 in camera C1. A red '+' is issued in current camera view to localize 2D position suggested by 3D Kalman prediction. In *frame 738*, person 1 is walking close to a mirror which is in upper right corner of the image. The 2D projection in P1 shows that the false people image introduced by the mirror is rejected as an outlier and person identity is maintained. There results show that a robust tracking can be achieved under complete occlusion or false candidates.

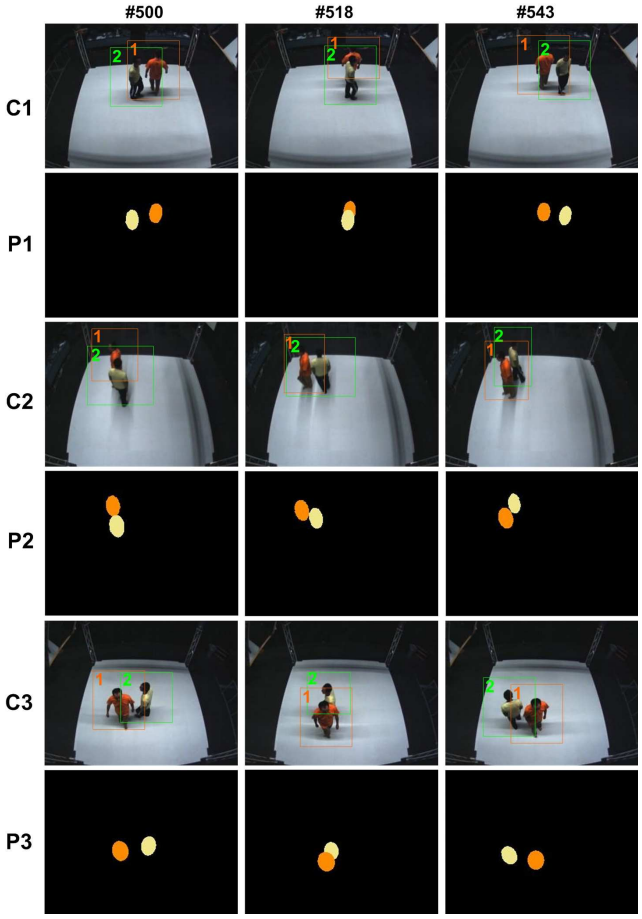
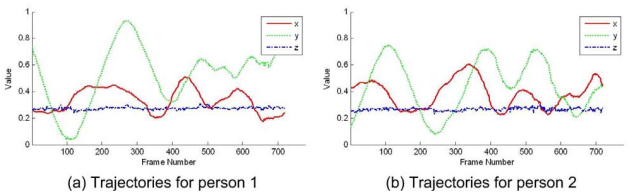


Fig. 2. Tracking under partial occlusions



(a) Trajectories for person 1 (b) Trajectories for person 2

Fig. 3. Tracking trajectory in X-Y-Z dimensions ($1unit \cong 427cm$)

5. CONCLUSIONS

In this paper, we present an approach to multi-view 3D people tracking. The proposed approach can improve tracking accuracy when the subjects partially occlude each other in some of the camera views. Experiment results show that our approach can robustly provide accurate and consistent tracking in the presence of partial or complete occlusions.

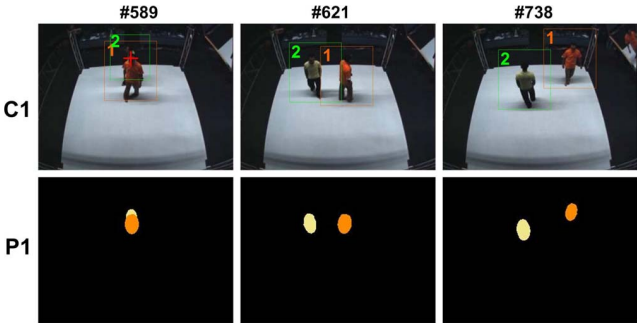


Fig. 4. Tracking under complete occlusion and outlier rejection in one camera view

6. REFERENCES

[1] J. Berclaz, F. Fleuret, and P. Fua, "Robust people tracking with global trajectory optimization," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2006.

[2] J. Black, T. Ellis, and P. Rosin, "Multi view image surveillance and tracking," in *IEEE Workshop on Motion and Video Computing*, 2002.

[3] O. Javed, Z. Rasheed, O. Alatas, and M. Shah, "Knight: A real time surveillance system for multiple overlapping and non-overlapping cameras," in *4th Int'l Conf. on Multimedia and Expo*, 2003.

[4] J. Kang, I. Cohen, and G. Medioni, "Tracking people in crowded scenes across multiple cameras," in *Asian Conf. on Computer Vision*, 2004.

[5] A. Mittal and L. Davis, "M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene using region-based stereo," *Int'l J. of Computer Vision (IJCV)*, 2003.

[6] K. Nummiaro, E. Koller-Meier, T. Svoboda, D. Roth, and L. Van Gool, "Color-based object tracking in multi-camera environments," in *25th Pattern Recognition Symposium, DAGM'03*, Berlin, Germany, September 2003, pp. 591–599.

[7] Z. Yue, S. Zhou, and R. Chellappa, "Robust two-camera tracking using homography," in *IEEE Intl Conf. on Acoustics, Speech and Signal Processing*, 2004, vol. 3, pp. 1–4.

[8] H. Jin, G. Qian, and S. Rajko, "Real-time multi-view 3d object tracking in cluttered scenes," in *2nd Int'l Symposium on Visual Computing (ISVC)*, Nov. 2006.

[9] T. Osawa, X. Wu, K. Wakabayashi, and T. Yasuno, "Human tracking by particle filtering using full 3d model of both target and environment," in *18th Int'l Conf. on Pattern Recognition*, August 2006.

[10] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[11] K. Kim, T. H. Chalidabhongse, D. Harwood, and L. Davis, "Real-time foreground-background segmentation using codebook model," *Real-time Imaging*, vol. 11, pp. 167–256, June 2005.