3D Human Tracking for Visual Monitoring

Tatsuya Osawa[†]*, Xiaojun Wu, Kaoru Wakabayashi, and Hideki Koike*

Abstract

In this article, we introduce a three-dimensional (3D) human tracking system for visual monitoring. It tracks human movement in three dimensions with high accuracy. A 3D environmental model that replicates the 3D structure of the real world is introduced to handle cases in which some objects obstruct the camera's view, i.e., occlusions. Experiments show that our system can stably track multiple humans who interact with each other and enter and leave the monitored area. This system is expected to be useful not only for surveillance but also for collecting marketing data.

1. Introduction

There is a lot of interest in visual monitoring systems to ease growing public fears. Tracking humans is one of the most critical aspects of visual monitoring because the movements of humans correlate well with human behavior. A human tracking system is expected to be useful not only for surveillance but also for collecting marketing data (**Fig. 1**).

One major problem for a practical three-dimensional (3D) tracking system is that other objects in the environment can obstruct the camera's view of the target. This phenomenon is known as occlusion. In practice, any tracking system should be robust when there are: (1) mutual occlusions caused by interacting targets, (2) occlusions caused by fixed objects in the environment, and (3) variable targets to be tracked representing entry to and departure from the monitored area.

Many methods of tracking targets on a two-dimensional (2D) image plane in these situations have been proposed. Methods of stably tracking multiple targets in the presence of occlusions caused by fixed objects in the scene and mutual occlusions have been reported [1], [2]. To track variable interacting targets, the MCMC (Markov chain Monte Carlo) particle filter has been used [3]–[5].

Compared with the 2D approach, the 3D approach is more effective in accurately estimating position in space and is more effective for handling the above-mentioned situations. However, few studies have attempted to utilize a 3D approach. 3D position has been estimated by integrating the tracking results on a 2D ground plane from multiple stereo camera modules [6], [7]. Unfortunately, if the tracking result from one stereo camera module is false, the whole system becomes unstable. 3D positions of humans in very cluttered environments have been tracked using multiple cameras located far from each other [8], [9]. However, in those studies, an ideal environment with no objects other than humans was assumed because the approaches were based on volume intersection.

The problems with the above methods mainly arise from the difficulty of solving the inverse problem (reconstructing 3D information from 2D images). To tackle the inverse problem, a method of tracking humans by directly predicting their 3D positions in a 3D environment model and evaluating the predictions using 2D images from multiple cameras was investigated [10]. This approach avoids the inverse problem because 3D information is not explicitly reconstructed. However, the computation cost is very high because multiple humans must be tracked by using several single-object trackers in parallel.

[†] NTT Cyber Space Laboratories
Yokosuka-shi, 239-0847 Japan
Email: osawa.tatsuya@lab.ntt.co.jp

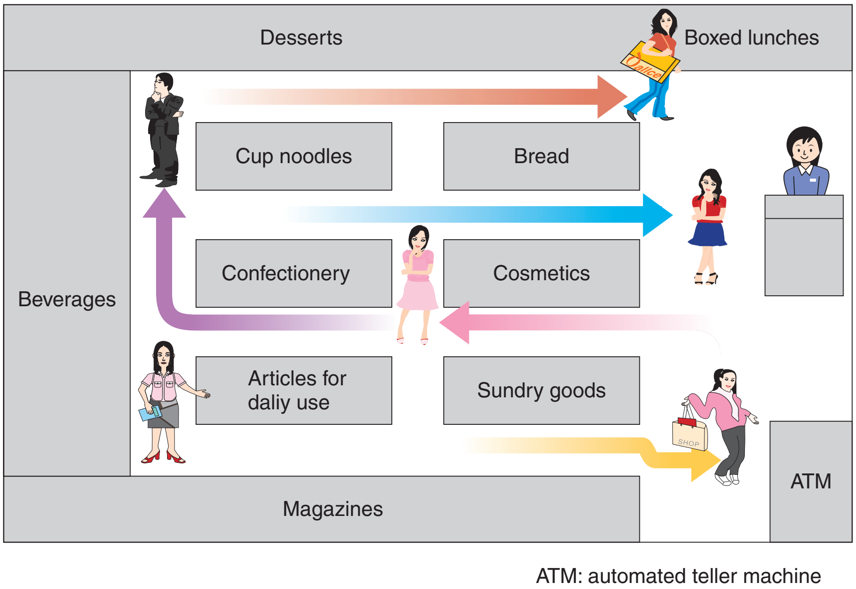


Fig. 1. Application for marketing.

Fig. 2. Example of an image sequence.

In this article, we present a new approach to the stable tracking of variable interacting targets in the presence of severe occlusions in 3D space. We formulate the state of multiple targets as the union state space of all the targets and recursively estimate the multibody configuration and the position of each target in the 3D space by using the framework of trans-dimensional MCMC [11]. In surveillance applications, environmental information is very useful because surveillance cameras are stationary in the environment.

This article is organized as follows: Section 2 introduces the 3D environment model and its application to tracking, Section 3 describes our tracking algorithm. Section 4 describes our experiments and presents our conclusions. Future work is mentioned in Section 5.

2. Our 3D environmental model

Our approach is to construct a 3D environment model that replicates the real-world's 3D structure in advance of tracking. We use this 3D environment model to handle occlusions caused by fixed objects in environment. We also define the entry and departure areas in the 3D environment model to enable reliable estimation of the number of targets in the monitored area because the areas through which people enter or leave are definitely fixed in the environment. If we know about such areas, we can suppress needless predictions of the appearance and disappearance of humans.

We construct a 3D environmental model that replicates the real-world's 3D structure from the image sequences captured by all the cameras in the environment. We capture an image sequence and move the viewpoint to the position used in the tracking process. We use the combination of a factorization method and multiview stereo [12] to reconstruct dense 3D points. A typical image sequence is shown in **Fig. 2**.

After reconstructing the 3D points from all the cameras, we integrate all 3D points in world coordinates and detect the ground plane as the plane with the largest area by applying the 3D Hough transform [13]; the 3D points are converted so that the X-Y plane lies on the ground. This allows us to use 2D coordinate values (x, y) to express the 3D positions of humans because human motion is strongly restricted to the 2D ground plane. A typical set of integrated 3D points is shown in **Fig. 3**.

Finally, the 3D surface is approximated by a triangular mesh; depending on the environment, we can set the entrance and departure areas manually.

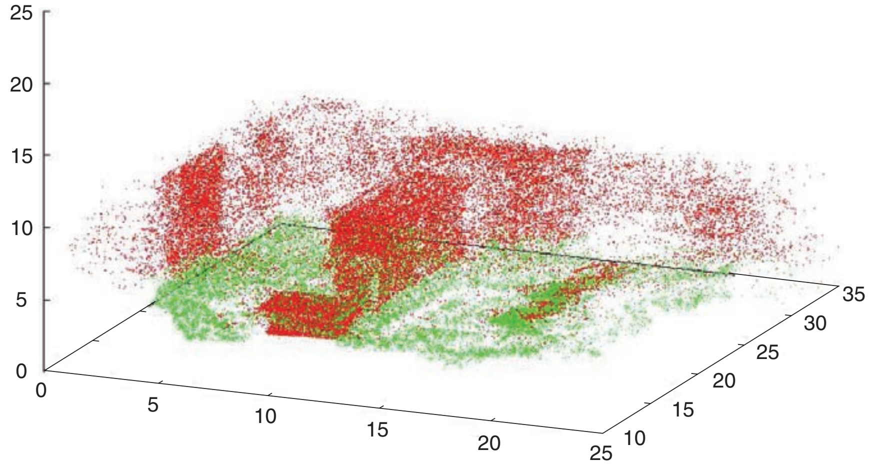


Fig. 3. Reconstructed 3D points.

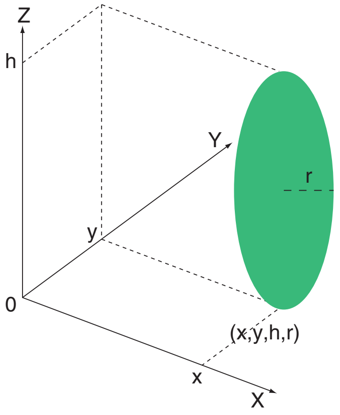


Fig. 4. 3D human model.

3. Tracking with the 3D environmental model

In this section, we introduce a multiple human tracking method with a 3D environmental model. Tracking means the sequential estimation of the state of multiple humans S_t; it represents the 3D position of humans in the 3D environmental model at time t. State S_t is estimated by directly predicting the 3D position of humans in the 3D environmental model and the predictions are validated by using 2D images from multiple cameras. In addition, we can restrict the predictions of the positions of humans because the 3D environmental model replicates the real-world's 3D structure and we are aware of the entrance and departure areas and of occlusions caused by fixed objects in the environment.

3.1 Handling multiple humans

We track humans using a 3D model that represents the human body as an ellipsoid. The state of human i is represented as a 4D vector $\mathrm{M}_i = (x_i, y_i, h_i, r_i)$, where (x_i, y_i) is the position on the 2D ground plane and (h_i, r_i) give the ellipsoid's height and radius (this allows us to handle shape differences between individuals), as shown in **Fig. 4**.

The state of multiple humans is defined as the union state space of all the humans. Consider a system tracking K people in the t-th image frame. S_t is represented as the $4K$-dimensional vector $S_t = (M_1, M_2, \ldots, M_K)$.

3.2 MCMC-based tracking algorithm

The number of dimensions of the state space estimates varies with the number of humans being tracked. To deal with this trans-dimensional state space, we use an estimation algorithm based on trans-dimensional MCMC [11].

First, in each time step, we compute the initial state of the Markov chain at time t using the state of previous time S_{t-1} according to the motion model.

After initialization, we generate $B + P$ new samples by changing the current state depending on a random selection of move type (MCMC sampling step) to obtain P samples because the first B samples are assumed to vary widely. We use four move types: entry of the target into the space, departure of the target, update of the target's position, and update of the target's shape. We decide to accept or reject a new sample as a new state by computing the likelihood of the new sample. After $B + P$ iterations, we compute state S_t as the maximum a posteriori (MAP) state using samples generated using the last P samples. The flow of our MCMC-based tracking algorithm is given below.

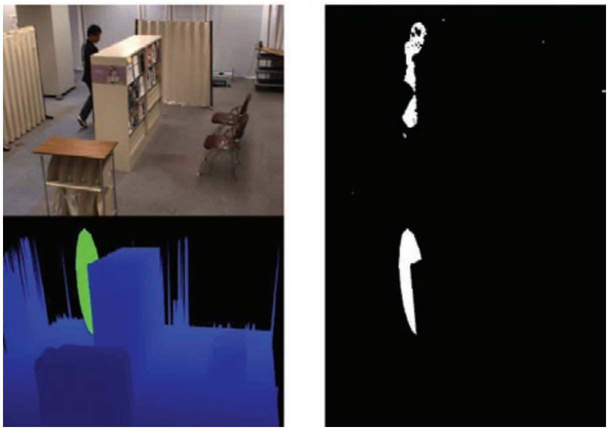


Fig. 5. Real camera image (top left), background-subtracted image (top right),virtual image (bottom left), and ellipsoid-detection image (bottom right).

(1) Initialize the MCMC sampler

(2) Perform MCMC sampling

a) Select the move type randomly and generate a new sample

b) Compute the likelihood of the new sample

c) Decide whether to accept or reject the new sample by computing the acceptance ratio

(3) Estimate the MAP

(1) *MCMC sampler initialization*: We initialize the MCMC sampler at time t using the state of previous time S_{t-1} according to the motion model, for which we use simple linear prediction. The initial state of the Markov chain at time t $\hat{S}_{t,0}$ is computed by

$$\hat{S}_{t,0} = S_{t-1} + V_{t-1}, \qquad (1)$$

where V_{t-1} is the previous velocity of the state. If the system tracks K humans at time t–1, vector V_{t-1} has $4K$ dimensions.

(2) *Move type*: We use the following move types to traverse the union state space:

1) Target Addition (entry of a new target)

2) Target Removal (departure of the target)

3) Position Update (update of the target's position)

4) Shape Update (update of the target's shape)

In each iteration, one of the above move types is selected randomly. If the present state $\hat{S}_{t,k}$ is null space, we always select Target Addition; Target Removal is not selected unless $\hat{S}_{t,k}$ includes a human state positioned in an entrance or departure area.

a) Target Addition: A new human state M_n is added to the present state $\hat{S}_{t,k}$. The position of M_n is limited to the entrance and departure areas because we assume that humans cannot enter or leave except via these areas. We use μr, μh as the average shape parameters of humans. The new human state M_n is generated by

$$M_n = (\delta_x, \delta_y, N(\mu_h, \sigma_h), N(\mu_r, \sigma_r)), \qquad (2)$$

where δ_x, δ_y are white noise limited to the entrance and departure areas and $N(\mu_r, \sigma_r)$, $N(\mu_h, \sigma_h)$ are Gaussian noises with means μ_r, μ_h and variances σ_r, σ_h, respectively.

b) Target Removal: A selected human state M_i is removed from the present state $\hat{S}_{t,k}$. Target i is randomly selected from the targets in the entrance and departure areas.

c) Position Update: The position parameters (x_i, y_i) of randomly selected human state M_i are updated by

$$(x_i, y_i) = (x_i + N(0, \sigma_x), y_i + N(0, \sigma_y)), \qquad (3)$$

where $N(0, \sigma_x)$, $N(0, \sigma_y)$ Gaussian noises with mean *0*, *0* and variance σ_x, σ_y, respectively.

d) Shape Update: The shape parameters (h_i, r_i) in randomly selected human state M_i are updated by

$$(h_i, r_i) = (h_i + N(0, \sigma_h), r_i + N(0, \sigma_r)), \qquad (4)$$

where $N(0, \sigma_h)$, $N(0, \sigma_r)$ are Gaussian noises with mean 0, 0 and variance σ_h, σ_r, respectively.

3) *Likelihood of the state*: The state is simulated by using 3D models of humans and the environment. We capture this scene using virtual cameras that have the same camera parameters as real cameras. The likelihood of the state is computed by comparing the real camera image with the corresponding virtual camera image. We can predict how the target will be occluded by objects in the environment because we use a full 3D model that includes the targets and the environment.

A real camera image, background subtracted image, virtual camera image, and ellipsoid detection image are shown in **Fig. 5**. We compare the background-subtracted image with the ellipsoid detection image using

$$V(S) = \frac{1}{C} \sum_{N=1}^{C} \frac{\Sigma_{k,l} B_{gN}(k,l) \cap S_{mN}(k,l)}{\Sigma_{k,l} B_{gN}(k,l) \cup S_{mN}(k,l)}, \qquad (5)$$

where C is the number of cameras and $B_{gN}(k, l)$ and $S_{mN}(k, l)$ are the (k, l) pixels of the background-subtracted image (as seen from camera N) and the corresponding ellipsoid detection image, respectively.

In addition, we introduce the following penalty functions using 3D information.

a) Penalty based on position in the environment: The probability that humans are floating above the

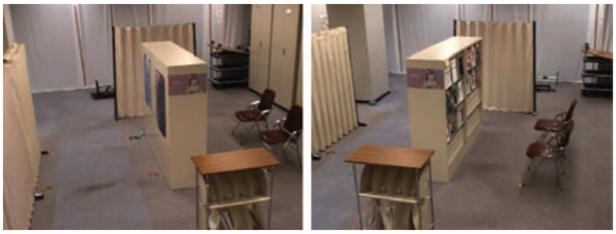


Fig. 6. Images from the two cameras.

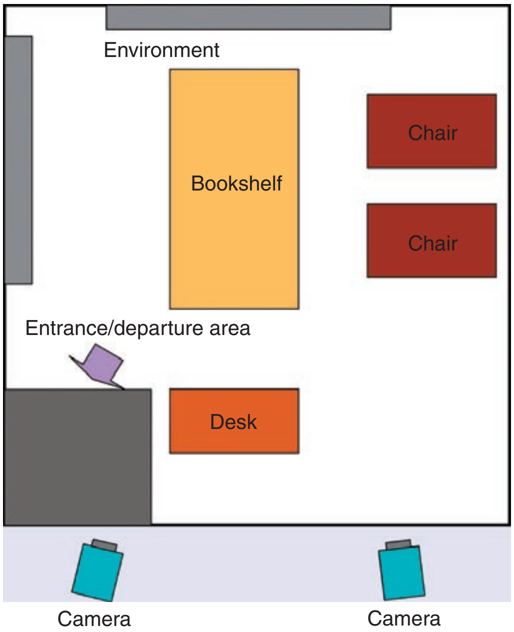


Fig. 7. Experimental environment.

floor is low, so we define penalty function $E\,(S)$ based on the position in the environment as

$$E(S) = \prod_{i=1}^{K} H(M_i), \tag{6}$$

where $H(M_i)$ is 1 if the position of state M_i in the 3D environmental model is on the floor (not on the objects); otherwise, $H(M_i)$ is 0.

b) Penalty based on relative distance among targets: Since multiple humans cannot occupy the same position, we define penalty function $R\,(S)$ based on the relative distance among targets as

$$\begin{aligned} R(S) &= \prod_{i,j} \psi(M_i, M_j) \\ \psi(M_i, M_j) &= 1 - \exp^{-\lambda D_{i,\,j}} \end{aligned} \tag{7}$$

where $D_{i,\,j} = \sqrt{(x_i - x_i)^2 + (y_i - y_i)^2}$ is the distance between targets i and j and λ is a threshold parameter.

Finally, likelihood L is computed by

$$L(S) = E(S) \times R(S) \times V(S). \tag{8}$$

This likelihood provides efficient estimation by restricting the human movement to just the floor and preventing target conflict in 3D space.

4) *Acceptance ratio*: We decide whether to accept or reject the state by using acceptance ratio a, which is given by

$$a = min(1, \frac{L_{new}}{L_{old}}), \tag{9}$$

where L_{old} is the likelihood of the previously accepted state and L_{new} is the likelihood of the state being considered. If $a \geq 1$, we accept the new state; otherwise, we accept the new state with probability a. If we reject the new state, we keep the current state.

5) *MAP estimation*: After repeating sampling $B + P$ times, we compute state S_t by

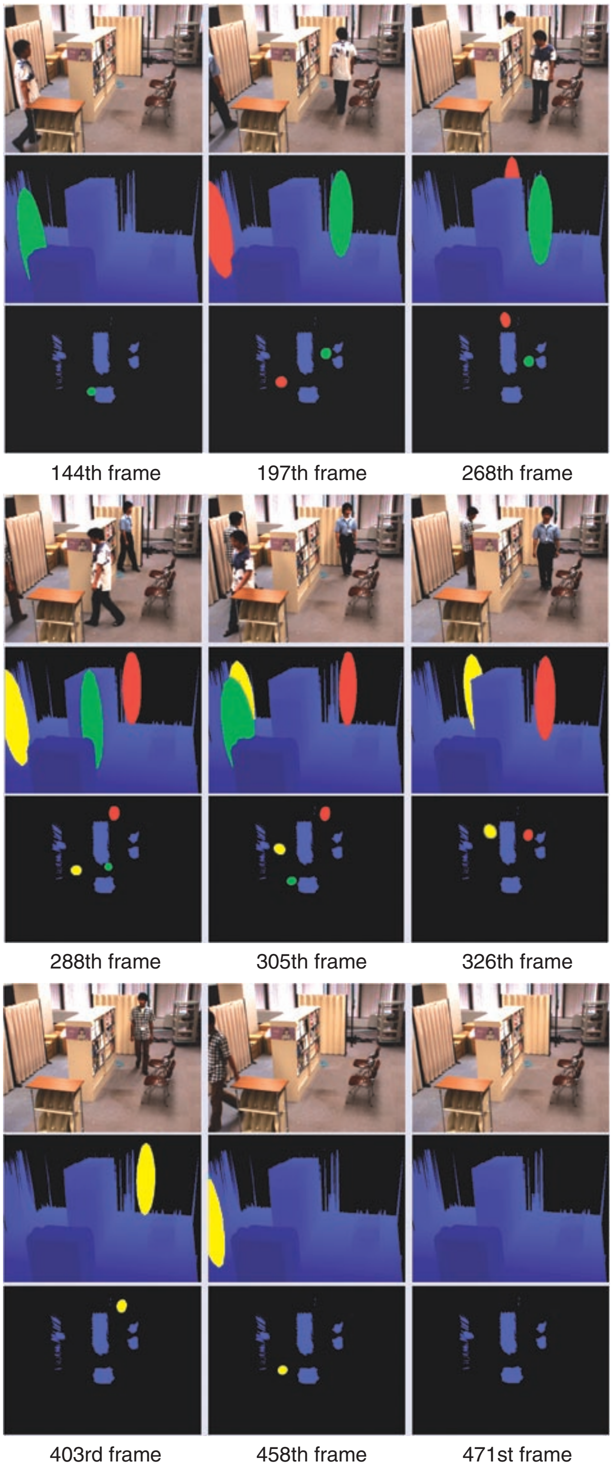
$$S_t = \frac{1}{P} \sum_{i=B}^{B+P} \hat{S}_{t,\,i}. \tag{10}$$

We use only the last P samples to compute state S_t because the first B samples are assumed to vary widely and include different target configurations.

4. Experiment

4.1 System and conditions

Our system consisted of a personal computer (CPU: AMD Athlon 64 × 2 4800+) and two color CCD cameras (FLEA made by Point Grey Research). Each captured image had a resolution of 640 × 480. The intrinsic and extrinsic camera parameters were estimated in advance. In this experiment, the number of iterations $B + P$ was set to 300. For MAP estimation, we use the last P=100 samples. The system ran at 5 frames per second in this non-optimized implementation. The images from the cameras are shown in **Fig. 6** and a bird's eye view of the experimental environment is shown in **Fig. 7**. We defined the shaded region in Fig. 7 as the entrance/departure area.



144th frame 197th frame 268th frame

288th frame 305th frame 326th frame

403rd frame 458th frame 471st frame

Fig. 8. Estimation results (images) of sequence #1. Top row: real camera image. Middle row: estimated virtual camera image. Bottom row: bird's eye view of estimated virtual camera image.

4.2 Multiple human tracking

To evaluate the basic tracking performance of our system, we used an image sequence in which three humans entered and left the monitored area at different times (sequence #1).

The images selected from the monitoring results of sequence #1 are shown in **Fig. 8**. Our system could correctly capture the movements of the three humans. In the 268th frame, most of one subject's body was occluded by a shelf, and in the 305th frame, two humans completely overlapped. Even under these severe occlusions, the tracking error was not significant as a result of our use of the entrance/departure area constraint.

The trajectories on the X-Y plane overlaid by reconstructed 3D points are shown in **Fig. 9**. The continuation of trajectories even in the case of severe occlusions caused by fixed objects in the environment and mutual occlusions demonstrates the robustness of our system to occlusions.

4.3 Evaluation of the tracked position

For a rough evaluation of the position tracked by the system, we used an image sequence in which a subject walked around a prearranged route (sequence #2). We compared the estimated motion trajectory with the actual trajectory on the ground (ground truth trajectory). The estimated motion trajectory and the ground truth trajectory on the X-Y plane overlaid by 3D points are shown in **Fig. 10**. The estimated and ground truth trajectories are very close, so this result confirms that our system offers high accuracy. The mean and maximum errors of the estimated distance were 4.86 and 29.43 cm, respectively.

5. Conclusions and future work

In this article, we introduced a 3D human tracking system that can track variable interacting targets in the presence of severe occlusions caused by both fixed objects in the environment and target movement. The next step is to extend the system to cope with crowded scenes, which we expect to be fairly difficult. Evaluations of such systems should lead to a better system design in terms of factors such as camera locations.

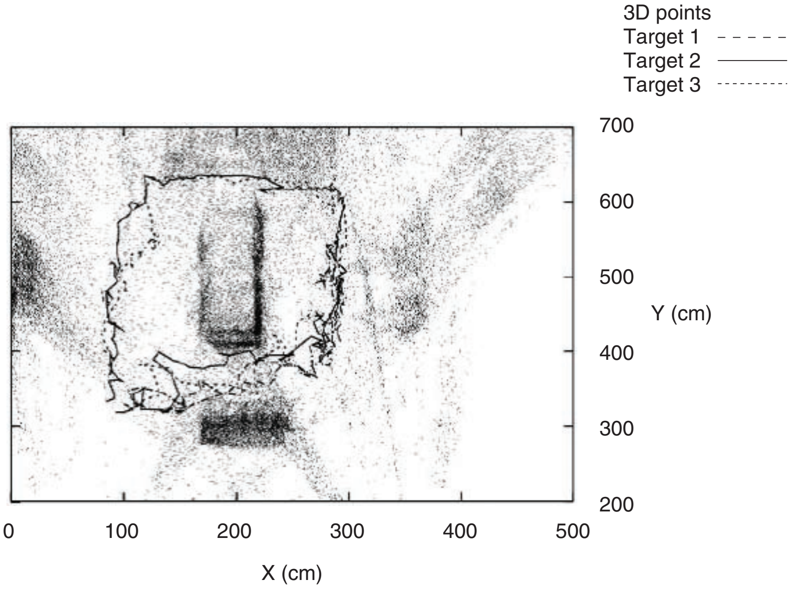


Fig. 9. Motion trajectories on X-Y plane of sequence #1 overlaid by 3D points.

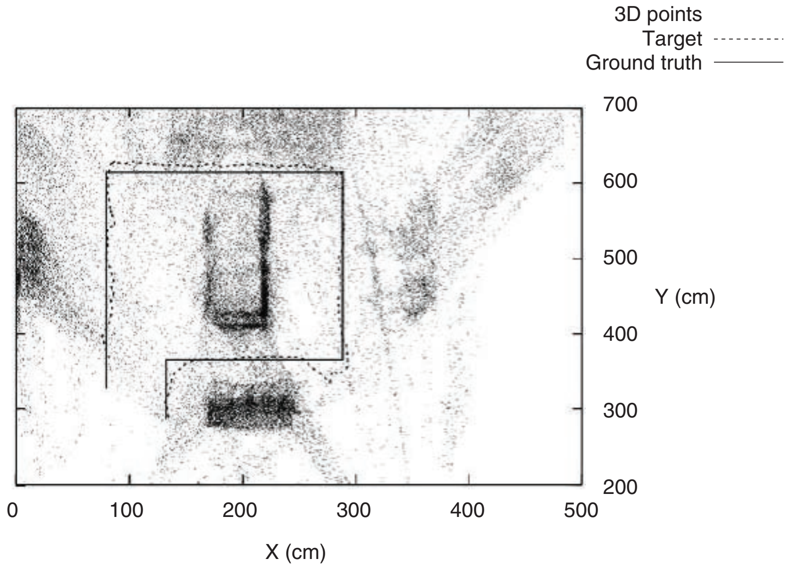


Fig. 10. Motion trajectory on the X-Y plane of sequence #2 and ground truth trajectory overlaid by 3D points.

References

[1] B. Wu and R. Nevatia, "Tracking of multiple, partially occluded humans based on static body part detection," Proc of CVPR'06, New York, Vol. 1, pp. 951–958, 2006.

[2] T. Yang, S. Z. Li, Q. Pan, and J. Li, "Real-time multiple objects tracking with occlusion handling in dynamic scenes," Proc. of CVPR'05, San Diego, Vol. 1, pp. 970–975, 2005.

[3] Z. Khan, T. Balch, and F. Dellaert, "MCMC-based particle filtering for tracking a variable number of interacting targets," IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 27, No. 11, pp. 1805–1819, 2005.

[4] K. Smith, D. Gatica-Perez, and J.-M. Odobez, "Using particles to track varying numbers of interacting people," Proc. of CVPR'05, San Diego, Vol. 1, pp. 962–969, 2005.

[5] Z. Tao and R. Nevatia, "Tracking multiple humans in crowded environment," Proc. of CVPR'04, Washington, DC, Vol. 2, pp. 406–413, 2004.

[6] T. Darrell, D. Demirdjian, N. Checka, and P. Felzenszwalb, "Plan-view trajectory estimation with dense stereo background models," Proc. of ICCV'01, Vancouver, Vol. 2, pp. 628–635, 2001.

[7] J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. Hale, and S. Shafer, "Multi-camera multi-person tracking for Easyliving," Proc. of 3rd IEEE International Workshop on Visual Surveillance, 2000, Dublin, pp. 3–10, 2000.

[8] A. Mittal and L. S. Davis, "M2Tracker: A Multi-view Approach to

Segmenting and Tracking People in a Cluttered Scene Using Region-Based Stereo," Proc. of ECCV'02, Copenhagen, Vol. 1, pp. 18–36, 2002.

[9] K. Otsuka and N. Mukawa, "Multiview occlusion analysis for tracking densely populated objects based on 2-D visual angles," Proc. of CVPR'04, Washington, DC, Vol. 1, pp. 90–97, 2004.

[10] T. Osawa, X. Wu, K. Wakabayashi, and T. Yasuno, "Human Tracking by Particle Filtering Using Full 3D Model of Both Target and Environment," Proc. of ICPR'06, Hong Kong, Vol. 2, pp. 25–28, 2006.

[11] P. J. Green, "Trans-dimensional Markov chain Monte Carlo, Highly Structured Stochastic Systems," Oxford Univ. Press, 2003.

[12] T. Osawa, I. Miyagawa, K. Wakabayashi, K. Arakawa, and T. Yasuno, "3D Reconstruction from an Uncalibrated Long Image Sequence," Proc. of IVCNZ'06, Great Barrier Island, pp. 473–478, 2006.

[13] K. Okada, S. Kagami, M. Inaba, and H. Inoue, "Plane segment finder: Algorithm implementation and applications," Proc. of ICRA'01, Robotics and Automation, Seoul, Vol. 2, pp. 2120–2125, 2001.

Tatsuya Osawa

Research Engineer, Visual Media Communications Project, NTT Cyber Space Laboratories.

He received the B.S. degree in physics and the M.E. degree in energy science from Tokyo Institute of Technology, Tokyo, in 2002 and 2004, respectively. Since joining NTT Laboratories in 2004, he has been engaged in research on computer vision. The current focus of his research is on human behavior recognition with distributed cameras. He has received several awards, including the Best Paper Award in ICPR from the International Association for Pattern Recognition, the Funai Best Paper Award from the Funai Foundation for Information Technology, and the Best Paper Award in IVCNZ from the National Group for Image and Vision Computing in New Zealand. He is currently pursuing a Ph.D. degree at Tokyo Institute of Technology. He is a member of IEEE and the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan.

Xiaojun Wu

Research Engineer, Visual Media Communications Project, NTT Cyber Space Laboratories.

He received the B.S. degree in electrical and electronic engineering, the M.S. degree in informatics, and the Ph.D. degree in informatics from Kyoto University, Kyoto, in 1998, 2000, and 2005, respectively. Since joining NTT Laboratories in 2005, he has been engaged in research on computer vision, focusing on 3D shape reconstruction of the human body using multiviewpoint cameras. He is a member of IEICE.

Kaoru Wakabayashi

Senior Research Engineer, Visual Media Communications Project, NTT Cyber Space Laboratories.

He received the B.E. degree in electro-communications from the University of Electro-Communications, Tokyo, in 1982 and the Ph.D. degree in electronic engineering from the University of Tokyo, Tokyo, in 1999. Since joining Nippon Telegraph and Telephone Public Corporation (now NTT) in 1982, he has been engaged in research on facsimile communications networks, binary image processing, map information processing, cognitive mapping and understanding, and visual monitoring systems. He received the 1993 NTT President's Award, the 1998 AM/FM International Japan Best Speaker Award, the 2006 ICPR Best Paper Award, the 2006 Funai Best Paper Award, and the 2006 IVCNZ Best Paper Award. He is a member of IEICE and the Information Processing Society of Japan.

Hideki Koike

Senior Research Engineer, Supervisor, Group Leader, Visual Media Communications Project, NTT Cyber Space Laboratories.

He received the M.S. degree in mathematics from Tohoku University, Miyagi, in 1985. He joined NTT Labs. in 1985 and engaged in research on image processing. He was transferred to NTT COMWARE in 2001 and engaged in research on RFID. He moved to NTT Cyber Space Labs. in 2007 and is engaged in research on computer vision. He is a member of IEICE.