Systems and Computers in Japan, Vol. 38, No. 2, 2007
Translated from Denshi Joho Tsushin Gakkai Ronbunshi, Vol. J88-D-II, No. 8, August 2005, pp. 1592–1600

Incorporating Environment Models for Improving Vision-Based Tracking of People

Tatsuya Suzuki,[1] Shinsuke Iwasaki,[1] Yoshinori Kobayashi,[1] Yoichi Sato,[1] and Akihiro Sugimoto[2]

[1]Institute of Industrial Science, University of Tokyo, Tokyo, 153-8505 Japan

[2]National Institute of Informatics, Tokyo, 101-8430 Japan

SUMMARY

This paper presents a method for real-time 3D human tracking based on the particle filter by incorporating environment models. We track a human head represented with its 3D position and orientation by integrating the multiple cues from a set of distributed sensors. In particular, the multi-viewpoint color and depth images obtained from distributed stereo camera systems and the 3D shape of an indoor environment measured with a range sensor are used as the cues for 3D human head tracking. The 3D shape of an indoor environment allows us to assume the existing probability of a human head (we call this probability the environment model). While tracking the human head, we consider the environment model to improve the robustness of tracking in addition to the multi-camera's color and depth images. These cues including the environment model are used in the hypothesis evaluation and integrated naturally into the particle filter framework. The effectiveness of our proposed method is verified through experiments in a real environment. © 2007 Wiley Periodicals, Inc. Syst Comp Jpn, 38(2): 71–80, 2007; Published online in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/scj.20612

Contract grant sponsor: Ministry of Education, Science and Technology research grant for the Special Study "Realization of Flexible Human-Machine Interaction Based on Understanding of Human Intentions and Behavior" (No. 13224051).

Key words: human tracking; particle filter; multiple cues; environment model.

1. Introduction

Tracking people by using camera footage is one of the most important tasks in computer vision. The reason is that detecting people and recognizing their motions are tremendously important issues in many application systems especially in surveillance systems.

To date, many human tracking systems have been proposed. In the case of object tracking by using an unreliable observation, time-series filtering is known to be effective. Time-series filtering is a method which estimates the value based on motion prediction and sensor observation. Among these, in the last decade, the particle filter framework [3, 4] has been proposed and reported to be effective. This allows us to realize robust human tracking in a complex background and a situation in which the observed values are non-Gaussian distributed.

The particle filter framework is also known as a Bayesian filter, Condensation or sequential Monte Carlo. In the particle filter framework the tracking target is represented as a discrete probability density by using a finite set of hypotheses with state variables and likelihoods. Tracking is realized by propagating the representation by means of a stochastic model. This allows for robust tracking against observation noise and abrupt changes in the target's motion.

On the other hand, in order to improve performance in such a tracking framework, it is necessary to improve the accuracy of human head detection by using sensor observation, for which various methods have been proposed.

For the evaluation of the human head using a single fixed camera, a method has been proposed in which the head is assumed to be an ellipse and a color histogram of the head is used [7]. It is also proposed that the contour similarity indicates the reliability of the human head [1, 13]. Human detection methods using distance images, which are less affected by changes of illumination, have been proposed [2, 8]. Methods of improving the accuracy of observation by integrating these multiple observation cues have also been proposed [5]. Methods which observe an object from multiple viewpoints to reduce the unobservable regions and to realize robust tracking against occlusion are proposed. Some of these methods are based on visual volume intersection [6, 11]. Some methods propose to merge multiple cues from multi-viewpoint image on a 2D plane [12, 14].

However, these relatively simple methods are not enough to be effective for tracking a human head in 3D such as an indoor environment. In order to achieve 3D tracking using multiple cameras, it is necessary to evaluate the likelihood that an object is a human head, including cases in which the head is not always directed toward the camera. Furthermore, illumination changes greatly in real environments. In order to achieve stable tracking even against a complex background in which the colors of the skin or hair cannot be discriminated from those of the floor or walls, it will be effective to use environmental cues concerning the 3D shape of an indoor environment.

Consequently, this paper proposes a method of real-time 3D human head tracking by using stereo-vision cameras and a range sensor. The method integrates multi-viewpoint images and utilizes the 3D shape of an indoor environment obtained from the range sensor, so that the distribution of the existing probability of the head in the indoor space is considered.

Color and distance information are obtained from stereo-vision cameras (Fig. 1). Each pixel of a stereo-vision camera image has color information on the RGB values (represented as a 3D vector) of the point in the 3D space which is projected to that pixel. Each pixel also has distance information consisting of the 3D coordinates in the camera coordinate system of the point in the 3D space which is projected to that pixel. The 3D shape of an indoor environment such as wall and desk is obtained from the range sensor (Fig. 2).

Fig. 1. Color and depth images captured from a stereo-vision camera system.

Fig. 2. Three-dimensional shape of an environment measured by using a range finder.

In particular, the proposed method is as follows. The tracking target is the head, modeled as an ellipsoid with its orientation. A particle filter is used as the time-series filter for tracking. For observation, stereo-vision cameras are placed at the four corners of an indoor ceiling, directed toward the center of the room (Fig. 3). Color and distance information are acquired, and the likelihood of a human head is evaluated from each of these. By integrating the multiple cues obtained from the multi-viewpoint images, the position of the human head in the 3D space, and also the orientation of the head, is evaluated. Then, using the 3D shape of an indoor environment obtained from the range sensor, an environment model which considers the existing probability of a human head in the indoor space with walls

Fig. 3. The placement of the sensors used in our system.

and desks, is incorporated in the evaluation. Stable tracking is accomplished by these procedures. Experiments in a real environment were performed to track humans by the above procedure, and the effectiveness of the proposed method was demonstrated.

2. Particle Filter

The particle filter described in Ref. 3 is used in this study. The algorithm of the particle filter is briefly described below.

2.1. Time-series filtering

Let $\mathbf{x}_t$ be the state variable of the tracking target at time t, and $\mathbf{z}_t$ be the observation obtained from the image. Let the observation obtained up to time t be $\mathbf{Z}_t = (\mathbf{z}_1, \ldots, \mathbf{z}_t)$. The problem is to estimate the probability density function $P(\mathbf{x}_t|\mathbf{Z}_t)$ of state $\mathbf{x}_t$ when observations up to time t have been obtained.

When the probability density function $P(\mathbf{x}_{t-1}|\mathbf{Z}_{t-1})$ of the tracking target at time $t-1$ and the motion model $P(\mathbf{x}_t|\mathbf{x}_{t-1})$ from time $t-1$ to t are given, the a priori probability $P(\mathbf{x}_t|\mathbf{Z}_{t-1})$ at time t is expressed as follows, assuming a Markov process:

$$P(\mathbf{x}_t \mid \mathbf{Z}_{t-1}) = \int P(\mathbf{x}_t \mid \mathbf{x}_{t-1}) P(\mathbf{x}_{t-1} \mid \mathbf{Z}_{t-1}) d\mathbf{x}_{t-1} \quad (1)$$

When the likelihood $P(\mathbf{z}_t|\mathbf{x}_t)$ at time t is estimated from the image, the probability density function $P(\mathbf{x}_t|\mathbf{Z}_t)$ at time t is expressed as follows, according to the Bayes theorem:

$$P(\mathbf{x}_t \mid \mathbf{Z}_t) \propto P(\mathbf{z}_t \mid \mathbf{x}_t) P(\mathbf{x}_t \mid \mathbf{Z}_{t-1}) \quad (2)$$

2.2. Weighted sampling

In the particle filter, the probability density function $P(\mathbf{x}_t|\mathbf{Z}_t)$ at time t is represented in discrete form, using the finite set of hypotheses $\{\mathbf{s}_t^{(1)}, \ldots, \mathbf{s}_t^{(N)}\}$ for state $\mathbf{x}_t$ and the weights of the hypotheses $\{\pi_t^{(1)}, \ldots, \pi_t^{(N)}\}$. Let $\mathbf{s}_t^{(n)}$ be the state variable for the n-th hypothesis at time t. The weight is evaluated by $\pi_t^{(n)} = P(\mathbf{z}_t|\mathbf{x}_t = \mathbf{s}_t^{(n)})$.

2.3. Tracking by particle filter

The proceeding applied to the set of hypotheses consists of the following three parts. Tracking is accomplished by repeating this process.

(1) Let the distribution $P(\mathbf{x}_{t-1}|\mathbf{Z}_{t-1})$ of the state variable $\mathbf{x}_{t-1}$, when observation $\mathbf{Z}_{t-1}$ is obtained at time $t-1$, be represented by a set of N weighted hypotheses $\{(\mathbf{s}_{t-1}^{(n)}, \pi_{t-1}^{(n)}), n = 1, \ldots, N\}$. The set of hypotheses $\{\mathbf{s}'^{(1)}_{t-1}, \ldots, \mathbf{s}'^{(N)}_{t-1}\}$ is selected according to the proportion of the weights $\{\pi_{t-1}^{(1)}, \ldots, \pi_{t-1}^{(N)}\}$ for the hypotheses.

(2) The selected set of hypotheses is propagated in accordance with the motion model $P(\mathbf{x}_t|\mathbf{x}_{t-1} = \mathbf{s}'^{(n)}_{t-1})$ to generate the set of N hypotheses $\mathbf{s}_t^{(n)}$ at time t corresponding to $P(\mathbf{x}_t|\mathbf{Z}_{t-1})$.

(3) By estimating the weight $\pi_t^{(n)}$ from the image, the weight $\pi_t^{(n)} = P(\mathbf{z}_t|\mathbf{x}_t = \mathbf{s}_t^{(n)})$ for the new sample $\mathbf{s}_t^{(n)}$ is derived. The weight $\pi_t^{(n)}$ is normalized so that $\Sigma_{n=1}^{N} \pi_t^{(n)} = 1$. As a result, $\{(\mathbf{s}_t^{(n)}, \pi_t^{(n)}), n = 1, \ldots, N\}$ is obtained, which is an approximate representation of $P(\mathbf{x}_t|\mathbf{Z}_t)$ at time t. The expectation of the set of hypotheses is used as the estimation value of the state variable for the tracking target.

3. Estimation of Position and Orientation of Human Head

3.1. Model of human head

An ellipsoid is assumed as the model for the human head. The 3D world coordinates XYZ are defined in the indoor space. The coordinate system is represented with their X and Y axes aligned on the floor and the Z axis normal to the floor. The shape of the human head is assumed to remain invariant, and the position is represented by the center coordinates (x, y, z) of the ellipsoid. Assuming that a human does not tilt his/her head, the orientation is represented by using only the angle of rotation θ around the Z axis, with the X axis as the reference. Figure 4 shows the situation.

Thus, the head is represented by the state variable $\mathbf{s} = (x, y, z, \theta)$. The n-th hypothesis at time t for the state variables of the human head is represented as $\mathbf{s}_t^{(n)} = (x_t^{(n)}, y_t^{(n)}, z_t^{(n)}, \theta_t^{(n)})$.

Let the ellipsoid $\Gamma_t^{(n)}$ representing the n-th hypothesis at time t be projected onto the image of the i-th camera, and

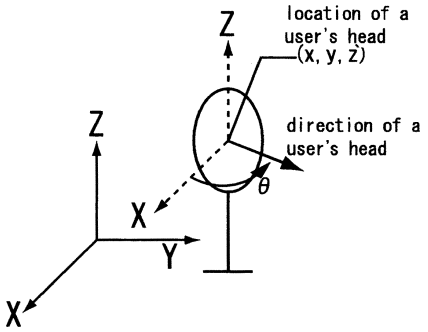


Fig. 4. The model of a user's head.

let the obtained region be $\Omega_{i,t}^{(n)}$. Letting the projection function be F_i, the region is expressed as follows:

$$\Omega_{i,t}^{(n)} = F_i(\Gamma_t^{(n)}) \tag{3}$$

Since the head is modeled as an ellipsoid, $\Omega_{i,t}^{(n)}$ is an ellipse. The total number of pixels in region $\Omega_{i,t}^{(n)}$ is written as $|\Omega_{i,t}^{(n)}|$.

3.2. Evaluation of likelihood as human head based on color information

The weight $\pi_{i,t}^{color,(n)}$ based on the color information of hypothesis $\mathbf{s}_t^{(n)}$ is determined from the color image of the i-th camera. In this paper, weight means the likelihood of being a human head.

In order to reduce the effects of the background and individual differences of hair color, only the skin color is used in the evaluation. The HSV color space is used as the color space, because many studies indicate that this color space is one of the most suitable spaces for extracting the skin color region [9, 10]. In this study, however, V is ignored in order to reduce the effect of illumination changes, and the 2D space composed of H and S is used in the identification of the skin color region. Based on learned data on human skin color, the skin color region is defined beforehand. If the chrominance of a pixel in HS color space is included in the skin color region, the pixel is judged to be skin-colored.

The weight is evaluated by the similarity between the ratio of the skin color pixels in the ellipse $\Omega_{i,t}^{(n)}$ and the ratio of the skin color pixels in the human head, which is defined beforehand. In the i-th color image obtained by observation, let the number of pixels contained in the skin region inside the ellipse $\Omega_{i,t}^{(n)}$ be $c_{i,t}^{(n)}$, and its ratio be $\bar{c}_{i,t}^{(n)}$. We use

$$\bar{c}_{i,t}^{(n)} = \frac{1}{|\Omega_{i,t}^{(n)}|} c_{i,t}^{(n)} \tag{4}$$

It should also be noted that the skin color region is large when the head is oriented toward the center of the camera, and is small when it is oriented away from the camera. To deal with this problem, the quantity $\hat{c}_{i,t}^{(n)}$, which is a function of state variable θ, taking its maximum value when the human is oriented in the frontal direction and its minimum value when the human is oriented in the reverse direction, is provided beforehand. The weight $\pi_{i,t}^{color,(n)}$ is set higher when the difference is smaller, and is given by the following evaluation function:

$$\pi_{i,t}^{color,(n)} = a_i^{color} - b_i^{color} \left| \bar{c}_{i,t}^{(n)} - \hat{c}_{i,t}^{(n)} \right| \tag{5}$$

where a_i^{color} (> 0) and b_i^{color} (> 0) are constants. In the experiment, these values are defined empirically. The weights are set higher when the similarity between the prediction by the hypothesis and the observation is high.

3.3. Evaluation of likelihood as human head based on distance information

The weight $\pi_{i,t}^{depth,(n)}$ for the hypothesis $\mathbf{s}_t^{(n)}$ based on the distance information is derived from the distance image of the i-th camera. For the distance image obtained by observation from the i-th camera, let the camera coordinates of pixel p inside the ellipse $\Omega_{i,t}^{(n)}$ be $\tilde{\mathbf{w}}_{i,t}^{(n)}(p)$. Let the center $\mathbf{v}_t^{(n)} = (x_t^{(n)}, y_t^{(n)}, z_t^{(n)})$ of the ellipsoid be projected onto the camera coordinate system of the i-th camera, and let the coordinates of the result be $\tilde{\mathbf{v}}_{i,t}^{(n)}$. Let the distance between $\tilde{\mathbf{w}}_{i,t}^{(n)}(p)$ and $\tilde{\mathbf{v}}_{i,t}^{(n)}$ be $d_{i,t}^{(n)}(p)$. Then, the following relation applies:

$$d_{i,t}^{(n)}(p) = \left| \tilde{\mathbf{w}}_{i,t}^{(n)}(p) - \tilde{\mathbf{v}}_{i,t}^{(n)} \right| \tag{6}$$

The distance of the point on the human head ellipsoid corresponding to pixel p from the center of the ellipsoid is known, and is written as $\hat{d}_{i,t}^{(n)}(p)$. Then the weight $\pi_{i,t}^{depth,(n)}$ is set higher as the sum of the differences for these pixels becomes, and is given by the following evaluation function:

$$\pi_{i,t}^{depth,(n)} = a_i^{depth} - b_i^{depth} \left(\frac{1}{|\Omega_{i,t}^{(n)}|} \sum_{p \in \Omega_{i,t}^{(n)}} \left| d_{i,t}^{(n)}(p) - \hat{d}_{i,t}^{(n)}(p) \right| \right) \tag{7}$$

where a_i^{depth} (> 0) and b_i^{depth} (> 0) are constants. In the experiment, as in the case of the color information, these values are empirically set, and the weight is set higher with increasing similarity between the prediction by the hypothesis and the observation.

3.4. Introduction of environment model

In this study, the environment model is further introduced to stabilize the tracking process. The system acquires the 3D shape of an indoor environment from the range sensor in advance, and evaluates the existing probability of a human head in the indoor space from the arrangements of the walls and desks. Such a representation of the existing probability of a human head in the space is called the environment model.

Based on the environment model, when evaluating the set of hypotheses in the particle filter framework, it is possible to suppress the weights of the hypothesis for regions in which a human head cannot exist, such as inside objects in the background, or in regions where a human head is unlikely to exist, such as the top of a desk or shelf.

As a result, the generation of hypotheses for regions with a low existing probability of a human head is suppressed, and more stable tracking is expected.

A model composed of the following three kinds of regions is considered as the environment model.

- Region A: outside the walls and inside still objects, such as desks and shelves, where a human head cannot exist.
- Region B: vertical direction of still objects, such as desks and walls, where the existing probability of a human head is lower than in other regions.
- Region C: low regions and regions higher than the human height, where the likelihood of existence of a human head is lower than in other regions.

Figure 5 shows an example of such an environment model. Based on the above classification of regions, the weight $e_t^{(n)}$ is determined on the basis of the existing probability of a human head for the state variable $\mathbf{s}_t^{(n)}$. Let the existing possibilities for the human head in the regions be $e_t^{A,(n)}$, $e_t^{B,(n)}$, and $e_t^{C,(n)}$, respectively. Let the existing possibility of the human head for the region not belonging to any of these regions be α. When regions overlap, the minimum value is selected. In other words, let

$$e_t^{(n)} = \min(e_t^{A,(n)}, e_t^{B,(n)}, e_t^{C,(n)}, \alpha) \qquad (8)$$

In the experiment, we set $e_t^{A,(n)} = 0$, $e_t^{B,(n)} = 0.5$, $e_t^{C,(n)} = 0.2$, and $\alpha = 1.0$.

3.5. Integration of multiple cues

Color and distance information from multiple cameras, and also the environment model, are integrated. Based on color information, it is possible to judge the orientation of the head. However, since the human head has a symmetrical structure, it is impossible for a single camera to distinguish between states with left–right symmetry. The tracking target and the background can be discriminated by using distance information. However, the evaluation based on distance information can be high not only when the hypothesis indicates the position of the human head, but also when the hypothesis indicates a region close to the wall. By further introducing the environment model, higher weights can be assigned to hypotheses with higher existing probability of the human head in the indoor space. Thus, in order to realize stable 3D tracking, it is necessary to integrate these three items of information so that the hypothesis indicating the correct position is finally given a high evaluation value.

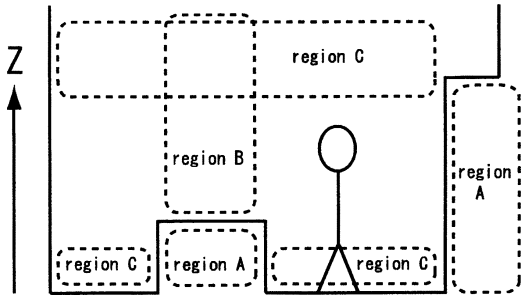


Fig. 5. The model of an environment.

The integration procedure is simplified so as not to degrade real-time operation. For hypothesis $\mathbf{s}_t^{(n)}$, the product of the weight $\pi_{i,t}^{color,(n)}$ based on the color information from the camera, the weight $\pi_{i,t}^{depth,(n)}$ based on the distance information, and the existing probability $e_t^{(n)}$ of the human head defined in the environment model, is calculated:

$$\pi_t^{(n)} = e_t^{(n)} \prod_i \pi_{i,t}^{color,(n)} \prod_i \pi_{i,t}^{depth,(n)} \qquad (9)$$

Thus, the weight for the hypothesis is obtained with allowance for the environment model.

4. Experimental Results

4.1. Overall processing flow

The system was composed of one server PC and four client PCs (CPU Intel Pentium4 2.0 GHz, memory 1.0 Gbyte). These PCs were connected through Gigabit Ethernet with a 1 Gbps rate. The stereo-vision camera was a Point Grey Co. Digiclops. The range sensor was a model LMS200 manufactured by SICK. The sensors were calibrated beforehand. The results of observation by the sensors could be fused in the preset world coordinate system. A Digiclops camera was attached to each client PC so that the color information and the distance information could be acquired. The environment model was constructed beforehand on the basis of the indoor configuration obtained from the range sensor.

The overall processing flow is as follows.

(1) The server generates a set of hypotheses $\mathbf{s}_t^{(n)} = (x_t^{(n)}, y_t^{(n)}, z_t^{(n)}, \theta_t^{(n)})$.

(2) The server sends the set of hypotheses $\mathbf{s}_t^{(n)}$.

(3) Client i receives the set of hypotheses $\mathbf{s}_t^{(n)}$.

(4) Client i acquires the color information and the distance information.

(5) Client i evaluates the weights for the set of hypotheses $\mathbf{s}_t^{(n)}$. It evaluates the weight $\pi_{i,t}^{color,(n)}$ for $\mathbf{s}_t^{(n)}$ based on the color information, and evaluates the weight $\pi_{i,t}^{depth,(n)}$ for $\mathbf{s}_t^{(n)}$ based on the distance information.

(6) Client i sends the weights $(\pi_{i,t}^{color,(n)}, \pi_{i,t}^{depth,(n)})$.

(7) The server receives the weights $(\pi_{i,t}^{color,(n)}, \pi_{i,t}^{depth,(n)})$.

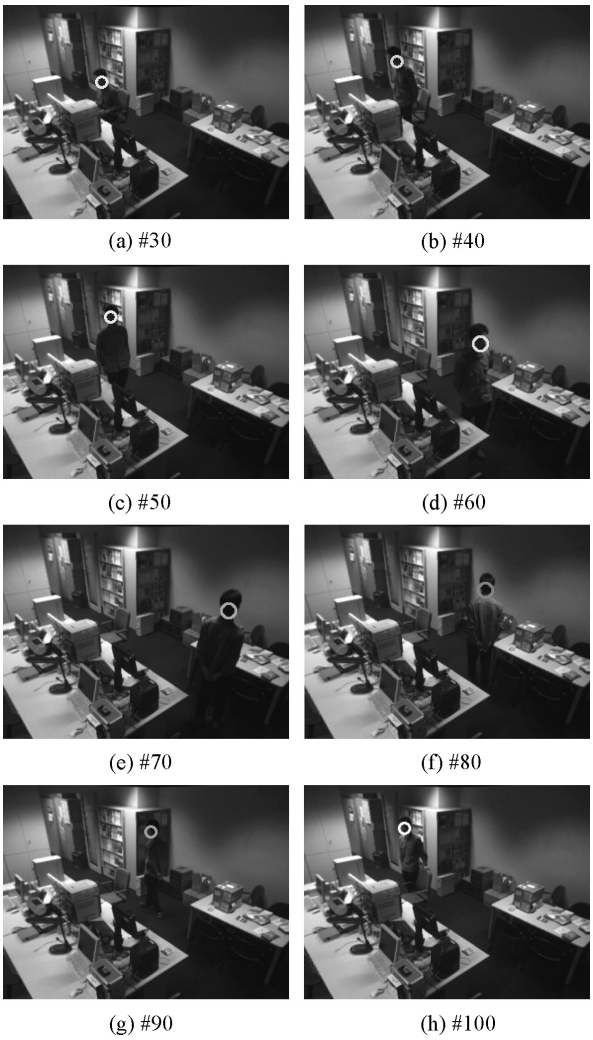
(8) The server integrates the environment model $\pi_t^{e,(n)}$, and evaluates the weight $\pi_t^{(n)}$.

(9) The server acquires $(\mathbf{s}_t^{(n)}, \pi_t^{(n)})$. It generates a set of hypotheses $\mathbf{s}_{t+1}^{(n)}$, and the procedure returns to step (1). It estimates the state variable by expectation.

In tracking, it is assumed that the initial value of the state variable for the human head is known. The number of hypotheses is set as $N = 512$, and head tracking is started. The frame rate in tracking is approximately 4 fps.

4.2. Result of tracking

Based on the above preparations, the method proposed in this paper was used in a human head tracking experiment in a real environment. The tracking experiment was performed in the following sequence.



(a) #30 (b) #40 (c) #50 (d) #60 (e) #70 (f) #80 (g) #90 (h) #100

Fig. 6. Tracking results 1.

The subject started from sitting posture (#30), stood up (#40), changed orientation (#50), walked straight (#60), changed orientation (#70), gradually changed orientation while walking (#80), turned back to walk toward the original position (#90), and stopped (#100). Figure 6 shows an example of the images obtained in the experiment.

The result of estimation based on the expectation of the state variable for the human head after information integration is shown by a circle superimposed on the color image, with the brightness increasing as the subject is oriented more toward the camera. The weight of the hypothesis is evaluated by using the color information, the distance information, and the environment model.

We can see from Fig. 6 that the center of the human head is well estimated on the basis of the expectations for the set of hypotheses. This is attributed to the fact that hypotheses with higher weights after information integration are generally concentrated to the neighborhood of the human head. It is also seen in the evaluation based on the color information that the color model for the human head based on the camera orientation is effective, and that estimation of the position and orientation of the human head works effectively in the integration of the results observed from multiple cameras.

In order to investigate quantitatively the tracking accuracy by the proposed method, the human head in the image was manually specified. The 3D coordinates obtained by inverse projection from multiple images was assumed to be the true position, and was compared to the result of estimation. Figure 7 shows the result of estimation, and also the corresponding trajectories of the true position of the human head in the 3D space and on the *XY* plane. Table 1 shows the mean and standard deviation of the error on the *Z* axis and on the *XY* plane.

The detection error on the *Z* axis is approximately 6 cm, deviating downward. The reason seems to be that the evaluation with color information includes the neck, and the evaluation with the distance information includes their clothes. The detection error on the *XY* plane, on the other hand, is approximately 4 cm, and the tracking can be considered sufficiently accurate.

Next we investigated the robustness of the method proposed in this paper. Using entirely the same system, the head tracking experiment was performed for a subject different from the subject in the previous experiment. The subject walked around a table in the room, changing the direction of the head. Figure 8 shows a selection from among 60 frames obtained in the same way as in the previous experiment.

Figure 9 shows the result of estimation, and also the trajectory of the corresponding true position of the human head on the *XY* plane. Table 2 shows the mean and standard

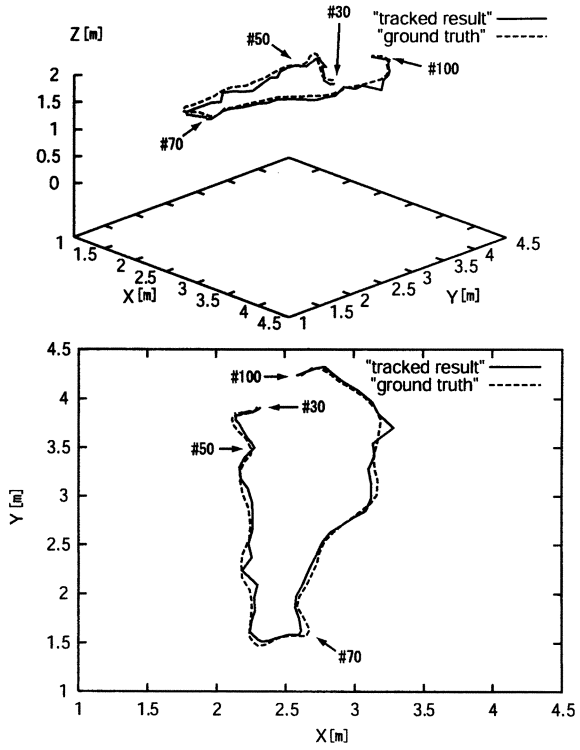


Fig. 7. Trajectory of a user's head position 1.

Fig. 8. Tracking results 2.

deviation of the error on the Z axis and on the XY plane. The detection errors on the Z axis and XY plane are approximately 2 cm and 3 cm, respectively. It is thus seen that accurate tracking was obtained, as in the previous experiment.

We also evaluated the variation of the estimated position due to the approximation of the probability density by the set of hypotheses in the particle filter. The detection algorithm was applied 30 times to an input image sequence stored in advance. The standard deviation of the estimated position in the frames was calculated and was averaged for all frames. Table 3 shows the results. The value for the frame having the greatest standard deviation is also shown.

In the frame (#37) with the largest variation on the XY plane, the standard deviation was approximately 0.7 cm, which is sufficiently small compared to the detection error. Thus, the position of the human was estimated stably by using a sufficient number of hypotheses.

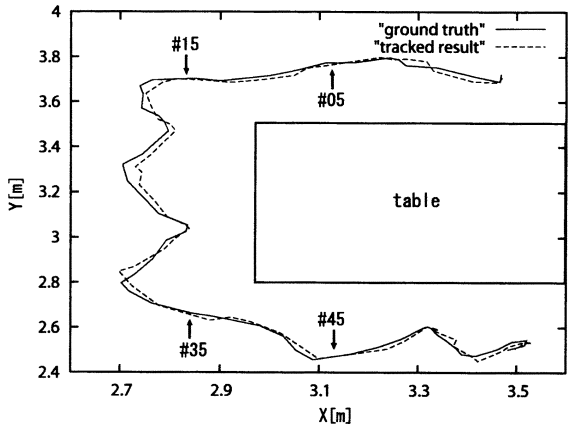


Fig. 9. Trajectory of a user's head position 2.

Table 1. Tracking error 1

	average [cm]	standard deviation [cm]
Z axis direction	6.02	2.62
XY plane	4.34	2.70

Table 2. Tracking error 2

	average [cm]	standard deviation [cm]
Z axis direction	3.25	1.77
XY plane	2.17	1.20

Table 3. Variation of a user's head position in every frame using the same image sequences

	average standard deviation [cm]	maximum standard deviation [cm] (frame)
Z axis direction	0.27	0.50 (#36)
XY plane	0.36	0.68 (#37)

4.3. Effectiveness of integrating color information and distance information

An experiment was performed to investigate the effectiveness of integrating color information and distance information for the same image sequence. In the weight calculation in the particle filter framework, the environment model was also integrated.

Figure 10 shows the observational result for the color information and the distance information in frame #40. For each frame, the hypothesis generated by the particle filter is projected onto the image of each camera, and the points corresponding to the result of evaluation by the color information and the distance information are shown superimposed on the acquired color image. The brightness represents the weight determined from the observational result and is higher for higher weights.

The following observations are made regarding the evaluation by the color information. When the head is oriented close to the direction of the camera, a model with a large skin region is applied as the color information model. Thus, the weight is higher near the center of the head. However, if there is an object in the background with a color close of that of the skin, the weight of the hypothesis projected toward that object is evaluated as high.

As regards the evaluation of the distance information, the following tendency is observed. The weight is larger for hypotheses not only for the human head, but also for the human torso, which is close to the human head position as seen from the camera. On the other hand, the weight is lower for objects in the background.

Thus, the weight based on the color information and the weight based on the distance information are both larger for hypotheses close to the true value, making the weight larger after integration.

(a) #40 (Color) (b) #40 (Depth)

Fig. 10. Evaluation result based on both color and depth information.

(a) #58 (Camera 1) (b) #58 (Camera 2)

Fig. 11. Evaluation result without an environment model.

4.4. Effectiveness of environment model

To investigate the effectiveness of introducing the environment model, an experiment was performed for the same image sequence, with and without the use of the environment model.

When the environment model was not used (#60), the hypotheses diverged to the outside of the view field of the camera, even though the tracking target stayed in the view field of each of the four cameras. Thus, tracking failed. Figure 11 shows the state of the particle filter in the immediately preceding frame, together with the evaluation of the color information. For comparison, Fig. 12 also shows the result when the environment model was used.

When the environment model was not used, the following situation occurred in Fig. 11. False detection occurred in camera 1 (#58) due to the color evaluation of an object with a color close to that of the skin. Thus, the hypothesis was attracted toward a region other than the human head, and false detection occurred in the orientation of the hypothesis. Viewing this situation from another camera 2 (#58), we see that very few hypotheses were generated near the human head. Thus, tracking failed when the environment model was not used.

When the environment model was used, on the other hand, we see from Fig. 12 that a hypothesis was also generated near the human head in camera 2 (#58). Due to

(a) #58 (Camera 1) (b) #58 (Camera 2)

Fig. 12. Evaluation result with an environment model.

the effect of the environment model in which other objects such as desks or the height of the region is considered, the generation of hypotheses by false detection was suppressed. Hypotheses were efficiently generated in the region with high existing probability of the human head, and tracking was continued.

Thus, the stability of human tracking is improved by introducing the environment model.

5. Conclusions

This paper has proposed a method for real-time tracking in a 3D space considering the orientation of the human head, by integrating cues from multiple sensors.

In the tracking of the human head by using a particle filter, the likelihood of a human head is evaluated on the basis of color and distance information obtained from multiple stereo-vision cameras. In addition, an environment model obtained from the range sensor is incorporated. With this approach, the estimation of the position and orientation of the human head is realized stably, considering the distribution of the existing probability of the human head in the indoor space.

Successful tracking was demonstrated in a tracking experiment for the human head in a real environment. In the natural motions of a human in an indoor space with a complex background, such as standing up and walking in arbitrary directions, the orientation of the head was estimated even when the orientation of the head with respect to the camera direction changed. This is due to the integration of color and distance information, so that the effect of false detection in individual evaluations is reduced, even if there is an object in the background with a color close to that of the skin, and larger weights are assigned to hypotheses closer to the true value. It was also shown that by incorporating an environment model containing the existing probability of the human head, more stable tracking was realized by reducing the effect of false detection of other objects, such as desks or walls.

In future studies, we plan to integrate multiple cues by considering observation reliability, such as distinguishing cases in which the tracking target is hidden by occlusion. A method will also be considered to acquire the behavioral history of the person by long-term observation, and to incorporate data such as the frequent routes into the environment model.

Acknowledgment. Part of this study was supported by a Ministry of Education, Science and Technology research grant for the Special Study "Realization of Flexible Human-Machine Interaction Based on Understanding of Human Intentions and Behavior" (No. 13224051).

REFERENCES

1. Birchfield S. Elliptical head tracking using intensity gradients and color histograms. Proc IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'98), p 232–237.
2. Haritaoglu I, Harwood D, Davis LS. A real-time system for detecting and tracking people in 2 1/2 D. Proc 5th European Conference on Computer Vision (ECCV'98), p 877–892.
3. Isard M, Blake A. Condensation—Conditional density propagation for visual tracking. Int J Comput Vis 1998;29:5–28.
4. Isard M, Blake A. ICondensation: Unifying low-level and high-level tracking in a stochastic framework. Proc 5th European Conference on Computer Vision (ECCV'98), Vol. 1, p 893–908.
5. Loy G, Fletcher L, Apostoloff N, Zelinsky A. An adaptive fusion architecture for target tracking. Proc 5th IEEE International Conference on Automatic Face and Gesture Recognition (FG '02), p 261–265.
6. Martin WN, Aggarwal JK. Volumetric descriptions of objects from multiple views. IEEE Trans Pattern Anal Mach Intell 1983;5:150–158.
7. Swain M, Ballard D. Color indexing. Int J Comput Vis 1991;7:11–32.
8. Taycher L, Darrell T. Range segmentation using visibility constraints. Int J Comput Vis 2002;47:89–98.
9. Terrillon J, Pilpre A, Niwa Y, Yamamoto K. Analysis of human skin color images for a large set of color space and for different camera systems. Proc IAPR Workshop on Machine Vision Applications (MVA '02), p 20–25.
10. Zarit BD, Super DJ, Quek FKH. Comparison of five color models in skin pixel classification. Proc International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, p 58–63, 1999.
11. Ukita M, Matsuyama R. Real-time collaborative tracking of multiple objects by a set of active vision agents. Trans CVIM Inf Process Soc 2002;43:64–79.
12. Ohtsuka K, Takekawa N. Analysis of mutual occlusion and sequential estimation of state for multiple objects by multi-point observation. Trans CVIM Inf Process Soc 2003;44:109–125.
13. Sugimoto A, Taniuchi K, Matsuyama R. Tracking human heads based on interaction between hypotheses with certainty. Proc of Scandinavian Conference on Image Analysis, p 617–624, 2003.
14. Nakajima T, Hamazaki K, Okaya T, Deguchi K. Tracking of multiple human objects by fusion of multi-viewpoint images by CONDENSATION. Tech Rep IEICE MIRU 2002;2:317–322.

AUTHORS (from left to right)





Tatsuya Suzuki received an M.E. degree in Information and Communication Engineering from the University of Tokyo in 2004. His thesis concerned human tracking techniques using multiple cameras in indoor environments. He is currently at Hitachi Ltd.

Shinsuke Iwasaki completed the M.E. program in Information and Communication Engineering at the University of Tokyo in 2005 and joined Toyota Motor Corporation.

Yoshinori Kobayashi completed the M.E. program in the Department of Information Management Science at the University of Electro-Communications in 2000 and joined the Design Systems Engineering Center of Mitsubishi Electric Corporation. He is now in the doctoral program in Information and Communication Engineering at the University of Tokyo.

Yoichi Sato received a B.S. degree in mechanical engineering from the University of Tokyo in 1990 and M.S. and Ph.D. degrees in robotics from the School of Computer Science, Carnegie Mellon University, in 1993 and 1997. He then joined the Institute of Industrial Science at the University of Tokyo, where he is currently an associate professor. His primary research interests are in the fields of computer vision (physics-based vision, image-based modeling), human–computer interaction (perceptual user interface), and augmented reality. He is a member of IEEE.

Akihiro Sugimoto received his B.S., M.S., and D.Eng. degrees in mathematical engineering from the University of Tokyo in 1987, 1989, and 1996. After working at Hitachi Advanced Research Laboratory, ATR, and Kyoto University, he joined the National Institute of Informatics where he is currently a professor. He is interested in mathematical methods in/for engineering. In particular, his current main research interests are discrete mathematics, approximation algorithm, vision geometry, and modeling of human vision. He received a Paper Award from the Information Processing Society in 2001.