# Human Head Tracking in Three Dimensional Voxel Space

Haruki Kawanaka
Faculty of Info. Sci. and Tech.
Aichi Prefectural University
Aichi, 480-1198, Japan
kawanaka@ist.aichi-pu.ac.jp

Hironobu Fujiyoshi
Dept. of Computer Science
Chubu University
Aichi, 487-8501, Japan
hf@cs.chubu.ac.jp

Yuji Iwahori
Dept. of Computer Science
Chubu University
Aichi, 487-8501, Japan
iwahori@cs.chubu.ac.jp

## Abstract

*This paper proposes a new approach to track a human head in 3-D voxel space. Information of both color and distance is obtained from multiple stereo cameras and integrated in 3-D voxel space. Formulating a likelihood function from voxel location and its color information can achieve stable tracking with particle filtering in 3-D voxel space.*

## 1. Introduction

For human tracking, particle filtering [3] is one of the techniques for robust tracking in the presence of occlusion and noise. Particle filtering is a maximum posteriori estimation method based on the past and present observation. It also achieves robust tracking for the case which the observed distribution is non-Gaussian. It approximates the discrete probability density where the random variable is represented by many particles.

Previous techniques of tracking by multiple cameras perform 2D-image based tracking on each camera image and obtain the 3D position of the object by the integration of these results[6][2]. However, contradiction occurs when tracking fails in some camera or the tracking results do not correspond to the identical person. In addition, in 2D-image based tracking, the shape and size of the person in an image varies in each camera, making it necessary to always consider the shape and size while tracking.

In this paper, a new head tracking method which uses particle filtering in 3-D voxel space with multiple stereo cameras is proposed. We do not track in each camera image. Instead, we construct a voxel -based representation of the object by integration of the multiple depth maps of all the stereo cameras (hereafter referred as “3D voxel shape”), this is because the size of human head is always constant in 3D voxel space. We then track this 3D voxel shape using particle filtering.

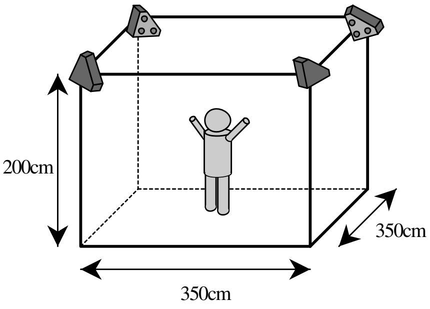

**Figure 1. Tracking Space**

Integration of the information from all cameras and reconstruction of 3-D voxel shape is computationally expensive. We reduce computational cost by reconstructing a low resolution 3D voxel shape. Although, low resolution tracking is inaccurate in general, the use of particle filtering enables robust tracking.

## 2. Integration of images and head tracking

### 2.1. Multiple stereo camera system

The four stereo cameras are arranged at the four corners of a 350[cm] × 350[cm] square on the ceiling of the room at a height of 200[cm], as shown in Fig.1. All cameras are calibrated [5] in advance, so that the internal and external parameters of all the cameras are known. This enables the transformation from the camera coordinate system to the world coordinate system.

### 2.2. Integration of information into voxel space

Color information of skin and hair is a good clue to track the human head. However, color information alone is not

0-7695-2521-0/06/$20.00 (c) 2006 IEEE

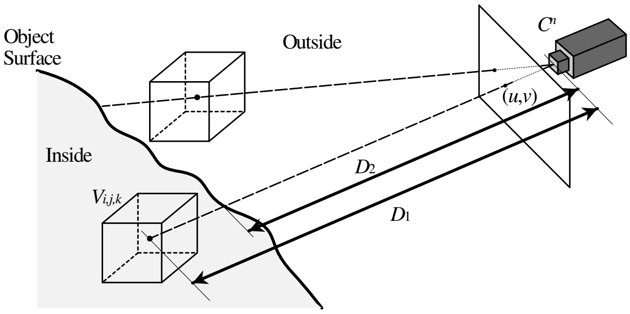

**Figure 2. Decision the inside or outside**

sufficient because the background color may be very similar to foreground objects. The use of distance information from stereo cameras enables us to distinguish the object and the background (or another object). Therefore, the integration of color and distance information enables robust tracking in 3-D space.

Each stereo camera provides a depth map (with associates color information). We back project these depth maps into the 3D voxel space to reconstruct the 3D voxel-shape of the object.

Then, let $D_1$ be the distance from camera $C^n$ to the voxel center $(i,j,k)$ as shown in Fig.2. We project the voxel center $(i,j,k)$ to the camera $C^n$ to obtain image point $(u,v)$. Let $D_2$ be the depth value corresponding to this point $(u,v)$. We can use the values of $D_1$ and $D_2$ to compute a "location" value $L^{C^n}(V_{i,j,k})$ (Eqn 1), which is the "location" value of a voxel computed using information from only one camera $C^n$. $L^{C^n}(V_{i,j,k})$ has one of three possible states "*Inside*", "*Outside*" or "*Surface*".

$$L^{C^n}(V_{i,j,k}) = \begin{cases} Outside & D_1 < D_2 - th \\ Inside & D_1 > D_2 + th \\ Surface & |D_1 - D_2| \leq th \end{cases} \qquad (1)$$

where $th$ is constant.

We integrate the "location" states $L^{C^A}(V_{i,j,k})$ and $L^{C^B}(V_{i,j,k})$ from neighbor cameras $C^A$ and $C^B$ to get $L(V_{i,j,k})$ which is the integrated "location" state of the voxel $V_{i,j,k}$, by the following equation, where the lookup table refers to Table1:

$$L(V_{i,j,k}) = LookUpTable(L^{C^A}, L^{C^B}) \qquad (2)$$

For example, if a voxel state of camera $C^A$ is "*Surface*" and that of camera $C^B$ is "*Inside*", then the state $V_{i,j,k}$ is determined as "*Surface*". Here, the actual number of cameras is four. This look up table is repeatedly used for two among four cameras.

**Table 1. Look Up Table for Integration $V_{i,j,k}$**

| | | $V^{C_A}_{i,j,k}$ | | |
|---|---|---|---|---|
| | | *Outside* | *Surface* | *Inside* |
| $V^{C_B}_{i,j,k}$ | *Outside* | Outside | Outside | Outside |
| | *Surface* | Outside | Surface | Surface |
| | *Inside* | Outside | Surface | Inside |

Further, we assign an associated color value $C(V_{i,j,k})$ (RGB value) to each "*Surface*" voxel. For each "*Surface*" voxel, we have the distance values from each camera. We find the camera with the shortest value of $D_2$ and we use the color value from that camera as the the color value $C(V_{i,j,k})$ of the voxel. Each low resolution voxel projects to a set of pixels. We compute the median value of the color of these pixels as the value of $C(V_{i,j,k})$.

## 2.3. Head tracking using particle filtering

Here, particle filtering [3] is applied to the tracking of object in 3-D voxel space. Particle filtering gives the non-parametric approximation of the probability density function which represents the state of the tracked object as particles.

$$\mathbf{s}^{(i)} = \left\{\mathbf{x}^{(i)}, \pi^{(i)}\right\} \qquad (i = 1, \ldots, N). \qquad (3)$$

Here, $N$ is the number of particles.

For each particle which has the property $\mathbf{x}$ in the state space, its probability becomes proportional to the weight $\pi$. Let $\mathbf{x}_t$ be the property of the tracked object at time $t$ and $\mathbf{Z}_t$ be the observation result. Here, the result obtained by time $t$ is represented as $\mathbf{Z}_t = \{\mathbf{z}_1, \ldots, \mathbf{z}_t\}$. Particles change based on the state space model given in advance. The posterior probability distribution $p(\mathbf{x}_t|\mathbf{z}_t)$ of state $\mathbf{x}_t$ is based on the Bayesian estimation given by the following equation.

$$p(\mathbf{x}_t|\mathbf{Z}_t) = \alpha p(\mathbf{z}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{Z}_{t-1}) \qquad (4)$$

where, $\alpha$ is the constant for normalization.

$p(\mathbf{x}_{t-1}|\mathbf{z}_{t-1})$ is obtained from $\{\mathbf{s}^{(n)}_{t-1}, \pi^{(n)}_{t-1}\}_{n=1,\ldots,N}$ at $t-1$. The hypotheses for the next step are selected according to the ratio of weights $\{\pi^{(1)}_{t-1}, \cdots, \pi^{N}_{t-1}\}$. In this paper, let the world coordinate $\mathbf{x} = (x_w, y_w, z_w)$ of the tracked head position be the state of the tracked object. We update the state information using $\mathbf{x}_t = \mathbf{x}_{t-1} + \omega_t$ , where $\omega_t$ is the gaussian noise. It generates the $N$ hypotheses $\mathbf{s}^{(n)}_t$. The weight $\pi^{(n)}_t$ of the new samples $\mathbf{s}^{(n)}_t$ is obtained from the information of the voxels. Note that the sum of weights is normalized to 1. As a result, the approximation of $p(\mathbf{x}_t|\mathbf{z}_t)$ at time $t$ is obtained. In addition, expectation of the hypotheses is adopted as an optimum estimation of the state quantity. The weight $\pi_t^{(i)}$ is calculated using the likelihood function $L(\mathbf{z}_t|\mathbf{x}_t)$. Then particles are resampled based on the *Sequential Importance Sampling* [1].

Next, $p(\mathbf{z}_t|\mathbf{x}_t)$ is calculated as the likelihood $L(\mathbf{z}_t|\mathbf{x}_t)$. $L(\mathbf{z}_t|\mathbf{x}_t)$ is computed from two histograms, a location histogram and color histogram, in order to track robustly. A location histogram of the number of the cubes (with $a$ voxels on a side) whose center is the voxel which includes $\mathbf{x} = (x_w, y_w, z_w)$ becomes $H_{id}^{location}(id =$ "*Inside*", "*Surface*", "*Outside*"$)$. Here, the histogram $H^{ref\_location}$ for the reference is obtained beforehand from the voxel "location" information of the person head. The similarity $S^{location}$ between $H^{location}$ and $H^{ref\_location}$ is calculated using *Swain histogram intersection* [4].

$$S^{location} = \sum_{id=1}^{3} \min(H_{id}^{location}, H_{id}^{ref\_location}) \qquad (5)$$

Similarly, the color histogram $H^{ref\_color}$ for the reference is obtained beforehand from color information of the voxel of the person head. And let $H_{id}^{color}$ be a histogram in the cube territory ($id = 1, \ldots, T^3$). The similarity $S^{color}$ between $H^{color}$ and $H^{ref\_color}$ is given as

$$S^{color} = \sum_{id=1}^{T^3} \min(H_{id}^{color}, H_{id}^{ref\_color}) \qquad (6)$$

where, $T$ is the tone of color. In the experiment, tone is taken to be 16 (4bit).

The histograms which consist of color information and voxel location information are the observed values. A likelihood function is given as

$$L(\mathbf{z}_t|\mathbf{x}_t) = \exp(k_1 S^{location}) \exp(k_2 S^{color}) \qquad (7)$$

where, $k_1$ and $k_2$ are the constants. This makes it possible to assign a high weight to the hypothesis which has a high probability for the existence of the person in the room.

## 3. Experiment

The experimental system consists of 4 units of the IEEE1394 stereo camera $Digiclops^{TM}$ (Point Grey Research Inc.) and a PC of CPU Intel Xeon 3.6 GHz with Main Memory 2.0 Gbyte. All cameras are synchronized by a SyncUnit and the cameras are calibrated. Both color image and depth map (320×240pixels) are obtained from the stereo camera at 24[fps]. Under the condition that the initial value for the status of the tracking head is known, the number of particles was taken as $N = 100$ and the voxel size of head was taken as 6×6×6[voxel] (1[voxel] = 3.5[mm]).

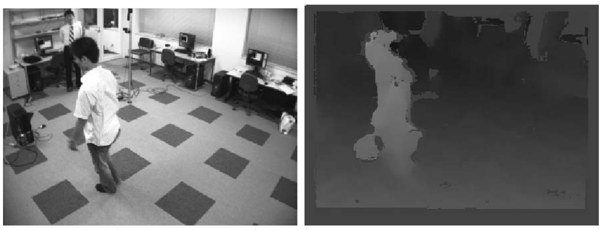

**Figure 3. Color information and depth map**

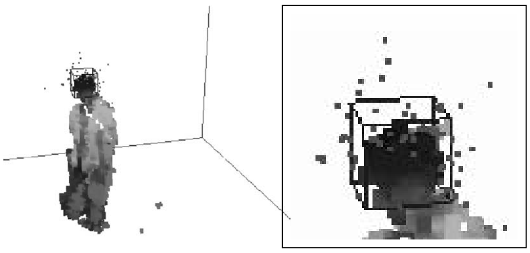

**Figure 4. Appearance of voxel and particles**

To check the performance of tracking with the particle filtering in the voxel space, we evaluated our method for some image sequences. The calculation time was around $200[ms]$ for the integration process and around $2[ms]$ for the tracking process.

We tested tracking performance for various image sequence which included (a) walking from the center of the room in counterclockwise direction, (b) crouching down and (c) standing up. Examples of the image and the depth map used in the experiment are shown in Fig.3. An example of the voxel space obtained after the integration processing is shown in Fig.4. The points around the human head show the particles of the particle filtering. A wire frame cube shows an estimated location of the head. The tracking process is shown in Fig.5. We see that the center of the human head can almost be estimated using the expectation of the multiple hypotheses. This is because the hypotheses which have high weights gather around the human head.

The trajectory of the tracking sequence is shown in Fig.6. In the case that the likelihood function has only voxel location information (Eqn.5) or only color information (Eqn.6), the method produces wrong tracking trajectories. However, our method achieved reliable tracking because we use both

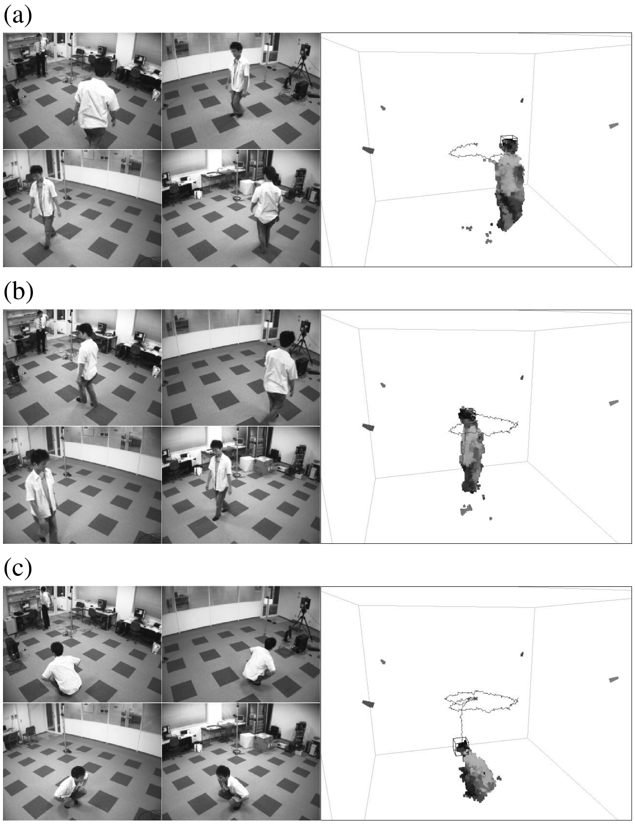

**Figure 5. Camera images and voxel space with trajectory. ( frame number : (a) 134, (b) 208, (c) 343 )**

the location and color histogram. Although the voxel data is low resolution (as depth maps are not very accurate), the method could successfully track a human head continuously over 1000 frames.

## 4. Conclusion

This paper presents a new approach which integrates information from multiple stereo cameras into the voxel space and tracks the human head by particle filtering.

The use of both the voxel location information and the color information obtained from multiple stereo cameras to estimate the position of human head in real-time, enables our method to achieve reliable tracking in 3D voxel space with low resolution. The method is experimentally validated by testing on various sequences involving (a) walking from the center of the room in counterclockwise direction, (b) crouching down, and (c) standing up.

Increasing the speed of the “integration process” is remained as future work.

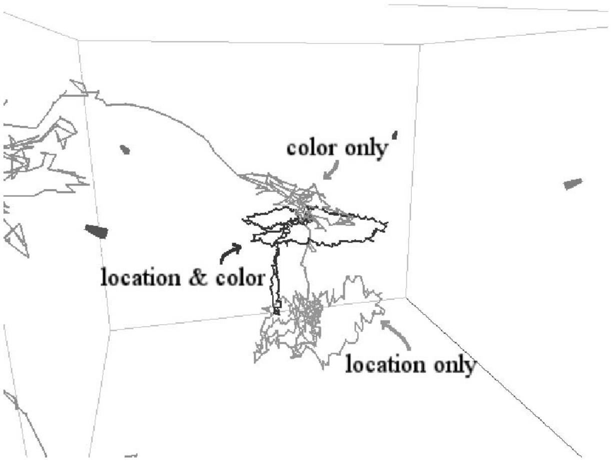

**Figure 6. Trajectory of the image sequence**

**Acknowledgment** This research was partly supported by a grant of the “High-Tech Research Center” Project from Ministry of Education, Culture, Sports, Science and Technology (2005-2009). Iwahori’s research is supported in part by the Grant-in-Aid for Scientific Research (No.16500108) from Japan Society for the Promotion of Science (JSPS) and Chubu Univ. Grant.

## References

[1] A. Doucet, S. Godsill, and C. Andrieu. On sequential monte carlo sampling methods for bayesian filtering. *Statistics and Computing*, 10:197–208, 2000.

[2] H. Hamasaki, T. Nakajima, T. Okatani, and K. Deguchi. Understanding 3-dimensional dynamic environment fusing multiple images by mixed-state condensation algorithm. *Soc. of Instrument and Control Eng. 2002*, pages 1510–1515, 2002.

[3] M. Isard and A. Blake. Condensation- conditional density propagation for visual tracking. *Intl. J. Computer Vision*, 29(1):5–28, 1998.

[4] M. J. Swain and D. H. Ballard. Color indexing. *Int. J. Computer Vision*, 7(1):11–32, 1991.

[5] R. Y. Tsai. A versatile camera calibration technique for high accuracy 3-d machine vision metrology using off the shelf tv cameras and lenses. *IEEE J. of Robotics and Automation*, 3(4):323–344, 1987.

[6] N. Ukita and T. Matsuyama. Incremental observable-area modeling for cooperative tracking. *15th Intl. conf. on Pattern Recognition*, pages 192–196, 2000.