End Point Tracking for a Moving Object with Several Attention Regions by Composite Vision System

Kotaro Nagahama, Tomohiro Nishino, Mitsuharu Kojima,
Kimitoshi Yamazaki, Kei Okada and Masayuki Inaba
Department of Mechano-Informatics,
Graduate School of Information Science and Technology,
The University of Tokyo
7-3-1, Hongo, Bukyo-ku, Tokyo, JAPAN
Email: {nagahama, tomohiro, kojima, yamazaki, k-okada, inaba}@jsk.t.u-tokyo.ac.jp

***Abstract*— This paper describes an approach of multi-target tracking for gaze control to know the motions of end points on a moving object. In order to track several moving parts from image streams, three different types of tracker to observe temporal, spatial and appearance changes are combined. Also, we developed composite vision system on which two wide-angle cameras and two zoom-enabled cameras are mounted. We tested the gaze control system and the head system by observing a human working in daily environment. This results showed the effectiveness of our approach.**

***Index Terms*— Tracking, Composite Vision System**

I. INTRODUCTION

For robots, it is necessary to track multiple regions at one time. It is useful not only to estimate the correlation of tracked objects to each other but also to estimate the joint structure of a object [1].

The density of the attention regions should also be diverse. Fig.1 shows a situation in which a robot observes a human who is cleaning with a broom and a dustpan. The robot knows the correlation between two objects and the type of dust by watching and tracking in detail the 3-D motions of the end points of the tools. The robot should also pay attention to the motion of the whole person and changes in her attention. This enables the robot to estimate her next motion, and pass to her the necessary objects based on that estimation, for example.

Human beings can watch objects in detail while finding changes in the appearance in whole environments. This is because they have an unevenness distribution of visual cells on their retinas. For robots, retina-like lenses or cameras have been developed [2], [3]. But they are not suitable for estimating the trajectories of objects' 3-D motion because ordinary methods of stereo calculation based on disparity cannot be applied to them.

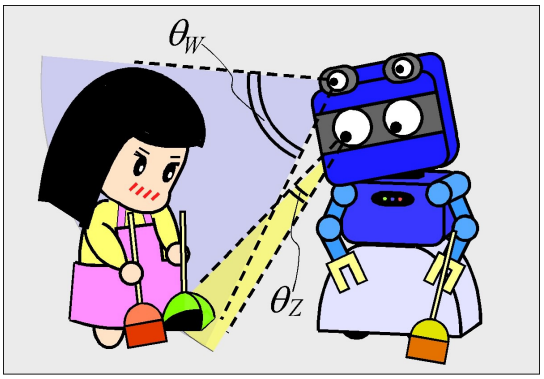

Fig. 1. Observing the motion of a person manipulating objects by multi-cameras. θ_W is the view angle of wide-angle cameras, θ_Z is that of zoom cameras

In order to solve above issues, we take the following approaches:

1) A tracking function to know the movements of end points of a object, for instance, human face, hands or handled tools, is implemented. Their 3-D trajectories are estimated with a high frame rate.
2) A composite robot head system which equips different types of cameras is built up. Two omni-directional cameras and two zoom cameras are mounted on the head. The former is useful to capture the images of large environments or to track the whole body of a person. The latter has a role to track attention region in detail.

For item 1), a tracking method based on continuous disparity filter is applied. 3-D motions of end points are tracked in parallel by this method using a pair of cameras. Tracking quality is stochastically evaluated by filtering three different types of information. (1) Spatial distance, (2) temporal change of texture pattern and (3) appearance information are used to calculate maximum likelihood of target regions. This information provides robust tracking results against area deformation and transient occlusion. Section II describes this in detail.

For item 2), we build a sensor head with two different types of cameras. In past research, composite camera systems have been proposed and implemented. Asfor et al. [4] developed a compound vision system with wide and narrow viewing cameras. Our camera system consists of two by two cameras (Fig.4). Two omni-directional cameras are used to track the whole motion of a person or to view a large environment (θ_W in Fig.1). Two zoom-enabled cameras are mounted and used to capture an attention region with appropriate resolution (θ_Z in Fig.1). Because the maximum magnifications of the zoom is 9.5, various scenes can be captured with diverse resolutions. In addition, to deal with targets at various distances, the zoom cameras are assembled into rotation axes to create convergence. Both stereo pairs are calibrated so that they can provide 3-D information based on their disparities. Section III

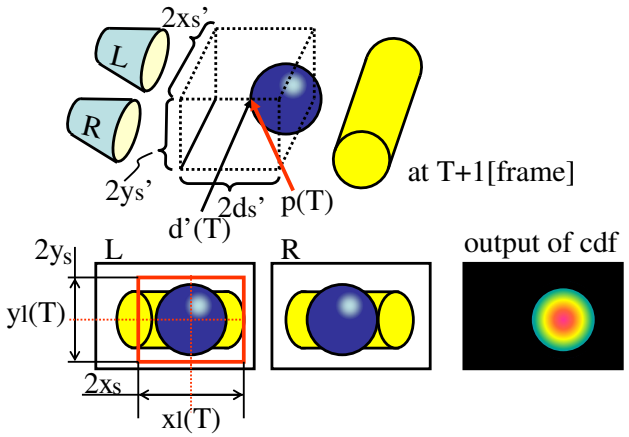

Fig. 2. Continuous disparity filter. At T+1[frame], only the pixels near $p(T)$ are filtered, so the region of the cylinder in the back is filtered out. x'_s, y'_s, d'_s and $d'(T)$ in 3-D space correspond to x_s, y_s, d_s and $d(T)$ in the stereo images, respectively.

describes this in detail.

The paper is organized as follows. In Section II, our visual processing system for tracking end points is explained. Section III provides the hardware configuration and the software system of our composite sensor head, "CV-head"(Composite Vision-head). In Section V, we evaluate our visual processing system and "CV-head" in three experiments in which the robot tracked human motions in daily environments.

II. A TEMPORAL-SPATIAL TRACKING METHOD BASED ON LOW LEVEL VISUAL FUNCTIONS

This section explains our end point tracker, a temporal-spatial tracking method based on low level visual functions. The tracker finds and tracks the motions of end points of a moving target such as a human head, hands and handling objects.

A. An End Point Tracking Method Based on Temporal-Spatial Shape Continuity

At each frame, the input of the end point tracker is stereo images at that time and the former attention position or initial attention position provided by other visual processes. The position is used as the center of the local attention region in which 3-D filtering is done. The output of the tracker is attention position at that point, which is calculated based on the evaluation of shape continuity and other visual features.

The end points trackers can be run in parallel while they communicate their positions each other.

B. Evaluation of 3-D Shape Continuity and Spatial Filtering

One of the basic visual functions for the end point tracking is to move the tracker so that it follows temporally and spatially-continuous regions. Our tracking method first evaluates the continuity of a 3-D shape in the local region near the position of the tracker at the former frame (Fig.2). Then, only the continuous region is used to process other visual functions. We call this "Spatial Filtering".

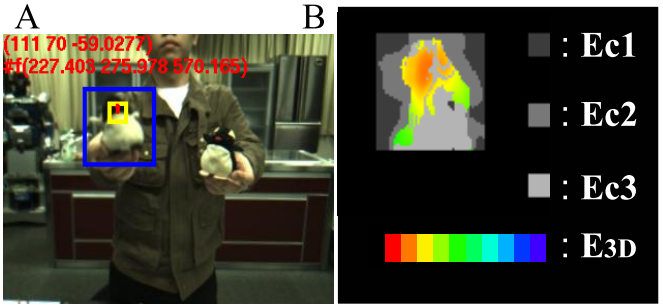

Fig. 3. An output of the continuous disparity filter. A is a left camera image and blue rectangle in it is the search area ($2x_s \times 2y_s$). B is the output of the filter. The color pixels show the points which are not filtered out. Evaluation on distance (E_{3D}) with bias $\boldsymbol{v}_{bias}$ = (0mm, -10mm, -10mm) at camera coordinate system is shown by color graduation.

For spatial filtering, we use "continuous disparity filtering". This is an expansion of "zero disparity filtering" [5], which abstracts features in zero disparity fields. Because the continuous disparity filter is able to abstract features from any disparity field, it can be used to general parallel tracking.

The input of the continuous disparity filter is the set of (1) the position in the left camera image and the disparity $\boldsymbol{p}_I(T)$, or 3-D position $\boldsymbol{p}(T)$ of the end point at time T[frame], and (2) stereo images at T+1[frame]. They are described as follows,

$$\begin{aligned}
&\mathbf{p}_I(T) = (x_l(T), y_l(T), d(T)),\\
&\mathbf{p}(T) = (x(T), y(T), z(T)),\\
&\mathbf{p}_I(T+1,i) = (x_l(T+1,i), y_l(T+1,i), d(T+1,i)),\\
&\mathbf{p}(T+1,i) = (x(T+1,i), y(T+1,i), z(T+1,i)),
\end{aligned} \tag{1}$$

where $(x_l(T),\ y_l(T))$ is the position of the end point in the left image at T[frame], $d(T)$ is the disparity and $\boldsymbol{p}(T)$ is the 3-D position of the point on the camera coordinate. i indicates the index of the pixel in filtered region.

Let $2x_s$ and $2y_s$ be the width and the height of the search area for the continuous disparity filter in the left image. Let $2d_s$ be the disparity range of the search area. If there exists a disparity $d(T+1,i)$ that satisfies the equation below,

$$d(T) - d_s \leq d(T+1,i) \leq d(T) + d_s, \tag{2}$$

and in which stereo calculation succeeds, the position with the disparity becomes a candidate of the next end point. It is not filtered out by the continuous disparity filter. Note that the point $(x_l(T+1,i), y_l(T+1,i))$ must satisfy the equations below in the left image

$$\begin{aligned}
x_l(T) - x_s \leq x_l(T+1,i) \leq x_l(T) + x_s\\
y_l(T) - y_s \leq y_l(T+1,i) \leq y_l(T) + y_s.
\end{aligned} \tag{3}$$

The continuous disparity filter outputs the 3-D positions of the next end point candidates $(x(T+1,i), y(T+1,i), z(T+1,i))$ when an appropriate $d(T+1,i)$ exists.

SAD (Sum of Absolute Difference) is applied to calculate correlation between left and right images. Then, subpixel registration is executed by means of the quadratic estimator [6]. The stereo calculation is adequately performed in the limited disparity $d(T+1,i,j)$ which satisfies $d(T,i) - ds \leq$

$d(T+1,i,j) \leq d(T,i) + ds$, where j is an index of the disparity, through the following three evaluation steps:

1) Evaluation on the position of the disparity which indicates the minimum SAD in the search disparity region (E_{C1}):
   At the left image point $(x_l(T+1,i), y_l(T+1,i))$, if the disparity $d(T+1,i)$ that indicates the minimum SAD is at the edge of the search range, in other words $d(T+1,i) = d(T) - ds$ or $d(T+1,i) = d(T) + ds$, the point is filtered out.
2) Evaluation on the maximum correlation value (E_{C2}):
   If the maximum correlation of the point is lower than the constant minimum value, in other words the minimum SAD is larger than the constant value, the point is filtered out.
3) Evaluation on the position of the disparity which indicates the second highest correlation (E_{C3}):
   If there exists a disparity which is far enough from the disparity with the highest correlation and is evaluated as too high, the point is filtered out. This serves to filter out low textured points for low reliability of the stereo calculation.

The colored points in Fig.3 are the outputs of the continuous disparity filter, and the gray points indicate regions filtered out by the three evaluation steps (E_{C1}, E_{C2}, E_{C3}).

The evaluation of the 3-D distance between $\mathbf{p}(T)$ and $\mathbf{p}(T+1,i)$ is expressed by the following equation:

$$E_{3D}(T+1,i) = exp\{-C||\boldsymbol{p}(T+1,i) - \boldsymbol{p}(T) - \boldsymbol{v}_b||\}, \quad (4)$$

where C is a constant positive number, $\boldsymbol{p}(T)$ is the position of the end point at $t = T$[frame], and $\boldsymbol{p}(T+1,i)$ is the position of the candidate i of next end point at $t = T+1$[frame]. $\boldsymbol{v}_b$ expresses the velocity directivity of the region around a target point. E_{3D} takes the value from zero to one. The shorter the distance, the higher the value is. If all the local candidates are filtered out by the spatial filter at $t = T+1$[frame], the position at $t = T$[frame] is used as the result at that time. This enables the recovery from the temporary failure of stereo calculations in the local region by an occlusion, for example.

C. Multiple Evaluations of Filtered Image

In order to determine the next end point, the tracker evaluates the spatially-filtered image by the procedure mentioned above.

1) Evaluation of 2-D Temporal Change:
For the evaluation of temporal change, for example, the evaluation using similarity of the texture is realized by the following E_{temp}:

$$E_{temp}(T+1,i) = 1 - S(T+1,i), \quad (5)$$

where $S(T+1,i)$ is the normalized SAD between the local image of the former position of the tracker at $t = T$[frame] and that of the position i at $t = T+1$[frame]. The value is divided by the number of the pixels in the local image. It takes the value between 0 and 1.

2) Evaluation Using Registered Appearance:
For evaluation using registered appearance, color information is applied. In our case, the evaluation method is set as follows:

$$E_{apr}(T+1,i) = E_{color}(T+1,i) = N_{color}(T+1,i)/N_{all}, \quad (6)$$

where the N_{all} is the number of the pixels in the target region, and the N_{color} is the number of the pixels of the registered color in the region.

3) Integration of Multiple Visual Functions:
For the integration of multiple evaluations by various visual functions, the following $E_{all}(T+1,i)$ is used.

$$E_{all}(T+1,i) = E_{3D}(T+1,i)^{C_{3D}} \cdot E_{temp}(T+1,i)^{C_{temp}} \cdot \prod_k E_{apr_j}(T+1,k)^{C_{apr_k}}, \quad (7)$$

where k is the index of multiple evaluations using registered appearances. The C_* is a constant positive number which represents the weight in integrating multiple evaluations. C_* changes the action of the end point tracker. The candidate's position i where the integrated evaluation E_{all} has the highest value is used as a target end point at time T+1[frame].

III. CONFIGURATION OF COMPOSITE VISION SYSTEM "CV-HEAD"

We developed a composite vision system with two types of stereo cameras. The system, which we call "CV-head" hereafter, also have three degrees of freedom which enable vergence eye movement.

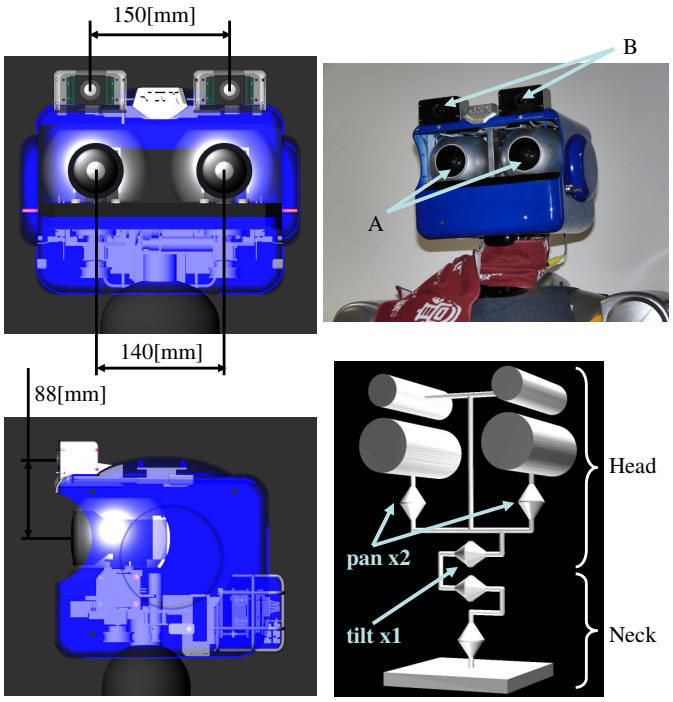

Fig. 4. Configuration of CV-head. A is a set of zoom cameras, and Omni-directional cameras are mounted at B.

A. Hardware System of CV-head

Fig.4 and Table I show the hardware specification of CV-head. One of the stereo cameras, mounted on the top of the CV-head, is made up of two "NM33" cameras made by OPT Co.,Ltd. They have omni-directional lenses, and can capture not only fish-eye images but also square images of arbitrary angles of view. In the experiments, we used these cameras as wide view angle (about 90[degree]) cameras.

Another stereo camera is made up of two "FCB-IX11A" camera modules made by SONY Co.,Ltd. They have zoom lenses and their optical zoom factor can be changed from 1.0 to 9.5 times. Combined with $0.7\times$ wide conversion lens "VCB-HG0730X" by SONY Co.,Ltd, their view angle can be controlled from 65.7[degree] to 6.9[degree]. With a VGA (640[pixel] $\times$ 480[pixel]) image size and its maximum zoom factor ($9.5\times$), about 20/13 vision worth of resolution (USA unit) is acquired.

Each zoom camera can move independently around its pan joint, and the tilt joint is shared by four cameras. For two pan joints and one tilt joint, AC 40[W] motors by Maxon Co.,Ltd and Harmonic Drive gears with a 300:1 gear ratio are used.

For the two pan joints of CV-head, we chose "MEH-30-4500 PST16 (72000)" rotary encoders made by Microtech Laboratory Inc. With these and the 300:1 Harmonic gears, the maximum accuracy of the joint angle of the pan axis is 0.00125[degree/count]. This value is sufficiently smaller than the value of the view angle per pixel at the maximum zoom factor, $6.9/640 = 0.01078$[deg/pixel] at the resolution of QVGA.

B. Software System of CV-head: Calibration and Recalculation of Internal, External Camera Parameters

In the optical zoom stereo camera system with vergence axes at the CV-head, the extrinsic camera parameters are changed by the vergence eyes' motion, and the intrinsic camera parameters are changed by zoom factor. Therefore, the system recalculates both extrinsic and intrinsic camera parameters based on the zoom factor and vergence angles at that point in every visual process. The calibration and recalculation methods are described in this subsection.

1) Calibration and Recalculation of Intrinsic Camera Parameters: First of all, we preliminarily calculated the position

TABLE I

HARDWARE SPECTATION AND KINEMATICS OF CV-HEAD

External size	width 270[mm] $\times$ height 245[mm] $\times$ depth 250[mm]
Weight	3.6[kg]
Actuators	40[W] AC Motor $\times$3 (pan $\times$2, tilt $\times$1)
Movable range of pan joint	-35[degree] to 35[degree]
Movable range of tilt joint	-45[degree] to 45[degree]
View angle of omni-directional cameras	180[degree]
View angle of zoom cameras	about 65.7[degree] to 6.57[degree]
Time to zoom from 1.0X to 9.5X	about 1.1[second]
Height of the zoom cameras when CV-head is mounted on HRP-2V	about 1550[mm]

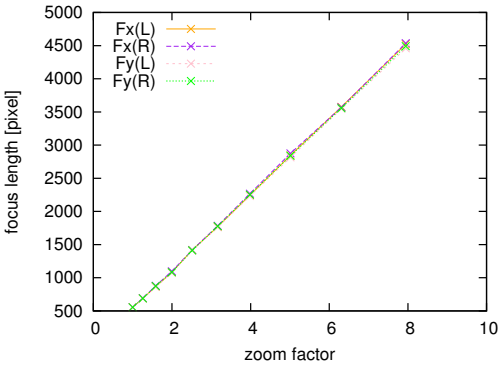

Fig. 5. Correlation between zoom factor and focal point distance

of the zoom center as follows: a) We determined the orbits of four corner points of a fixed calibration pattern in the camera images while we changed the zoom factor. b) We calculated the intersection point of the four orbits and used it as the zoom center and the image center (C_x, C_y).

Second, we determined ten focal point distances F_x, F_y and radial distortion parameters k_1 at different zoom factors ($10^{i/10}\times, (i = 0, 1, 2, ..., 9)$) using the camera calibration procedure proposed by Zhang [7]. Note that distortion parameter k_1 is expressed as

$$\begin{aligned} x &= x_c\{1 + k_1(x^2 + y^2)\} \\ y &= y_c\{1 + k_1(x^2 + y^2)\}, \end{aligned} \tag{8}$$

where (x, y) is the position of a pixel in the raw image, and (x_c, y_c) is that in the undistorted image.

Fig.5 shows the correlation between the zoom factor and the focal point distance of two zoom cameras. Because they have a linear correlation, the camera calibration system recalculates intrinsic camera parameters with linear approximation using the zoom factor at that point.

2) Calibration and Recalculation of Extrinsic Camera Parameters: Extrinsic camera parameters of zoom cameras depend on the coordinate transformation from a pan axis to another pan axis, the transformation from each pan axis to the zoom camera and the joint angles of two pan axes.

We used the kinematic calibration procedure by Welke, et al. [8], and calculated the transformation from each pan joint coordinate system to its camera coordinate system. Then we applied normal extrinsic calibration between a set of zoom cameras. Using this procedure and the kinematic calibration procedure above, the transformation between the two pan joints' coordinate systems were estimated. Therefore the system memorizes the transformation between the two pan joints' coordinate systems and the transformation between each pan joint coordinate system and its zoom camera.

In the experiments, extrinsic camera parameters of the zoom stereo cameras are recalculated based on these transformation data and joint angles of two pan joints at that point.

Fig.6 shows the undistorted and rectified stereo images of zoom cameras and depth images calculated using the recalculated internal and external parameters at differenct zoom factors (z) and pan joint angles (θ_L, θ_R).

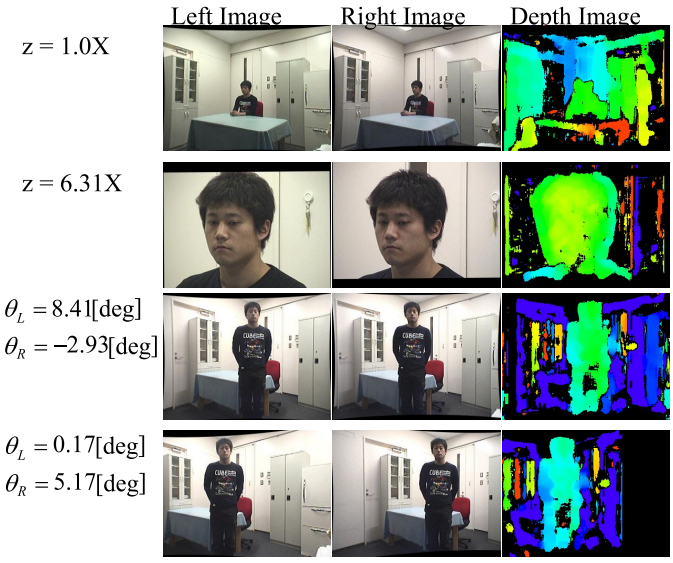

Fig. 6. Undistorted and rectified stereo images of zoom cameras and depth images calculated using the recalculated internal and external parameters at different joint angles of pan axes and zoom factor.

IV. EXPERIMENTS

In this section, we tested the end point tracking system and the CV-head by observing a human working in daily environment.

A. Experiment 1: Tracking a Human Head by Zoom Cameras

1) Experimental setup: In order to evaluate the temporal 3-D tracking performance of the proposed end point tracking method, we tested the method in an experiment of tracking a human head.

After a person faced the robot, he walked at about 3[km/h] around a table in a room. The CV-head started tracking his head soon after it detected his face using a boosted cascade of features [9]. The 3-D position of detected face was used as the initial input of the end point tracker.

For the evaluation of the end point tracker, each of the three evaluation methods shown in section II was used independently in the experiment 1.1. Then, all three evaluations were combined by (7) in Experiment 1.2.

The image size was QVGA (320[pixel]×240[pixel]) and images were captured by zoom cameras. The window size of the continuous disparity filter was 16[pixel]×16[pixel], and the search region of the continuous disparity filter was width 31[pixel]×height 31[pixel]×disparity 7[pixel].

In Experiment 1.2, each constant number shown in (7) was set as $C_{3D} = 0.08$, $C_{apr} = 0.1$ and $C_{temp} = 1.1$. In addition, the velocity directivity was set as $\boldsymbol{v}_{bias} = (0, -30[mm], -30[mm])$.

2) Experimental result: In Experiment 1.1, all three tracking failed. Here, the "fail" means that the tracking result was outside of a walking person. Tracking processes based only on the continuity of 3-D shape, registered appearance and temporal change failed at 11[frame], 27[frame], and 30[frame] respectively.

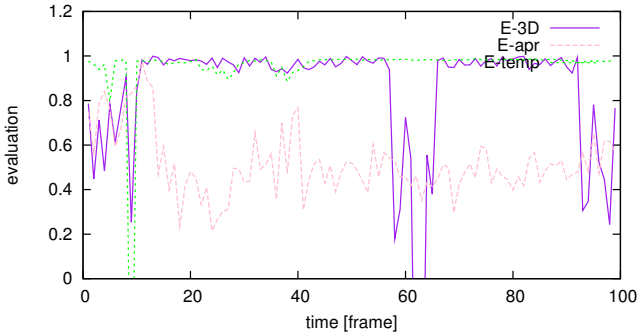

Fig. 7. The transit of the each evaluation in Experiment 1.1

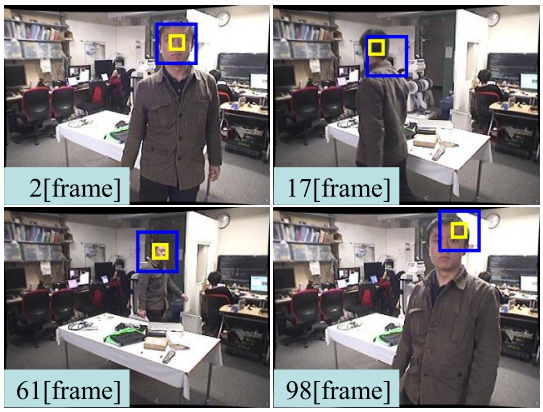

Fig. 8. Result of the experiment of tracking human head based on integrated evaluation

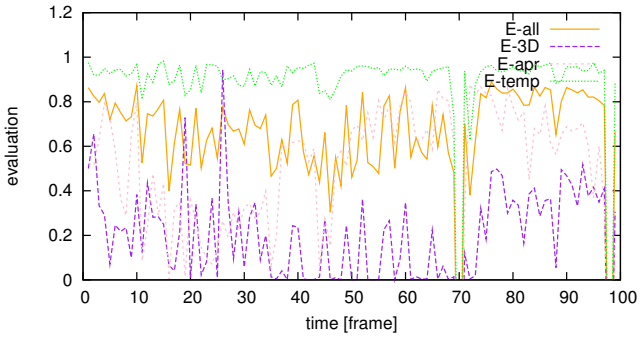

Fig. 9. The transit of the evaluation E_{all}, E_{3D}, E_{apr} and E_{temp} in the experiment 1.2

The transit of the each evaluation is shown in Fig.7. Each evaluation kept a high value, but it did not lead to the right result. Note that the point with zero evaluation value is when all the local candidates are filtered out by the spatial filter and a failure recovery process mentioned in Section II was done.

The result of Experiment 1.2 is shown in Fig.8 and Fig.9. Each independent evaluation is lower than that in Experiment 1.2 (Fig.7), but the tracking process succeeded by using the integrated score E_{all}.

B. Experiment 2: Tracking Human Head and Hands, and Tracing Handling Objects

1) Experimental setup: In order to evaluate the 3-D tracking and tracing performance of the end point tracker, we tested

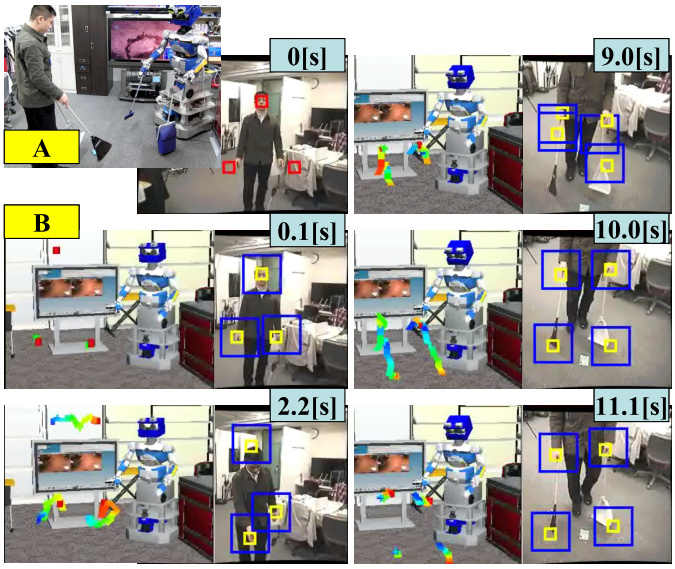

Fig. 10. Result of the experiment of end point tracking and tracing of handling objects of a person

the method in an offline experiment of tracking a person's head, both of his hands, and tracing his handling objects (Fig.10,A). After a human face was detected, the system was programmed not only to start an end point tracker (TR1) to track the face, but also to make two other end point trackers track both human hands (TR2, TR3). The initial tracker positions of TR2 and TR3 were set to the lower left and right of the initial position of TR1, respectively. The velocity directivity of the TR1, TR2 and TR3 were set to $\boldsymbol{v}_{bias} = (0, -B_{1y}, -B_{1z})$, $\boldsymbol{v}_{bias} = (B_{2x}, B_{2y}, -B_{2z}), (-B_{2x}, B_{2y}, -B_{2z})$, respectively, in the camera coordinate system. Furthermore, evaluation on the appearance memory (9) was used at TR2, TR3. It has almost the same meaning as evaluation method (6).

$$E_{apr}(T+1,i) = E_{col}(T+1,i) = exp\left\{C_1\left(\frac{N_{skin}(T+1,i)}{N_{all}} - 1\right)\right\} \quad (9)$$

$N_{skin}(T+1)$ is the number of skin color pixels in the tracked point window at T+1[frame] , and C_1 is a constant positive number.

Then, in the middle of the process, two other end point trackers (TR4, TR5) were made. They did not use appearance evaluations, and the velocity directivity of the TR4 and TR5 were set to $\boldsymbol{v}_{bias} = (B_{2x}, B_{2y}, -B_{2z}), (-B_{2x}, B_{2y}, -B_{2z})$ respectively. The image size was QVGA and the images were captured by zoom cameras. The tracked point's window size was 16[pixel]x16[pixel], and the search region of the continuous disparity filter was width 51[pixel] x height 41[pixel] x disparity 7[pixel].

2) Experimental Result: Fig.10,B shows the result of the experiment. The tracked positions of TR2 and TR3 went to both hands at 0.1[s], and tracked them (2.2[s]). TR4 and TR5 traced the broom and the dustpan held in human hands (9.0[s]), and stopped at the tip (10[s], 11.1[s]).

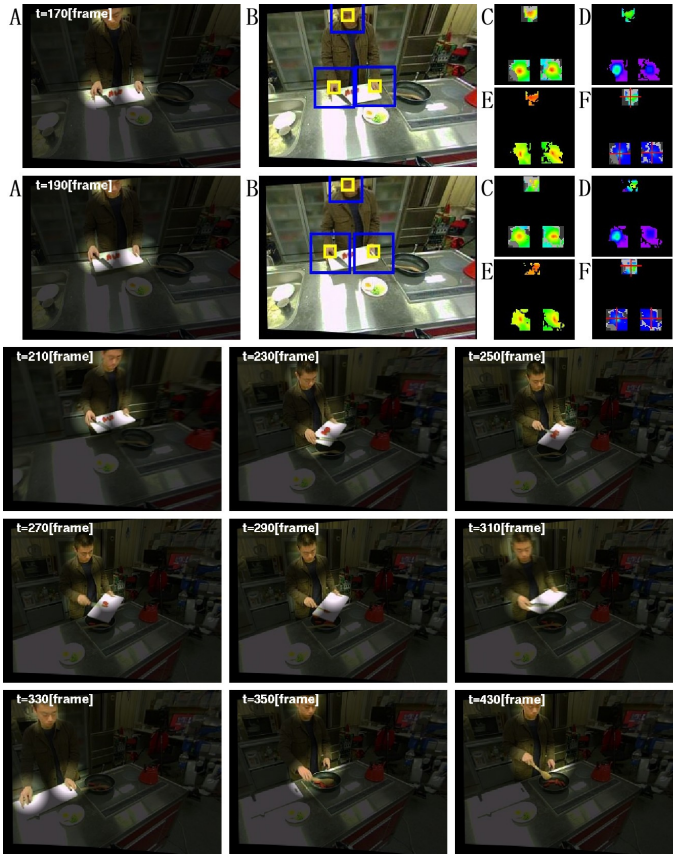

Fig. 11. Result of the experiment of tracking human head and hands

C. Experiment 3: Tracking a Human Head, Hands and Tools Using Both Wide-angle Cameras and Zoom Camera

1) Experimental setup: In order to evaluate the performance of the cooperation of two stereo cameras, We made an experiment to track a human head, both hands, and a tool he had. The wide angle stereo camera images and three end point trackers were used to track human motion. The end point trackers used the appearance evaluation (6) and evaluation of temporal change (5).

In addition, using the zoom cameras' images and a knowledge-based object recognition method using 2-D straight edge cues [10], a tool's coordinate was estimated. Note that the zoom factors of the zoom cameras were determined by a human.

2) Experimental result: Fig.11 shows the result of tracking a human face and both hands by the wide angle stereo camera and three end point trackers. Fig.11,A shows the gaze positions and B shows the windows of the gaze points and the search ranges. C, D, E show the evaluations of 3-D shape's continuity, appearance and temporal change respectively. F shows the integrated evaluation value. Three tracked points were able to track a human face and both hands using wide angle cameras while the coordinate of a cutting board was estimated using the zoom camera as shown in Fig.12.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we introduced a method for end point tracking and a composite camera system for tracking multiple end

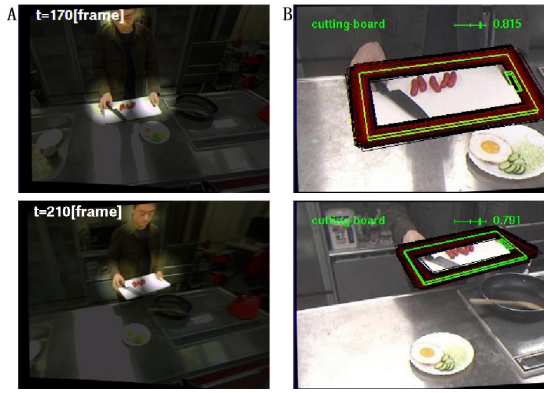

Fig. 12. Result of the experiment of estimating a tool's coordinate

points. Through the experiments, effectiveness of our approach was demonstrated.

For future work, we will try to develop a method to change zoom factors of zoom cameras based on the changes in the images of wide angle cameras. It will be useful to select important regions to be tracked and memorized in detail.

REFERENCES

[1] T. Mori, M. Inaba, and H. Inoue. Visual tracking based on cooperation of multiple attention regions. *In Proceedings of The 1996 IEEE International Conference on Robotics and Automation*, Vol. 4, pp. 2921 –2928 vol.4, apr. 1996.

[2] Y. Kuniyoshi, N. Kita, S. Rougeaux, T. Suehiro, and T. Mitsui. Active Stereo Vision System with Foveated Wide Angle Lenses. *Recent Developments in Computer Vision Lecture Notes in Computer Science*, Vol. 1035, pp. 191–200, 1995.

[3] G. Sandini and G. Metta. Retina-like sensors: motivations, technology and applications. In T.W. Secomb, F. Barth, and P. Humphrey, editors, *Sensors and Sensing in Biology and Engineering*. Springer Verlag, New York, NY, 2002.

[4] Tamim Asfour, Kai Welke, Pedram Azad, Ales Ude, and Rüdiger Dillmann. The karlsruhe humanoid head. In *Proceedings of IEEE-RAS International Conference on Humanoid Robots*, pp. 447–453, 2008.

[5] N. Kita, S. Rougeaux, Y. Kuniyoshi, S. Sakane. Thorough zdf-based localization for binocular tracking. In *Proceedings of IAPR International Workshop MVA'94*, pp. 190–195, 1994.

[6] Qi Tian and Michael N. Huhns. Algorithms for subpixel registration. *Computer Vision, Graphics, and Image Processing*, Vol. 35, No. 2, pp. 220–233, 1986.

[7] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transaction of Pateren Analysis and Machine*, Vol. 22, pp. 1330–1334, 2000.

[8] Welke, Przybylski, Asfour, and Dillmann. Kinematic Calibration for Saccadic Eye Movements. Technical report, Institute for Anthropomatics, University Karlsruhe, 2008.

[9] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, Vol. 1, pp. I–511 – I–518 vol.1, 2001.

[10] K. Okada, M. Kojima, S. Tokutsu, T. Maki, Y. Mori and M. Inaba. Multi-cue 3d object recognition in knowledge-based vision-guided humanoid robot system. In *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3217–3222, 2007.