MVA2002 IAPR Workshop on Machine Vision Applications, Dec. 11 - 13, 2002, Nara - ken New Public Hall, Nara, Japan

3—18

Identifying Body Parts of Multiple People in Multi-Camera Images

Masafumi Tominaga *
HOIP, Softopia Japan / JST

Hitoshi Hongo †
HOIP, Softopia Japan / JST

Hiroyasu Koshimizu ‡
SCCS, Chukyo University

Yoshinori Niwa §
HOIP, Softopia Japan / JST

Kazuhiko Yamamoto ¶
Faculty of Engineering, Gifu University

Abstract

In order to track and recognize the movements of multiple people using multiple cameras, each person needs to be segmented and identified in the image of each camera. We propose a method that tracks multiple people and identifies their body parts in multiple camera images.

Estimation of the human positions and identification among the multiple cameras is mainly based on the silhouettes method combining the background subtraction and the frame subtraction. The noise problem and the combine problem of the persons in the silhouettes method is addressed by use of the existence probability map. The existence probability map consists of the position probability map and height cumulus map. The gesture event is detected by the human block and motion block that are extracted from the voting space using the silhouettes method. We demonstrate the experimental results and the validity of the proposed method.

1 Introdution

Tracking human motion in an indoor environment is of interest in applications of surveillance, intelligent environment and so on. We use multiple cameras mounted in the area of interest to track and monitor the motion of individuals in sequences of color images. In this work, identification of the body parts among multiple cameras images is one of the most important tasks. To establish the correspondence of the detected body parts between different cameras, we propose a method that identifies the body parts between the different cameras, and tracks the motion of the targeted person among multiple people.

Much research has been performed for multiple people tracking. Research into human tracking based on multiple asynchronous camera images employing the Kalman Filter is being carried out [1]. Since the candidate region of a person is corresponded to by only one point, it becomes more complicated to identify multiple people who are occluded by each other. The method of object tracking by the model of ellipse [2] and range data [3], the method of the cooperative multi-target tracking by active vision [4], and the method of the object detection by components like a human [5] were researched. The identification of the body parts is complicated by the angle of the cameras and human position in such research. In order to track multiple people and recognize their behavior, it is important to identify the body parts captured in images even when they are occluded by other people.

On the other hand, the accuracy of the gesture recognition was influenced by the camera arrangement and photography conditions, because the gesture recognition uses registered gesture images in advance in this research. If the image does not contain the whole body, gesture recognition is difficult. Omnidirectional vision [6, 7] covers the whole field of vision, but does not solve the occlusion problem. Therefore, the use of cameras at multiple viewpoints addresses the occlusion problem, and makes it possible to identify the body parts in detail.

The estimation of the position is possible by the silhouettes method [8, 9, 10], when each camera's parameters are known. Identification of the same person is easily possible by using the silhouettes method in the different images. In this paper, we propose a method for identification of the body parts in multiple cameras for gesture recognition [11]. However, the silhouettes method suffers from the noise problem. Identification of the same person among the multiple cameras is based on the silhouettes method from the background subtraction, and identification of the motion among the multiple cameras is realized from frame subtraction. The noise problem is solved by the research of the method that restricts the voting space by the motion estimation using the Kalman Filter [12]. This research has good simulation results, but the method has high processing costs and the issue of dealing with rapid change of the motion remains. Our method is realized by the silhouettes method using the existence probability map. The existence probability map can delete noise, without movements of the body parts as rising hand. In addition, it can extinguish between people standing closely together. In this paper, we make a report of the experimental results and the validity of the proposed method.

2 Multiple People Tracking by a Silhouettes Method using the Existence Probability Map

We use the silhouettes method for multiple people tracking and motion estimation. The silhouette is extracted from multiple cameras images by the background image subtraction, which is then binarized. The silhouettes are labeled as candidate regions of a person. The voxels are voted by the silhouettes method from candidate regions of the multiple cameras images in the 3-dimensional voting space [10].

*Address: 4-1-7, Kagano, Ogaki City, Gifu 503-8569, Japan. E-mail: tomy@hoip.jp

†Address: 4-1-7, Kagano, Ogaki City, Gifu 503-8569, Japan. E-mail: hongo@hoip.jp

‡Address: 101 Tokodate, Kaizu-cho, Toyota City, Aichi 470-0393, Japan. E-mail: hiroyasu@sccs.chukyo-u.ac.jp

§Address: 4-1-7, Kagano, Ogaki City, Gifu 503-8569, Japan. E-mail: niwa@softopia.pref.gifu.jp

¶Address: 1-1 Yanagido, Gifu City, Gifu 503-1193, Japan. E-mail: yamamoto@info.gifu-u.ac.jp

A silhouettes method is utilized to reconstruct a 3-dimensional shape model from the set of 2-dimensional silhouette images [8, 9]. The precision is very important of reconstruction of the 3-dimensional shape model. But, the precision of the shape is not necessarily important for our objective. The range of the human position extraction and processing speed are important in our study. Therefore, the resolution of the voting space V (SIZE : $K \times L \times M$) is coarse, and the range of the human position defined as the region of the sum from a multiple cameras images. The voxels obtained by the voting have some problems in this case. The voxels of the noise are extracted and the voxels of some persons are combined by the occlusions in the case of multiple people.

These problems are solved by the existence probability map. The existence probability map consists of the position probability map and height cumulus map. These maps are used for deletion of noise and separation of a persons' voxels. The position probability map is used as horizontal resolution, and the height cumulus map is used as vertical resolution.

2.1 The Separation of the Persons by the Position Probability Map

The position probability map $PM(x, y)$ is defined by

$$PM(x,y) = \max_{1 \le k \le N} P_k(x,y) \times 2 - \sum_{i=1}^{N} P_i(x,y) \quad (1)$$

of the position (x, y) at the voting space for the horizontal resolution. Here, the $P_i(x, y)(i = 1, 2 \cdots N$:the number of people) is the probability for person i at the position (x, y). The probability of a near position from the previous frame is high, and a far position is low in the case of a short enough sampling interval to allow the movement of a person. A persons' position probability is controlled by the other person's position probability in the case of the approach of some persons.

The model of the resolution is defined as continuous decrease from the position of the previous frame. Figure 1(a) shows the position probability in the case of two persons, and Figure 1(b) shows the transformation of the position probability by the movement.

In the coarse resolution of the voting space, the moment speed and acceleration is not stable. Therefore, the position (fx, fy) of the next(t) frame is estimated by

$$(dx, dy) = (fx + v_x t + \frac{1}{2} a_x t^2, fy + v_y t + \frac{1}{2} a_y t^2) \quad (2)$$

from the average speed (v_x, v_y) of the 3 past frames and average acceleration (a_x, a_y) for a stable detection. The range of the position probability map is not changed, only the probability is transformed. The position probability is the decrease from these estimate positions. Figure 2 shows the example of the position probability map.

2.2 The Noise Deletion of the Voting Space by the Height Cumulus Map

The height cumulus map $HM(x, y)$ is defined as

$$HM(x,y) = \sum_{j=1}^{M} j \quad s.t.\ V(x,y,j) \neq 0 \quad (3)$$

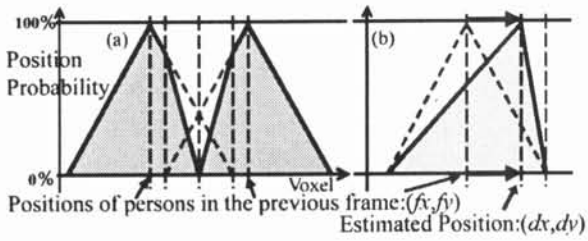


Figure 1: The position probability by two persons.

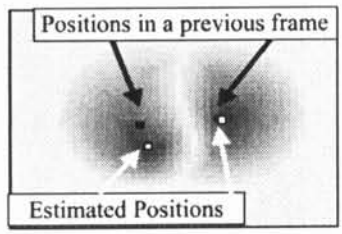


Figure 2: Position probability map.

from the high of the voting voxel $V(x, y, z)|_z$ at the position (x, y) in the voting space V (SIZE : $K \times L \times M$) in Figure 3. This height cumulus map is defined by consideration of the upper half of the body and the person's shadow. The voxels of the person with important body parts (arm, head, etc.) can delete a noise (include shadow) without deleting. Figure 4 shows the example of the height cumulus map.

The existence probability map $EM(x, y)$ is defined by

$$EM(x,y) = PM(x,y) \times HM(x,y) \quad (4)$$

from the position probability map PM and the height cumulus map HM. The deletion of the noise and the separation of the voxels are possible by deletion of voxels of less than threshold th by

$$\sum_{i=1}^{K} \sum_{j=1}^{L} \sum_{k=1}^{M} V(i,j,k) = 0 \quad s.t.\ EM(i,j) \le th. \quad (5)$$

The deletion of the noise is possible by deletion of voxels of less than threshold.

Figure 5(a), (b) shows the noise due to occlusion and result of segmentation. Figure 5(a) shows the voted voxels by the silhouettes method in the case of multiple people. This voting space has the noise voxels, and the voxels of some persons are combined by occlusion. Figure 5(b) shows the results of the noise deletion and separation of the persons in the voting space by the existence probability map. The block that is lump of voxsels corresponding to three persons has been detected in the voting space.

The voxel obtained by the voting is labelled and classified as the human block. The centre of gravity of the nearest block to the previous frame position is extracted as the position of the same person in the room.

3 Identifying Body Parts among the Multiple Cameras Images

A suitable image and region of the person is required for recognition of person. Therefore, the

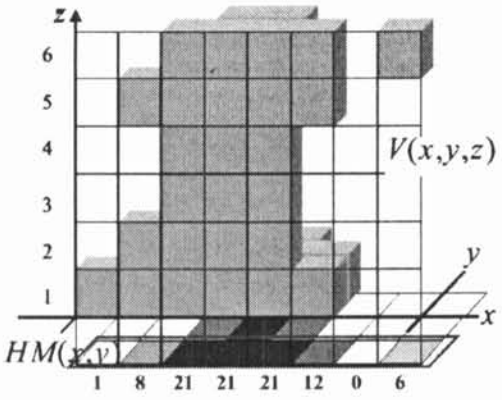
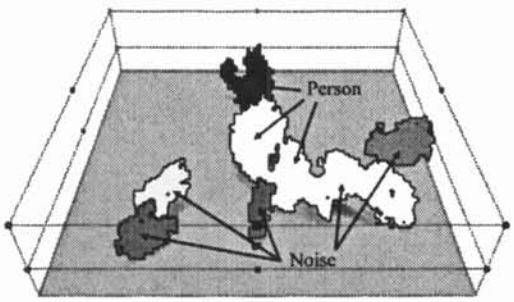
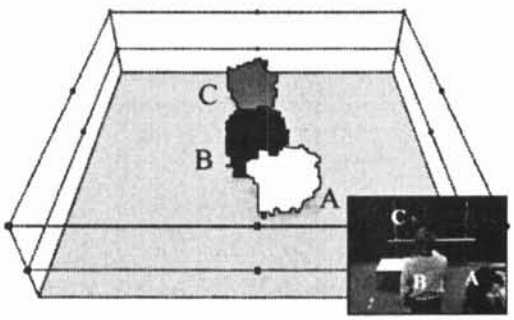


Figure 3: The cumulus by the height of the blocks.

Figure 4: Height cumulus map.



(a) Without existence probability map.



(b) With existence probability map.

Figure 5: Noise due to occlusion and result of segmentation.

correspondence as the same person between human block of the voting space and region of the image are required. And, identification of the body parts as being those of the same person among the multiple cameras images is required, too. Identification is defined by the label of the shortest distance block from each camera to persons' candidate region in the image. The lines given in

$$\frac{x - x_\alpha}{x_\beta - x_\alpha} = \frac{y - y_\alpha}{y_\beta - y_\alpha} = \frac{z - z_\alpha}{z_\beta - z_\alpha} \tag{6}$$

are generated from the camera $(x_\beta, y_\beta, z_\beta)(\beta = 1, 2, \cdots$: the number of camera) to the centre $(x_\alpha, y_\alpha, z_\alpha)(\alpha = a, b, c, \cdots$: the number of region) of the candidate region of the human as shown in Figure 6. The label γ for the candidate region is defined by the block $(x_\gamma, y_\gamma, z_\gamma)(\gamma = A, B, \cdots$: the label of block) neighbouring to the camera on the line.

$$Region\_Label = \gamma$$

$$s.t. \quad \min((x_\beta - x_\gamma)^2 + (y_\beta - y_\gamma)^2 + (z_\beta - z_\gamma)^2) \tag{7}$$

Here, the candidate region including some persons is not identified by only the centre of the candidate region of human in the images. Search is required for identification to all pixels (or some pixels) in the candidate region .

4 Identifying Motion block among the Multiple Cameras Images

Gesture recognition is possible by estimating the 3-dimensional position of the motion and extracting the target persons' motion region.

The motion block is extracted for the 3-dimensional position and motion estimation in the voting space by the silhouettes method from the frame subtraction in the region of the same person, determined by the human block. The label of the motion block is defined as the same label from region and human block. Figure 7 shows the extracted motion blocks. The gray of the floor indicates the existence probability map in Figure 7.

The gesture events are detected by comparing the human block and the motion block in the 3-dimensional space. The motion of the person is estimated by comparing the motion block MB with the human block HB in the same position. If the motion block was not detected, the person is in the state of standstill. The motion of the arm is estimated by the

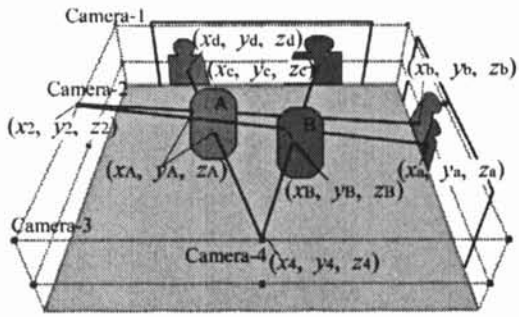


Figure 6: Identification of the region of the same person by the blocks.

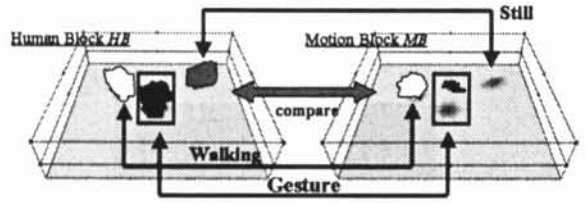


Figure 7: The human blocks and the motion blocks.

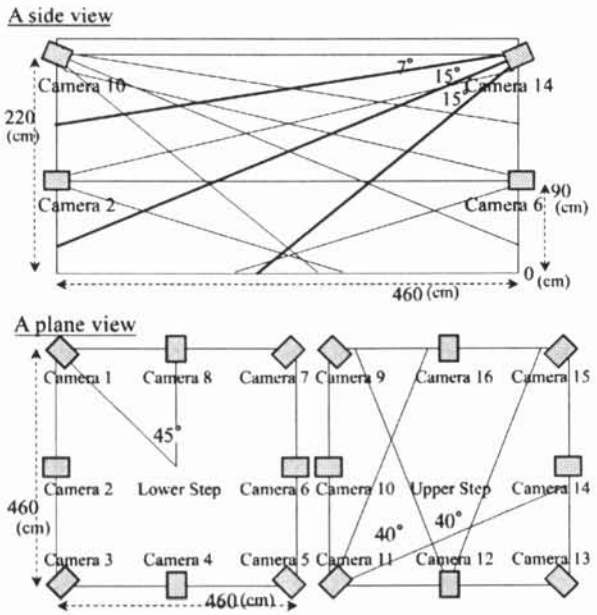


Figure 8: Camera configuration in the experimental room.

detection of the motion block in the air shown in the black rectangle in Figure 7. Using our method, the movement of a certain person can be analyzed. The gesture estimation is possible by determining a suitable judgment parameter of the human block and motion block from experimental results.

The identification of the motion region in multiple cameras images is also needed for view based gesture recognition. For example, one camera has the face image only, and other camera has the hand only. We can recognize which person is showing what kind of hand sign by integrating the result from two cameras.

5 Experiments

Figure 8 shows the camera configuration in the experimental room for multiple people tracking and motion detection. The room is square, with sides of 460 cm. This room has 16 cameras on it's walls. 8 cameras are installed at a height of 220 cm from the floor with a downward angle of 22°, and 8 more cameras are installed at a height of 90 cm from the floor level, with a horizontal view. The cameras are spaced by the intervals of regular 45° relative to the centre of the room.

Each camera is connected to a PC, and that PC acquires the VGA color images. One Main-PC aquires the VGA (640 × 480) color image that is combined

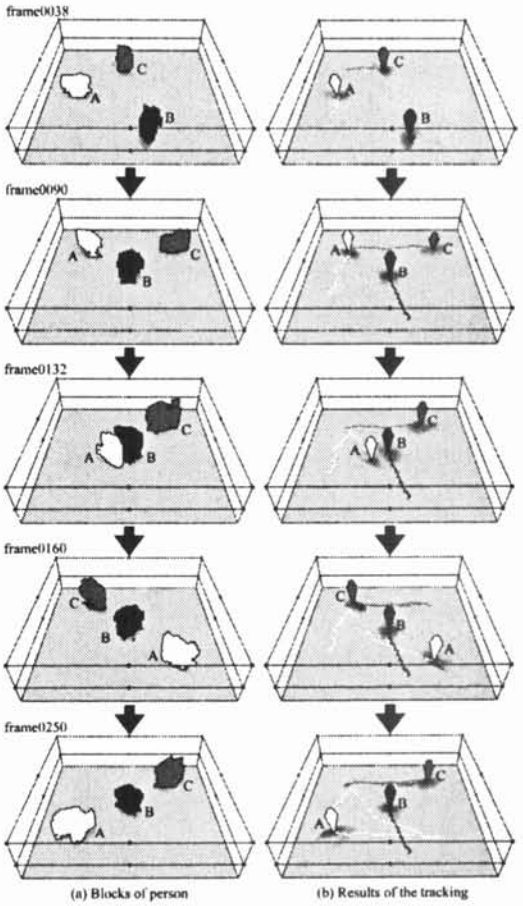


Figure 9: Results of the extracted blocks and the tracking.

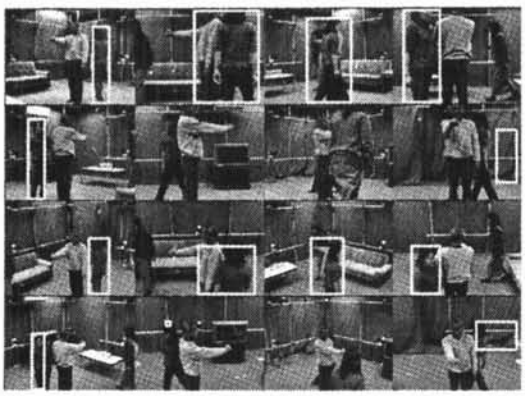
from 16 images by the switcher. The combined image is used for experiments in this proposed method of identification. The size of experimental voting space $(K \times L \times M)$ is $(92, 92, 22)$. The horizontal voxel resolution is 5 cm, and the vertical voxel resolution is 10 cm. The range of the existence probability map is defined 12 voxels (60 cm) for the human walking speed (1 m / sec) and human volume (40 cm). The human is usually need 100 msec (1.5 frame) for the movement of the 12 voxels range.

5.1 Results of the Multi People Tracking and Identification

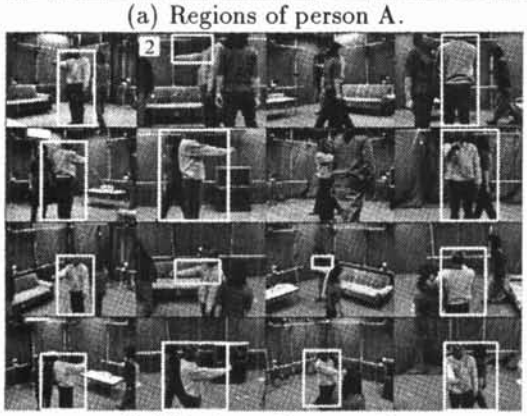
Figure 9 shows the results of the multi people tracking for 250 frames (16.7 sec). The column of Figure 9 (a) shows the results of the extracted blocks for three persons. The column of Figure 9 (b) shows the locus of the block's centre of gravity. The label shows the same person in these images. Person A moved in a triangular fashion around the room. Person B stopped in the centre of the room. Person C moved side-to-side. Figures 10 (a), (b) and (c) show the results of the identification of the body parts of the same person. The white rectangle region is the result of identification in this image. Although person B is obstructed by person A in camera 2 of Figure 10 (b). The arm is identifyed as that of person B.

Figure 11 shows the results of the identification in

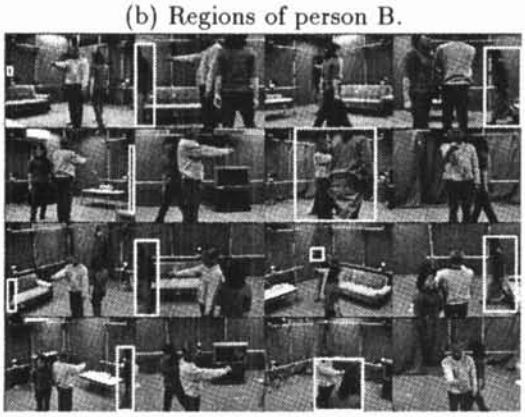
the case of five persons. Figure 11 (a) shows the results of the tracking, and Figure 11 (b) shows a region of the person D. The hand shape recognition of person D is difficult in the image of the camera 2. But, it is easy by the identification of the person D in other cameras.



(a) Regions of person A.



(b) Regions of person B.



(c) Regions of person C.

Figure 10: Results of the identification of the same person.

5.2 Results of the Motion Detection and Identification

We demonstrate the experiments of the motion detection for hand gesture. The hand gesture is defined as the motion of the rising hand only. The motion of raising the hand while the body is in motion (i.e. walking) is not included in the hand gesture.

Detection of the hand gesture event is possible from the position of the motion block MB, because the height of the top of the human block HB_h is defined as the stature of person. The range of the motion of the hand is defined as the upper half of the body except the head. Therefore, the motion of the hand gesture is detected by

$$HB_h/2 \leq MB_h \leq HB_h \times 7/8 \quad (8)$$

$$HB_h/2 \leq MB_l \leq HB_h \times 7/8 \quad (9)$$

when the human size is divided into eight. Momentary detection of a motion block is not a hand gesture event, that is noise by a little motion. The hand gesture event is detected by the detected motion with the continuous frame. The hand gesture start is defined by the continuous increase of the height of the centre of gravity of motion block $MB_g(x,y,z)|_z$, and the hand gesture end is defined by the continuous decrease.

Figure 12 shows the change in the height of the centre of gravity of the motion block $MB_g(x,y,z)|_z$. The hand gesture start is detected from frame 109 to frame 117, because the motion block shows a continuous increase. And, the hand gesture end is detected from frame 160 to 169, because the motion block shows a continuous decrease. Figure 13 shows the motion block of the hand gesture start and motion region of the camera 2. The motion block of the arm is detected in the air of the voting space. The identification of the region as the same motion is shown by the white rectangle in Figure 14. The region of the back is identified as the same motion region by the relative motion of the clothes in connection with the motion of the hand. The region for the gesture recognition is defined by the motion region at the hand gesture start, because the hand is stopping at the time of gesture.

6 Conclusion

In this paper, we proposed a method for the identification of the body parts of multiple people by silhouettes method using existence probability map from background image subtraction and frame subtraction in multiple camera images. We demonstrated the experimental results and the validity of the proposed method.

The processing time is average 1.5sec / frame for 3 persons on a Pentium 4 1.7GHz. Hereafter, prediction of the movement is required in the existence probability map for improvement of processing time and precision.

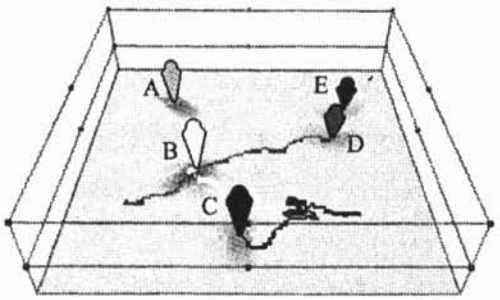
References

[1] Akira Utsumi, Hiroki Mori, Jun Ohya and Masahiko Yachida : "Multiple-Human Tracking using Multiple Cameras", Proc. of the Third International Conference on Automatic Face And Gesture Recognition, pp.498-503 (Apr.1998).

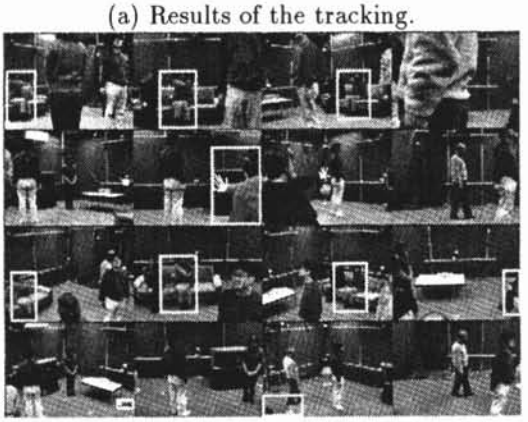
[2] Kiyotake Yachi, Toshikazu Wada and Takashi Matsuyama : "Human Head Tracking using Adaptive Appearance Models with a Fixed-Viewpoint Pan-Tilt-Zoom Camera", Proc. of the Fourth International Conference on Automatic Face And Gesture Recognition, pp.150-155 (Mar.2000).

[3] Leonid Taycher and Trevor Darrell : "Range Segmentation Using Visibility Constraints", Proc. IEEE Workshop on Stereo and Multi-Baseline Vision (Dec. 2001).

[4] Norimichi Ukita and Takashi Matsuyama : "Real-Time Cooperative Multi-Target Tracking by Communicating Active Vision Agents", Proc.



(a) Results of the tracking.



(b) Region of person D.

Figure 11: Results of the Identification in the case of 5 persons.

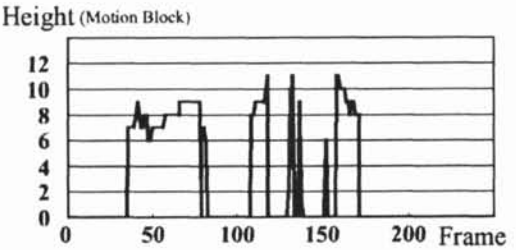
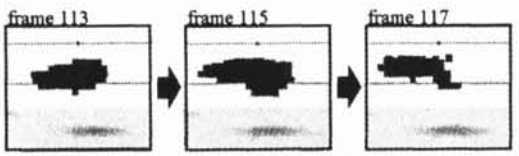
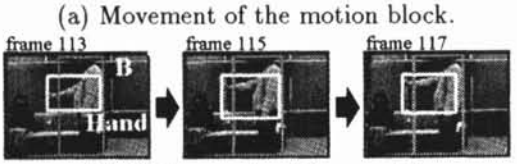


Figure 12: Transition of the height of the motion block (Person B).



(a) Movement of the motion block.



(b) Motion region (Camera 2).

Figure 13: Movement of the motion block and region (Person B).

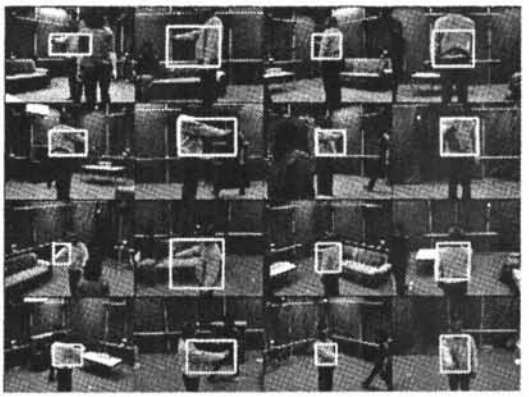


Figure 14: Result of the Identification of motion regions (Person B).

IAPR 16th International Conference on Pattern Recognition, Vol.I, pp.14-19 (Aug.2002).

[5] Anuj Mohan, Constantine Papageorgiou and Tomaso Poggio : "Example-Based Object Detection in Images by Components", IEEE transaction on PAMI, Vol.23, No.4 (Apr. 2001).

[6] Hiroshi Ishiguro : "Development of low-cost compact omnidirectional vision sensors and their applications", International Conference on Information Systems, Analysis and Synthesis, pp.433-439, (Jul.1998).

[7] Takuichi Nishimura, Takuya Sogo, Ryuichi Oka and Hiroshi Ishiguro : "Recognition of human motion behaviors using multiple omni-directional vision sensors", Proc. IEEE International Conference on Industtrial Electronics, Control and Instrumentation, pp.2553-2558 (Oct.2000).

[8] Toshikazu Wada, Xiaojun Wu, Shogo Tokai and Takashi Matsuyama : "Homography Based Parallel Volume Intersection : Toward Real-Time Volume Reconstruction Using Active Cameras", Proc. of Computer Architectures for Machine Perception, pp.331-339 (Sep.2000).

[9] Takashi Matsuyama and Takeshi Takai : "Generation, Visualization, and Editing of 3D Video", Proc. of symposium on 3D Data Processing Visualization and Transmission, pp.234-245 (Jun.2002).

[10] Masafumi Tominaga, Hitoshi Hongo, Hiroyasu Koshimizu, Yoshinori Niwa and Kazuhiko Yamamoto : "Estimation of Human Motion from Multiple Cameras for Gesture Recognition", Proc. IAPR 16th International Conference on Pattern Recognition, Vol.I, pp.401-404 (Aug.2002).

[11] Mamoru Yasumoto, Hitoshi Hongo, Hiroki Watanabe, Kazuhiko Yamamoto : "Face Direction Estimation and Face Recognition Using Multiple Cameras for Communication in a Virtual Enviroment", Proc. IEEE International Conference on Industtrial Electronics, Control and Instrumentation, pp.295-300 (Oct.2000).

[12] Mitsuharu Hayasaka, Hideyoshi Tominaga and Kazumi Komiya : "Multiple object tracking using back projection method and kalman filter", IEICE Technical Report, PRMU2001-132, pp.133-138 (Nov.2001) (in Japanese).