Binocular Omni-Stereo Based Human Tracking System in Indoor Environments for Intelligent Surveillance

Hong LIU[2], Wenkai PI, Hongbin ZHA
National Lab on Machine Perception
Peking University, [2]*Shenzhen Graduate School*
{liuhong, piwenkai, zha}@cis.pku.edu.cn

Abstract —**This paper proposes a new real-time human tracking system in indoor environments by using a binocular omni-stereo. Two catadioptric ODVSs (Omnidirectional Vision Sensors) are employed in our system to construct a binocular stereo, for the advantage of obtaining images with 360° view for global scenes. A method for adaptive background subtraction and shadow elimination is utilized to precisely segment the moving human body. Then, based on the analysis of different methods for human tracking and the special problems in binocular Omni-stereo, a novel method of features correspondence and a tracking strategy on the baseline are presented to determine the location of human body. Experiments show that the proposed system performs fast and robustly for tracing one person in indoor, complex environments.**

Key words - real-time;omni-stereo;human tracking

I. INTRODUCTION

Computer vision systems for human tracking are widely used in many applications, such as smart surveillance, virtual reality and motion analysis [1]. Single person tracking systems stand an important role in many applications, such as distant education, smart housing, nurse robot, intelligent cameras for tracking a specified athlete and so on. On the other hand, tracking of single person is the foundation of tracking multiple persons in smart surveillance.

There are three important issues that should be considered in a tracking system: Field Of View (FOV), computational complexity, and robustness. Most of conventional vision systems use single fixed camera to take image sequences of a scene, e.g. the "Pfinder" of Wren et al. [2], and "W^4" [3]. One disadvantage of these systems is that they have relatively narrow surveillance fields due to the limitation of their cameras' FOV. Tracking may fail when persons go beyond the boundary of the FOV.

Using multiple ordinary cameras located in different areas can partly enlarge the FOV of the system, such as Mittal and Davis's "M_2 Tracker" [4]. When a person disappears from the capture of one camera, another one can continuously track it. However, the images of the global scene can hardly be obtained. Furthermore, establishing feature correspondence among different cameras is a rather difficult problem.

Also many researchers rotate cameras to get the panoramic view. Ishiguro et al. [5] pan a camera at different locations in a room, thus construct the global map. Krishnan et al. get panoramic range data from focus by rotating their "NICAM" [6]. These systems can produce high-resolution data for static scenes. But considering their time-consuming, they can hardly be used in real-time human tracking systems.

In recent years, catadioptric ODVSs have been applied to many applications, for the reason of providing a 360° view angle of the environment in single image. With the advantages of panoramic, compact visual information and directive features, using the ODVS for human tracking is of great promising.

Nara Institute in Japan describes a human surveillance system using single ODVS [7]. It only performs well in simplex backgrounds, and cannot measure the locations of targets. The "LOTS" provided by VAST Lab in Lehigh University can adapt to the dynamic background [8]. However, since the targets being surveiled by the system must be very small and distant, the system is hard to be used in indoor environments.

Two hyperbolical ODVSs [9] are employed in our system to construct the omni-stereo. The two ODVSs are setup horizontally at different locations in our lab, to surveil one human body and determine its moving trajectory. Ishiguro use more than 4 ODVSs to surveil multiple human bodies [10], and our system only use two ODVSs for one person tracking. Furthermore, our system has two advantages: a) a method for adaptive background subtraction and shadow elimination is used to make the system work well even in dynamic and complex environments; b) a novel method of human feature correspondence is presented to determine more precise location of human body, even when the person moves on the baseline of two ODVSs.

The remainder of this paper is organized as follows: Section II describes the construction of our omni-stereo. A method for adaptive background subtraction and shadow elimination for obtaining the moving regions is described in section III. Section IV presents human tracking and trajectory estimation method using our omni-stereo. Experiments and conclusions are given in section V and section VI, respectively.

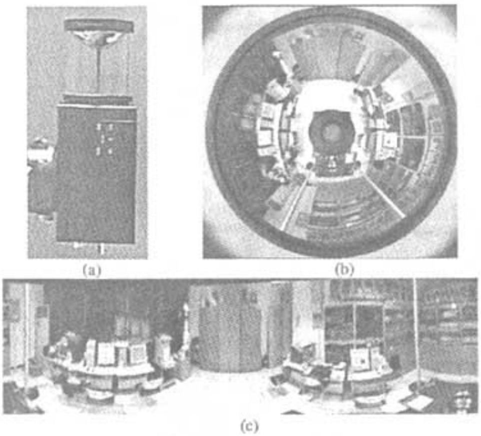


Fig 1 ODVS and omnidirectional images
(a) An ODVS (b)Omnidirectional image (c) Cylindrical panoramic image

II. BINOCULAR OMNI-STEREO

Each ODVS is composed of a hyperbolical mirror and a CCD camera, as illustrated in Fig.1 (a). Basically, there are two ways to construct omni-stereo. Many researchers setup two ODVSs together on a mobile robot, to reconstruct geometrical information of environment [11], or to realize robot self-localization [12]. And the omni-stereo constructed in our system is in another way. Two ODVSs are localized in different places in our lab, and the height and orientation of each one is precisely measured. By using the binocular principle, each person's location can be obtained. Measurement precision of target locations and baseline region between two ODVSs are two problems should be considered.

The former problem occurs when the target widely projected on the ODVSs. However, in our system, since the precise human body contour is obtained from previous steps, we can use some reliable feature points to correspond them by two ODVSs. Thus, the location of human body can be determined.

The latter problem is that when the target locates along the baseline of two sensors, the target location cannot be determined using the triangular principle. This problem does not occur in a conventional stereo, but in the omni-stereo, it cannot be avoided. In this case, another strategy using the right and the left boundaries of the target is presented to determine the target locations.

III. MOTION DETECTION FOR HUMAN BODIES

The omni-stereo is located in a laboratory room, which may contain some dynamic factors such as floating of curtains, blinking of computer screens, and variation of illuminations. Also, when a person is moving in the room, its mirror images maybe appear in the glass windows, and its shadows maybe appear on the ground. These problems make it difficult to obtain the accurate foreground region by simply subtracting a new frame from a settled background. Here, a method for adaptive background subtraction and shadow elimination [13, 14] is introduced in our system to establish a robust foreground segmentation mask.

A. Fast Recovery of Panoramic Images

Since the image directly acquired by the ODVS is naturally distorted, it needs to be transformed into the cylindrical panoramic image for target analysis, as shown in Fig.1 (b) and (c).

If the parameters of the hyperbolical mirror and the CCD camera's focus are known, each pixel in the cylindrical panoramic image can be exactly calculated from the circular omnidirectional image [7]. However, since the computation is much time-consuming, a fast and effective method is presented for the transformation in our previous work [15].

B. Adaptive Background Subtraction

Every point in the background is assumed to have a mean color value and a distribution about that value. Before any person entering the environment, the ODVS observes the scene for several seconds, and then the initial background model can be built up as follows. μ_i is defined as the mean color value of a point *i*, and σ_i^2 as the covariance of that point's distribution. Thus, (μ_i, σ_i^2) can be stored as the color background model for the point *i*. Since a color pixel has three components of R, G and B, μ_i and σ_i^2 are defined as vectors:

$$\mu_i = (\mu_i(r), \mu_i(g), \mu_i(b)) \tag{1}$$

$$\sigma_i^2 = (\sigma_i^2(r), \sigma_i^2(g), \sigma_i^2(b)) \tag{2}$$

The initial background model cannot be expected suitable for a long period due to the variations of the scene. For each new frame *t*, $y_i(t)$ is the current color of pixel *i*. The background model is updated on-line using the following formulas:

$$\mu_i(t+1) = \begin{cases} (1-\alpha)\mu_i(t) + \alpha y_i(t+1), & \text{If } i \text{ in background} \\ \mu_i(t) \ .\ . & \text{If } i \text{ in foreground} \end{cases} \tag{3}$$

$$\sigma_i^2(t+1) = \begin{cases} (1-\alpha)\sigma_i^2(t) + \alpha(y_i(t+1) - \mu_i(t+1))^2 \\ & \text{If } i \text{ in background} \end{cases} \tag{4}$$

If *i* in foreground$\sigma_i^2(t)$,

Here, the constant α $(0 < \alpha < 1)$ controls the adaptation rate.

C. Foreground Region Detection

Once the adaptive background model is obtained, each new frame can be subtracted from it to determine the foreground regions. The current pixel $y_i(t) = (r, g, b)$ is compared with the model. If $(y_i(t) - \mu_i > 3\sigma_i)$, the pixel can be regarded as foreground. Otherwise, it is regarded as background. Thus a mask is established, which is considered as a region of interest for further processing.

The method presented before can adapt to dynamic illuminative factors in the scene, but it cannot eliminate the shadows caused by people moving in the scene.

The shadow elimination method is based on the factor that an area cast into shadow often results in significant changes in intensity without much variation in chromaticity. Chromaticity can be computed as follows:

$$rc = r / (r + g + b) \quad (5)$$

$$gc = g / (r + g + b) \quad (6)$$

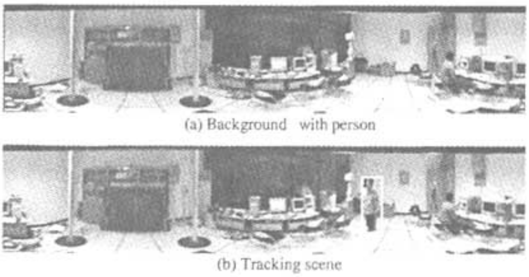
And each pixel's chromaticity is modelled using means and variances $(\mu_{rc}, \mu_{gc}, \sigma_{rc}^2, \sigma_{gc}^2)$. Adaptive background subtraction is performed again, but using chromaticity values for this time. Finally, a pixel is marked as foreground if both RGB and chromaticity information support that classification.

D. Human Body Extraction

The binary images, obtained directly from subtraction usually contain isolated points or lines caused by those dynamic factors. Morphological filters are used to erase them. First a 3 × 3 erode filter is used to wipe off the isolated points and lines, then a 3×3 dilation filter is used to recover the exact foreground region. Fig. 2 shows the result of detected foreground.

IV. HUMAN LOCALIZATION USING OMNI-STEREO

Two ODVSs measure the different azimuth angles of the moving human body from two directions. If the locations and the orientations of the ODVSs are known, the location of the target can be measured from the azimuth angles by triangulation principle.


(a) Background with person

(b) Tracking scene

Fig.2 Adaptive background subtraction

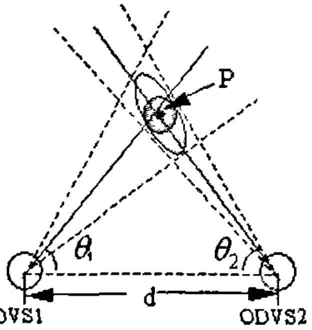


Fig.3 General algorithm for locating person

A. Measurement for Human Body

The centroid of the tracked human body can be estimated from the right and the left boundaries of human body in the images. But the boundary features are very different for body's different orientations around the ODVS. Only the position of human head can hardly be influenced in this kind of images. Therefore, a new feature of head position is combined to track the human body, which is shown in Fig.3. The dashed lines are for observing boundaries from ODVSs, which will result in much error when the human body is near to the ODVSs. The solid lines indicate the position of human head, which can be localized accurately to a great extent. Point P in Fig.3 indicates the human body's location. The two ODVSs are located at certain height to ensure that the head will be visible to both ODVSs.

B. Strategy on Baseline

Another important problem in the omni-stereo vision system with two ODVSs is the special fields of baseline. Theoretically, a point cannot be localized when it is just on the baseline. In the practical system, human body is not a point but a field with its size, although the error of location will be increased rapidly while the tracked body is just on the baseline.

When the human head cross the baseline, i.e., the azimuth of human head, θ_1 and θ_2 are bigger than a threshold angle θ_0, left and right boundaries of the human body can be still detected. Fig.4 (a)(b)(c) show the three moments of head crossing the baseline.

a) If both boundaries of the body do not cross the baseline, i.e. $\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22}$ are larger than θ_0, the position of human body P is the intersection of the baseline and the line between points of P_1 and P_2.

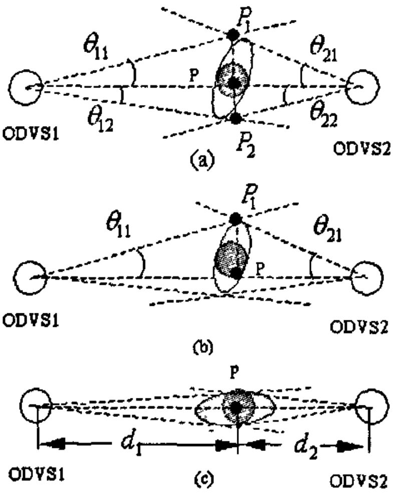


Fig.4 Strategy on baseline

b) If only one boundary of human body do not cross the baseline, i.e. θ_{11} and θ_{21} are larger than θ_0, only the position of P_1 can be got. The position of human body P can be approximately localized as the cross point of the baseline and the vertical line from P_1 to baseline.

c) If the two boundaries and the head of human body are all on the baseline, the position of head cannot be got from the above method. On this occasion, we suppose the distances from human body to two ODVSs are d_1 and d_2. And the areas projected by the body on the two ODVSs are s_1 and s_2. We find that the ratio of $d_1 : d_2$ is close to that of $s_1 : s_2$. This fact can be used to estimate the distances from the two ODVS.

The cases of the human body between the two ODVSs are described above. Also, the method is the same when the human body is on the same side of the two ODVSs.

C. Trajectory Estimation

While human wandering in the environment, its head position can be got by the location method described in section A, when the body is far away from the baseline. Head position can be got by the strategy described in section B when the body is near the baseline. Therefore, the trajectory of the moving human body can be marked as a series of points localized for each moment.

V. EXPERIMENTS

The experimental system is established in our lab room, which contains many dynamic factors. The moving target is assumed to be a human body. Two ODVSs fixed at a height of 1.5m, and a distance of 3m, are distributed in the room to construct the omni-stereo. Before tracking, the environment scene model is built by observing the whole scene without any person for several seconds. If human body is detected entering the scene, their detected region azimuth angles are obtained by each ODVS, and then its location can be estimated by using the method described in Section IV. Furthermore, the trajectory of the human body can be obtained. Omnidirectional images acquired in 24-bit RGB model are of 480×480 resolution. The system runs at average 12Hz on a Pentium IV 1.8GHz PC.

Fig. 5 shows the GUI of the tracking system. The right parts are for the two images from ODVSs, the base parts are for the results of foreground detection. The left-middle part is the bird-view of the experimental environment. The two bigger black circles are marked for the two ODVSs. The small point is the detected position of the tracked human body. Fig.6 (a) ~ (d) illustrate the head trajectories of 4 different persons in 10 testing persons (5 Men and 5 Women, indicated as M1~M5 and F1~F5, followed by their statures and ages) walking in the experimental environment. They walked along a common given path, which is pre-designed in our experiments. The points on the solid line show the head locations of each person at each frame. In Fig.6 (e), the broken lines indicate the pre-designed foot route. The errors between the two trajectories in each figure are due to not only the tracking error, but also head-foot distinctions and the differences on personal height, walking speed, gait, etc. Fig.6 (f) indicates the comparison of the 10-person average head trajectory and the pre-designed foot trajectory. Experiments show that the human body can be tracked in real-time for the complex environments, even in or near the fields of the omni-stereo's baseline.

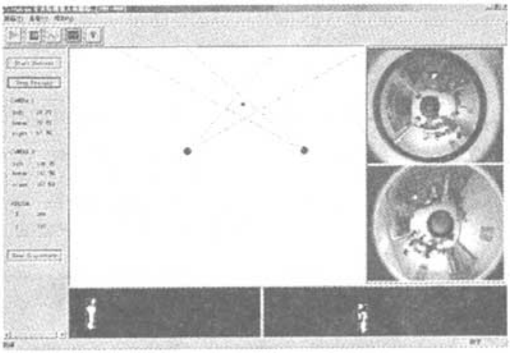


Fig.5 Graphic Interface of the tracking system

VI. CONCLUSIONS

A new real-time vision system based on omni-stereo is proposed in this paper for human tracking in indoor environments. Two ODVSs are used to obtain 360° view images of the scene to enlarge the monitored areas. A method for adaptive subtraction and shadow elimination is presented to improve the robustness of the system in dynamic environments. Moreover, a novel feature corresponding method and a tracking strategy on the baseline are presented to determine the human body location even in the field of omni-stereo baseline. Experiments show that the system can track single moving person with a large view field in a dynamic environment in real-time.

ACKNOWLEDGMENT

This work is supported by National Natural Science Foundation of China (NSFC), P. R. China.

REFERENCES

[1] D M. Gavrila, The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, vol. 73, no. 1, pp. 82-98, 1999

[2] C.R. Wren, A. Azarbayejani, T. Darrell, and A P Pentland, Pfinder: Real-time tracking of the human body *IEEE Trans on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 780-785, July 1997.

[3] I. Haritaoglu, D. Harwood, and L S. Davis, W4: Real-time surveillance of people and their activities *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.22, no. 7, pp. 809-830, 2000.

[4] A. Mittal and L.S. Davis, M_2 Tracker: A Multi-View Approach to Segmenting and Tracking People in a Cluttered Scene. *Int. J. of Computer Vision*, vol 51, no 3, pp 189-203, 2003

[5] H. Ishiguro, M Yamamoto, and S. Tsuji, Omni-directional Stereo. *IEEE Tran. on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 257-262, Feb. 1992.

[6] A. Krishnan, and N Ahuja, Range estimation from focus using a non-frontal imaging camera. *Int. J. of Computer Vision*, vol. 20, no. 3, pp. 169-185, 1996.

[7] Y. Onoe, K. Yamazawa, N. Yokoya, and H. Takemura, Visual surveillance and monitoring system using an omnidirectional video camera. *Proc. IEEE Intl. Conf on Pattern Recognition*, pp. 588-592, 1998.

[8] T.E Boult, R. Micheals, X. Gao, P. Lewis, et al. Frame-rate omnidirectional surveillance and tracking of camouflaged and occluded targets. *IEEE Workshop on Visual Surveillance*, pp. 48-55, June 1999.

[9] H Ishiguro, Development of low-cost compact omnidirectional vision sensors and their applications Intl. Conf. on Information Systems, Analysis, and Synthesis, pp 433-439, 1998.

[10] T Sogo, H Ishiguro, M M Trivedi, Real-time target localization and tracking by N-ocular stereo. IEEE Workshop on Omnidirectional Vision, pp 153-160, June 2000

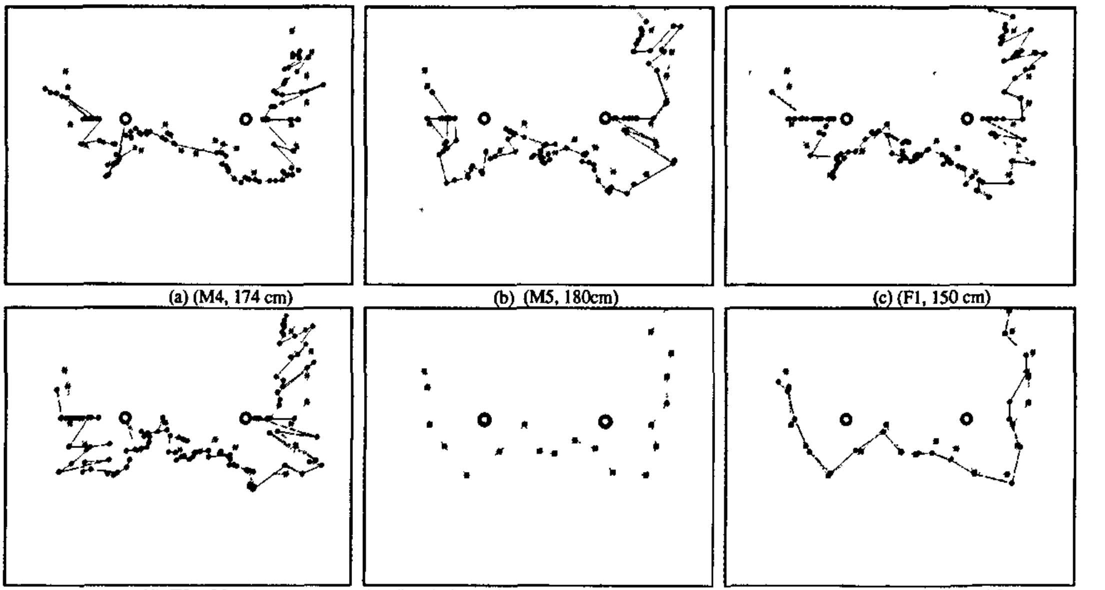
[11] J. H Kim and M. J. Chung, SLAM with Omni-directional Stereo Vision Sensor. IEEE Conf. on Intelligent Robots and Systems, vol. 1, pp. 442-447, Oct. 2003

[12] E. Jyun-ichi, T. Toshinobu, T. Jun-ichi, and H. Takumi, Self-positioning with an Omni-directional Stereo System IEEE Conf. on Robotics & Automation, vol. 1, pp 899-904, Sept 2003

[13] S. McKenna, S. Jabri, Z Duric, A Rosenfeld, and H. Wechsler, Tracking groups of people. Computer Vision and Image Understanding, vol 80, pp 42-56, 2000,

[14] C. Stauffer and W E L Grimson, Adaptive background mixture models for real-time tracking IEEE Conf on Computer Vision and Pattern Recognition, pp. 246-252, 1999

[15] H. Liu, W K. Pi, H B. Zha, Motoin Detection for Multiple Moving Targets by Using an Omnidirectional Camera IEEE Intl Conf. on Robotics, Intelligent Systems and Signal Processing, Oct. 2003.



(a) (M4, 174 cm)

(b) (M5, 180cm)

(c) (F1, 150 cm)

(d) (F2, 156 cm)

(e) Pre-designed foot route

(f) average head trajectory and the pre-designed foot trajectory

Fig. 6 Tracking Results of different human bodies