Multi-target Indoor Tracking and Recognition System with Infrared Markers for Virtual Reality

Wenhui Xu, Bo Wang, Yongshi Jiang
Institute of Automation Chinese Academy of Sciences, CASIA, Beijing, China
xuwenhui2014@ia.ac.cn, bo.wang@ia.ac.cn, yongshi.jiang@ia.ac.cn

***Abstract*—Virtual reality (VR) is a computer simulation technology, which can create a virtual world to allow users to immerse in the simulated environment, and be able to interact with objects in a nature way. This paper presents an indoor tracking and recognition system to meet the interactive of multiple users in virtual reality. As we all know, tracking in real time and accurately play a vital role in VR application. Interactive VR games require data update rate above 35HZ to make players fell well. In our system, we use infrared cameras, infrared markers and image processing techniques to acquire the users' position and orientation information in real time. We describe a relatively inexpensive, but can monitor the precise location and orientation information system. In our system, cheap infrared cameras are fixed on ceiling. Every user in the environment wears an infrared LED module. The distance between any two infrared LED in a LED module is different with others. We can distinguish every user and get their precise position and orientation in real time by stereo vision theory and our recognition algorithm. This system strikes a good balance between price and capability.**

Keywords—infrared cameras;markers;tracking;recognition

I. INTRODUCTION

Virtual reality is a comprehensive application of multiple technologies, including real-time three-dimensional computer graphics technology, wide-angle stereoscopic display technology, the head, eye and hand motion tracking technology, as well as voice input and output technologies [1], [2]. Among them, motion tracking technology is an important actor of information interaction between people and virtual environment. It is an important field of virtual reality technology development in recent years. In order to allow users to move freely in the virtual environment, and increase the flexibility of interaction, it is important to accurately monitor the users' position and orientation information in real time [3-5].

Many devices have been developed to acquire such three-dimensional position or orientation information, such as gyro orientation sensors, magnetic position sensors, accelerometer sensors, ultrasonic distance sensors, the global positioning system (GPS) and so on [2], [6]. Each sensor has its advantages and limitations. For example, gyro and accelerometer sensor can make a quick response to the user's movement, but they cannot detect the user's location and hand shape. Inertial and vision-based sensing techniques are used in paper [7] for accurate positioning and tracking in VR system. The inertial sensor apply accelerometer and gyroscope to measure rate and acceleration. The position and orientation are computed by vision-based sensing using camera and image processing techniques. The algorithm is efficient to achieve a quick response to the user's motion including spatial position and orientation. However, it is not suitable for precisely tracking in a large indoor environment and interacting in a multi-user environment. Another method for acquiring the users position and orientation is using infrared markers [8-10]. Weng D, Liu Y, Wang Y study an indoor tracking system based on primary and assistant infrared markers[10]. The system can track the user in a large workspace with high stability and accuracy.

In this paper, we describe a new indoor multi-user tracking system, which aims to improve tracking accuracy and stability in real time. In this system, we do not need to fix a lot of infrared markers on the wall, but put a few infrared markers on the head of users. The rest of devices are several infrared cameras fixed to the ceiling to monitor the mobile infrared markers. By processing with our algorithm, we can not only get the users' location and orientation, but also identify different users in a multi-user environment.

The paper is organized as follows. A multi-target indoor tracking and recognition system is described in section II . Section III describes the algorithm used in the system. In section IV, experimental results is shown by figures and tables. Finally, section V gives the conclusion and future work.

II. SYSTEM OVERVIEW

In our system, three or more infrared cameras are mounted on the ceiling. The larger the working areas are, the more infrared cameras are required. Each user wears a helmet-mounted display (HMD) and an infrared light emitting diode(LED) circuit board (LED module). HMD is connected to a laptop computer in a backpack worn by the user. When users moving in the room, these cameras detect the position information of the infrared LED markers on the users head, and transport to the desktop computer. After that, the computer calculate the spatial position of LED markers according to the theory of stereo vision. According to the distance between the infrared LED markers, the desktop computer can identify different infrared LED modules. Then we can acquire the location and orientation of different users, which can be expressed as 6 degree-of-freedom(6DOF), three translational variables and three rotational variables. The information can be delivered to laptop computer via wireless network. After that, the laptop computer can render the characters in virtual scene according to the position and orientation information. After transferring virtual scene to HMD, different users can see different virtual

scenes according to their position and orientation. Fig. 1 presents the overview of our system.

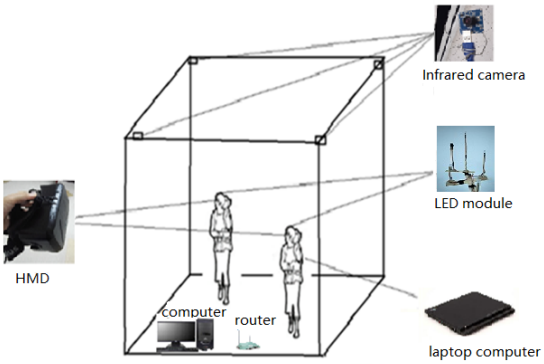

Fig. 1: System configuration

III. Measure users' position and orientation

A. Tracking Of The Infrared Markers

In this system, we use stereo vision theory to acquire infrared markers' position. To perform epipolar constraint and reconstruct three-dimensional coordinates of makers, we need to know the intrinsic parameters and the extrinsic parameters of cameras [11]. We use Zhang's approaches described in [12], [13] to precisely and fully calibrate cameras. Since we use multiple infrared filter cameras in this system to monitor working area, it is difficult to directly obtain all cameras' extrinsic in the common world coordinate system. To settle this issue, we can calibrate two of the cameras firstly, to get the external parameters of the two cameras in the common world coordinates. Then we calibrate one of the two cameras(the first camera) and another camera(the third camera) to acquire the conversion relationship between them. According to the conversion relationship and the first camera's external parameter, we can get the third camera's external parameter in the common world coordinates. In this way, we can obtain all cameras' extrinsic in the common world coordinates.

Since the infrared cameras are only sensitive to infrared light, the input images show dark background with some bright spots. It is sufficient to process the input images by background clipping and thresholding. Then we detect contour edge of the bright spots and calculate the center of contour gravity to obtain pixel coordinates of the infrared markers. Next we match the pixel coordinates of the same marker in different images to get the correct world coordinates by three-dimensional reconstruction. When the intrinsic parameters and the extrinsic parameters of cameras are calibrated, it is easily to deduce the epipolar constraint by formula 1.

$$q_r^t * F * q_l = 0; \tag{1}$$

In this equation, q_r and q_l are two corresponding points from left and right camera images respectively. The fundamental matrix F is the algebraic representation of the epipolar geometry [14].

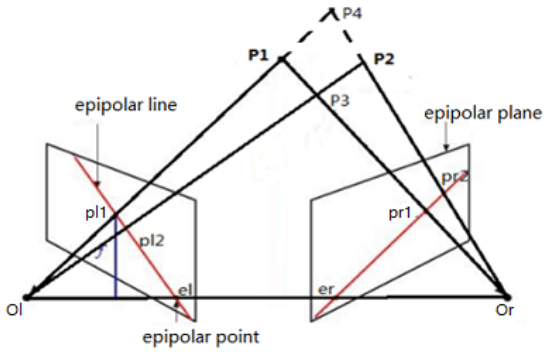

Fig. 2: Error match when epipolar constraint

However, under some circumstances, it could be possible that several points lie on the same epipolar line. As shown in Fig. 2, when the line that markers p_1 and p_2 lie on is parallel to the line that the cameras' optical center o_l and o_r are located at, the projected points of the two markers in the same projection plane lie on the same epipolar line. It is possible to produce false matches in this case, and get the wrong points p_3 or p_4. A solution for this issue is projecting all the points onto the third cameras projection plane, saving the points that find a match point in the third projection plane. The processing of matching the pixel coordinates of the same marker in different camera images is shown in table I.

TABLE I: MATCHING THE PIXEL COORDINATES OF THE SAME MARKER

Step1:

- Calculate the foundation matrix F according to the two cameras' intrinsic parameters and extrinsic parameters.

Step 2:

- For each point in the first image, find the corresponding point in the second image. The detail processing is as follows:
  1) For a point in the first image, calculate the distance vector of every point in the second image to it according to formula 1.
  2) Find the distance in the threshold and the corresponding point in the second image.
  3) If there is only one distance in the threshold, find the corresponding point in the second image to match the point in the first image. Recording the position of the found point and go back to 1) to process the next point in the first image.
  4) If there are two distance or more in the threshold, we need to calculate some world coordinates of the market according to the point in the first image and the points in the second image whose distance in the threshold. Then we project those world coordinates to a third camera image. Select the world coordinates whose project point can find the nearest point in the third camera image. Recording the position of the corresponding point in the second image and go back to 1) to process the next point in the first image.

Step 3:

- Sometimes different point in the first image is likely to match the same point in the second image, such as the number of the points in the second image is less than the first image. So, it is necessary to clear the matching overlap points. We can respectively calculate the epipolar constraint of the overlap point in the second image and different point in the first image, select the point whose epipolar constraint distance is nearest.

Once we find the corresponding points in different cameras image, we can obtain the world coordinates of each marker by

3D reconstruction described in [15]. The used formulas can be described as follows.

$$Z_{cn}\begin{pmatrix} u_n \\ v_n \\ 1 \end{pmatrix} = \begin{pmatrix} m_{11}^n & m_{12}^n & m_{13}^n & m_{14}^n \\ m_{21}^n & m_{22}^n & m_{23}^n & m_{24}^n \\ m_{31}^n & m_{32}^n & m_{33}^n & m_{34}^n \end{pmatrix} * \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} \quad (2)$$

The formula 2 describe the theory that the world point $X = (x, y, z)$ project to camera n to get pixel point (u_n, v_n). Combining all cameras' projecting formulas and dissolving the variable of Z_{cn}, we can acquire the formula 3.

$$KX = U; \quad (3)$$

$$K = \begin{pmatrix} u_1 m_{31}^1 - m_{11}^1 & u_1 m_{32}^1 - m_{12}^1 & u_1 m_{33}^1 - m_{13}^1 \\ v_1 m_{31}^1 - m_{21}^1 & v_1 m_{32}^1 - m_{22}^1 & v_1 m_{33}^1 - m_{23}^1 \\ \cdots & \cdots & \cdots \\ u_n m_{31}^n - m_{11}^n & u_n m_{32}^n - m_{12}^n & u_n m_{33}^n - m_{13}^n \\ v_n m_{31}^n - m_{21}^n & v_n m_{32}^n - m_{22}^n & v_n m_{33}^n - m_{23}^n \end{pmatrix}$$

$$U = \begin{pmatrix} m_{14}^1 - u_1 m_{34}^1 \\ m_{24}^1 - v_1 m_{34}^1 \\ \cdots \\ m_{14}^n - u_n m_{34}^n \\ m_{24}^n - v_n m_{34}^n \end{pmatrix}$$

Since we have found the cameras' projecting parameters and their corresponding pixel points, we can acquire the world point X by formula 4.

$$X = (K^T K)^{-1} K^T U \quad (4)$$

B. Recognition Of The Infrared Markers

In multi-user environment, we need to distinguish different users for VR application. So it is very important to correctly identify the infrared markers. In this system, we identify the infrared markers by their distance. The distances between infrared markers on different LED modules are different. Each LED module is composed by 4 or 5 infrared LED markers, as shown in Fig. 3. A1, B1, C1, D1 represent 4 infrared LED markers on a LED module. The value on the line represents the distance of the two markers.

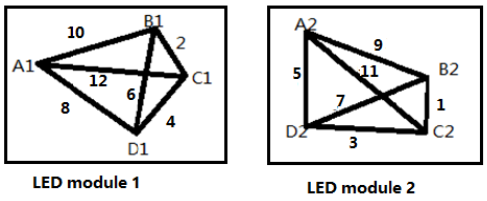

Fig. 3: Different LED module

After the LED module are made, we need to measure the distance among different markers on the same LED module in advance, and save the values as a vector whose elements sorted from small to large. For example, if there are 4 infrared LED on a LED module, we can obtain a 4 rows and 3 columns matrix. Each row of the matrix is sorted increasingly. If there are 3 LED modules, we can obtain a 12 rows and 3 columns matrix. Each row represents an infrared marker on the LED module.

Once the three-dimensional coordinates of all tracked markers have been computed, we can calculate the distances between different markers, and sort the distances increasingly. Then we delete the distance larger than the diagonal of the biggest LED module to eliminate interference when identifying markers. Next, we can compare the measuring distance vector of tracking markers to each row of the matched matrix according to the similarity to distinguish each markers. The similarity calculation of two vector (V1 and V2) is shown in table II. The size of V1 is not less than the size of V2. When there are multiple users in the system, different markers' distance vector may be similar. Therefore, after distinguish every infrared marker, we need verify whether the three markers who are closed to each other are recognized in the same LED module. We need recognize the marker once again whose nearest two markets are recognized in other LED module. In this algorithm, we can precisely distinguish the markers when the measuring distance vector of makers is accurate. At last, we eliminate identification overlapping markers just in case, and save the markers whose similarity in the threshold .

TABLE II: COMPUTER THE SIMILARITY OF TWO DISTANCE VECTOR

Step1:

- Set similarity s=0. Find the most similar element to V2[0] from vector V1, and save its location in V1 as t1.
- Find the most similar element to V2[m-1] from vector V1, and save its location in V1 as t2. m is the size of vector V2.

Step 2:

- If t1=t2, vector V1 is not similar to vector V2, set similarity to the max, return.
- If t2-t1=m-1, the similarity is the sum that abs (V1[t1+i]-V2[i]), which i is from 0 to m-1, return.

Step 3:

- Otherwise, for each element from V2, find the nearest element from V1, and ensure that the corresponding elements in order. Detailed steps are as follows:

1) Set label=-1, which means that none element in V1 has been used.
2) For each V2[i], where i is from 0 to m-1, calculate the distance of every element of V1 to V2[i] to get vector dist1 and sort it from small to large, then save the reverse order as order-num1. Find a number from vector order-num1 that is bigger than label, and set it to w1.
3) If i=m-1, s=s+dist1[w1], return.
4) Else, calculate the distance of every element of V1 to V2[i+1] to get vector dist2 and sort them from small to large, then save the reverse order as order-num2. Find a number from vector order-num2 that is bigger than label, and set it to w2.
5) If w2>w1, which means the element that is similar to V2[i] is different from the element that is similar to V2[i+1]. So s=s+dist1[w1], label=w1.
6) Else if label=-1:
   1 if w1=0, which means the most similar element to V2[i] and V2[i+1] is the first element of V1. So s=s+dist1[w1], label=w1;
   2 else s=s+dist1[w1-1],label=w1-1.
7) Else if w1=label, which means the most similar element to V2[i] and V2[i+1] is the same element of V1 that is similar to V2[i-1]. So s=s+dist1[w1+1],label=w1+1.
8) Else if w1=label+1,which means the most similar element to V2[i] and V2[i+1] is in the back of the element of V1 that is similar to V2[i-1]. So s=s+dist1[w1],label=w1.
9) Else s=s+dist1[w1-1],label=w1-1.
10) i=i+1,go back to 2).

C. Calculating Of The Users' Position And Orientation

After the infrared markers are distinguished, we use Kalman filter to optimize the world coordinates of the infrared markers. Kalman filter can improve the measurement accuracy and reduce the jitter of the coordinates data. Besides, if less than three infrared markers are distinguished in a LED module and the absent markers are detected in previous frame, we can predict the absent markers by Kalman filtering. Once three or more markers in a LED module are distinguished, we can acquire the transition matrix. The transition matrix represents the rotation and translation that current LED module relative to its initial status. We can extract its rotation matrix and transform it into quaternion or Euler angle to represent the orientation information of the LED module.

Since each LED module is fixed on different users head, the spatial relationship between LED module and user's head is unchanged. As a result, the change of the LED module's position and orientation can be used to represent the change of the user's position and orientation. So we can get each user' position and orientation by the markers on each LED module. At last, we transfer the position and orientation of all users to unity3d in the laptop computer by wireless network. Therefore unity3d can render VR scene according to its user's position and orientation and transfer the VR scene to HMD. So, different users can see appropriate virtual scenes according to their location and orientation.

IV. Experiments

A. Test Accuracy And Stability

We use three infrared cameras with resolution of 1280x720 pixels to conduct the experiment. There is a user with a LED module in the workspace who is nearly three meters away from the cameras. The user is stationary at the position (x, y, z)=(368.5mm, 471.5mm, 7.5mm) relative to the world coordinates that we have calibrated, rotation (yaw, pitch, roll)=$(1.0^o, 3.0^o, 5.0^o)$ measured by a calibrated gyro sensor. The measurement results are respectively shown in Fig.4, Fig. 5.

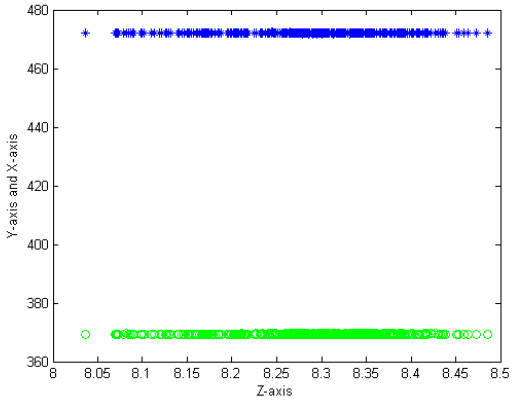

Fig. 4: The measure of user's position

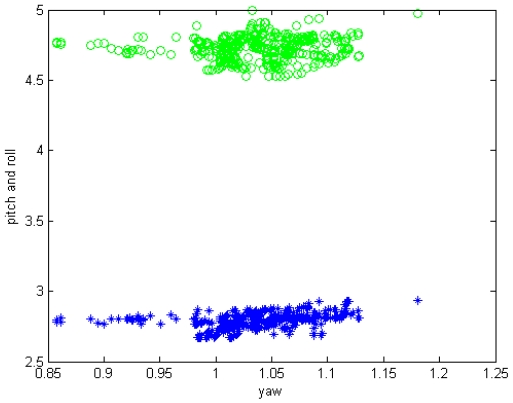

Fig. 5: The measure of user's orientation

By analyzing the data, we get the table III and table IV. We change the users' position and orientation and repeat the same experiment several times. By repeatedly analyzing the figures and tables, we conclude that our positioning accuracy is within 7mm, the jitter of position value is within 1mm. The accuracy of orientation is within 4 degree with the slight fluctuations of less than 1 degree. We can conclude that data is relatively stable from different frames. The fluctuations don't make an influence to VR scene display. The accuracy of position and orientation is sufficient for virtual scene display without confusion.

TABLE III: USER'S POSITION

	min	max	mean	variance
x(mm)	369.2700	369.6090	369.4185	0.0045
y(mm)	471.9090	472.3470	472.0449	0.0040
z(mm)	8.0358	8.4860	8.2917	0.0065

TABLE IV: USER'S ORIENTATION

	min	max	mean	variance
yaw(^{o})	0.8575	1.1806	1.0376	0.0026
pitch(^{o})	2.6694	2.9370	2.7944	0.0029
roll(^{o})	4.5292	4.9975	4.7261	0.0072

B. Test data refresh rate

The data refresh rate is influenced by a variety of factors. We conduct the experience in different resolution and number of infrared cameras or users. Here are some results of refresh rate in different experimental environment. Table V and table VI show that the more cameras, the more users, or the bigger resolution used in experiment will result in the lower refresh rate. Increasing number of cameras in this system is helpful to improve the accuracy of positioning and recognition and expand the workspace. Different resolutions have little effect on the accuracy in the indoor environment, but a great impact on the data refresh in real time. So we can appropriately reduce the resolution to improve the data refresh rate. With many

times experimental verification, three cameras with resolution of 1280x720 pixels can provide a sufficiently refresh rate and measurement accuracy for VR application in the indoor environment of 3m by 3m.

TABLE V: REFRESH RATE WITH 3 CAMERAS

	1920X1080	1600X1200	1280X720
one user(s^{-1})	57	68	110
two user(s^{-1})	34	50	73

TABLE VI: REFRESH RATE WITH 4 CAMERAS

	1920X1080	1600X1200	1280X720
one user (s^{-1})	34	46	77
two user(s^{-1})	20	34	52

V. CONCLUSION

This paper describes an indoor tracking and recognition system with the performance of low price, high accuracy, high stability and high data refresh rate. It is convenient to expand the workspace by increasing the number of cameras. The system allows multi-user to move in the workspace at the same time. However increasing the number of users and cameras can reduce the refresh rate of tracking data. Less than four users are supported to maintain the accuracy and real time with three cameras applied to the system currently. In the future, we will improve our algorithm to support more users in larger area with high refresh rate. Besides, we will continuous improve prediction in this system to reduce the impact that some markers are disappear.

VI. ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China under Grant No. 61403373.

REFERENCES

[1] Heim M. The metaphysics of virtual reality[M].Oxford University Press,USA,1993.

[2] Burdea G, Coiffet P. Virtual reality technology[J]. Presence: Teleoperators and virtual environments, 2003, 12(6): 663-664.

[3] Foxlin E, Naimark L. VIS-Tracker: A Wearable Vision-Inertial Self-Tracker[J]. VR, 2003, 3: 199.

[4] Hogue A, Jenkin M R, Allison R S. An optical-inertial tracking system for fully-enclosed vr displays[C]. Computer and Robot Vision, 2004. Proceedings. First Canadian Conference on. IEEE, 2004: 22-29.

[5] Vorozcovs A, Strzlinger W, Hogue A,et al.The hedgehog:a novel optical tracking method for spatially immersive displays[J]. Presence, 2006, 15(1): 108-121.

[6] Piekarski W, Thomas B H. The tinmith system: demonstrating new techniques for mobile augmented reality modelling[M]. Australian Computer Society, Inc., 2002.

[7] Koneru U, Redkar S, Razdan A. Fuzzy logic based sensor fusion for accurate tracking[C]. International Symposium on Visual Computing. Springer Berlin Heidelberg, 2011: 209-218.

[8] KANBARA R T M, YOKOYA N. A wearable augmented reality system using positioning infrastructures and a pedometer[C]. Proceedings of the seventh IEEE international symposium on wearable computers (ISWC03). 2003, 1530: 17-00.

[9] Maeda M, Ogawa T, Kiyokawa K, et al. Tracking of user position and orientation by stereo measurement of infrared markers and orientation sensing[C]. Wearable Computers, 2004. ISWC 2004. Eighth International Symposium on. IEEE, 2004, 1: 77-84.

[10] Weng D, Liu Y, Wang Y, et al. Study on an indoor tracking system based on primary and assistant infrared markers[C]. Computer-Aided Design and Computer Graphics, 2007 10th IEEE International Conference on. IEEE, 2007: 377-382.

[11] Faugeras O. Three-Dimensional Computer Vision: A Geometric Point of View[J]. 1993.

[12] Zhang Z. Flexible camera calibration by viewing a plane from unknown orientations[C]. Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on. Ieee, 1999, 1: 666-673.

[13] Zhang Z. Motion and structure from two perspective views: from essential parameters to Euclidean motion through the fundamental matrix[J]. JOSA A, 1997, 14(11): 2938-2950.

[14] Faugeras O. Three-dimensional computer vision: a geometric viewpoint[M]. MIT press, 1993.

[15] Rothwell C, Csurka G, Faugeras O. A comparison of projective reconstruction methods for pairs of views[C]. Computer Vision, 1995. Proceedings., Fifth International Conference on. IEEE, 1995: 932-937.