1. Introduction

The detection and tracking of a user's head movement or a part of their body (hands and faces) represent a very important step in the design of Human Computer Interaction (HCI) in virtual reality systems (VR).

Low-cost motion tracking systems are required for many different applications, such as health, industry, serious games, or collaborative virtual environments. The development of such systems is still challenging. A low-cost 3D tracking system is presented in this chapter. This system has been developed and tested in order to move away from traditional 2D interaction techniques (keyboard and mouse) in an attempt to improve user's experience while using a virtual environment. Such a tracking system is used to implement 3D interaction techniques that augment user experience, promote user’s sense of transportation in the virtual world as well as user’s awareness of their partners. The tracking system is a passive optical tracking system using stereoscopy, a technique allowing the determination of three-dimensional information of a real object from a couple of images. The infra-red rays are reflected by reflective markers, the resulting monochromatic images are processed in order to compute the 3D positions of the markers in real-time.

The presented tracking system is a cost-effective and flexible tracking system that detects and follows the movements of a user using a set of markers. The tracking system consists of two standard webcams equipped with infra-red projectors, a set of reflective markers and a detection algorithm. In contrast with existing tracking systems, this system is lightweight and very reasonable in price (less than 200 Euros), making it more accessible for research and low-cost application purposes. Another advantage of this tracking system is its flexibility. It can automatically detect up to 7 markers in the same scene because the detection algorithm implemented is not based on pre-existing forms. This 3D tracking system has been improved on a collaborative research platform for investigating 3D manipulation/interaction techniques in CVEs.

2. 3D tracking technologies

Many virtual reality applications require such data as type information, the 3D position or rotation of an object or user in the workspace. The conventional input device (mouse, keyboard, joystick) does not have such information. It is therefore necessary to use new...
adapted devices, called tracking devices or trackers. Usually, there are different technologies used for 3D motion tracking.

![Taxonomy of mostly used tracking sensors](image)

- Mechanical trackers: mechanical tracking devices measure the position and orientation using a direct contact between a point of reference and the target. A major benefit of this type of device is very high accuracy and a low latency. Indeed, the delay for the mechanical trackers is very short (less than 5 ms), update rate is quite high (300 updates per second) and mechanical trackers are accurate. Their main disadvantage is that user's movement is limited by physical constraints.
- Magnetic trackers: magnetic trackers use transmitters’ low-frequency magnetic field to determine the position and orientation of a small sensor (receiver). Generally, they track in a volume of approximately 1.5 to 10 meters. Unfortunately, metallic objects in the space of the tracking introduce distortions in far therefore, a loss of precision and latency system.
- Inertial trackers: inertial trackers use small gyroscopes to detect rotations. This type of devices suffers from an accumulation of errors when they are used for long periods.
- Optical trackers: optical tracking systems use generally two cameras in order to detect the position and orientation of a target with computer vision software. Optical trackers offer higher than the magnetic or mechanical trackers tracking space. This type of tracking also provides high update rates. However, the presence of an obstacle in a camera field of view makes impossible the target detection.
- Acoustic trackers: acoustic tracking systems emit high frequency ultrasonic waves and determine the position and orientation of a target by using the principle of the technique of flight time. Indeed, the distance between the target and the transmitter can be determined by measuring the propagation of a wave between the transmitter and the target time. With the distance, the position may be estimated. Similarly, the orientation is calculated by using multiple receivers (at least 3). This type of system has the advantage of being inexpensive, lightweight. However, they are very sensitive to outside noise.

Several systems such as magnetic, mechanical or acoustic systems are used for detection and tracking. While the precision of these systems is very good, the detection and tracking rely on devices that user wears (for example: cables, sensors, etc). These additional devices affect users’ experience as they limit their movements during their 3D interaction.
3. System description

The presented tracking system uses stereoscopy (Figure 2) where two cameras are used and equipped with infra-red projector. The infra-red rays are reflected by reflective markers, the resulting monochromatic images are processed in order to compute the 3D positions of the markers in real-time. The presented tracking system is a cost effective and flexible tracking system that detects and follows the movements of a user in a CVE using a set of markers. The tracking system consists of two standard webcams equipped with infra-red projectors, a set of reflective markers and a detection algorithm. In contrast with existing tracking systems, this system is light weight and very reasonable in price (less than 200 Euros), making it more accessible for research purposes. Another advantage of this tracking system is its flexibility. It can identify automatically up to 7 markers in the same scene, the detection algorithm implemented is not based on pre-existing forms.

Fig. 2. System block diagram

4. Hardware

Standard webcams are equipped with an infra-red filter. In order to enable tracking in the infra-red range, the IR filters are removed of the webcams, in doing so making them sensitive to infra red waves (Figure 2). The lighted waves outside the infra-red field (spectrum visible by the human eye) are filtered in that way to capture only the markers illuminated by IR lighting. Spherical markers are used as these markers reflect the light in the opposite direction from where it is projected independently from the orientation of the marker. The markers receive light from a source of light made of infra-red projectors (per camera) so the light is in direction of the cameras. These projectors are implemented in a printed circuit on which forty infra-red LEDs are soldered (Figure 3).
5. Calibration

Stereoscopic calibration consists of determining the matrix of transformation between the reference systems of the right and left cameras, by using a known 3D pattern. A solid pattern made of two faces composed of a matrix of visible points can be used (Figure 4). The position of each point as well as the distance between each point of the matrix is known. Here, these visible points are actually infra-red leds.

Fig. 3. Material used filters, webcams modified, and markers

Fig. 4. 3D pattern used for calibration
The calibration process starts by acquiring an image of the pattern using each camera. The pattern must be completely motionless. The cameras must be correctly positioned, so that the field of view of the two cameras intersects. The two cameras must be sufficiently distant from one to another. Once the two cameras have acquired images, points of interest in the two images can be detected in using the Harris detector (Harris & Stephens, 1988). Once this step is completed, the intrinsic and extrinsic parameters of the cameras are calculated using the Faugeras-Toscani method (Faugeras & Toscani, 1987).

6. Markers detection

After validation of the calibration process, real time detection of infra-red markers can be processed. First, binarization processing of each captured image is performed, in order to filter out the points corresponding to the markers. Once the binarization is completed, the image is searched pixel by pixel in order to detect potential “spots” of light. Once a pixel is detected as belonging to a marker, the pixel is colored by a coded way (in black for example) and the barycentre detection can start: each neighbouring pixel is checked to see if it belongs to the marker. If it does then its neighbouring pixels are also checked. Each time a point is detected as belonging to a marker, its coordinates are stored in an array. Once all the neighbouring pixels have been checked, the average of these coordinates is calculated in order to obtain the position in 3D of the barycentre of the marker of interest. Once these coordinates have been calculated, detection of other markers resumes (Figure 5).

---

Fig. 5. Markers detection steps
7. Automatic labelling of markers

The automatic markers labelling is a complex task. The markers are identical from the cameras point of view. Thus, a post-processing treatment is conducted where each marker is affected by a fixed number, in order to differentiate the markers.

Methods for tracking multiple markers can easily be included. For example, it is possible to use the iterative method (Magneau, Bourdot, & Gherbi, 2004), Heuristic methods (Haritaoglu, Harwood, & Davis, 1998), the mean shift (Comaniciu & Meer, 1999), or the Multiple Hypothesis Tracker MHT (Reid, 1979). These methods use various sources of information for tracking multiple points (for example kinematics or the shape of object can be used). A simple optimization method, based on the position, can be used. The speed and the weight of the markers are taking into account to improve the robustness of the labelling process.

A Kalman’s filter (Kalman, 1960) can be used to predict the marker position, in the current image at t+1, based on its coordinates in the previous image at time t. Once the coordinates are obtained, they are compared with the coordinates of the points detected in the current image. The size T of the spots of lights produced by each marker on the image of the camera is taken into consideration as well as the speed V of the spot from one image to another one. All these parameters allow the implementation of a cost function (1). The cost function needs to be minimized in order to obtain the best configuration of the possible positions of the markers. The cost for a marker i, detected in the image, compared with a marker j whose coordinates have been predicted by the Kalman’s filter can be represented by the following formula:

\[ C_i^j = \delta(P_i, P_j') + \delta(V_i, V_j') + \delta(T_i, T_j') \]  

(1)

The pairs minimizing the sum of every cost function identify the markers (Figure 6).

Fig. 6. Markers detected
8. 3D triangulation

Knowing the intrinsic and extrinsic parameters of a system, it is possible to calculate the position of a point in a scene starting from two points of view. In using the pinhole model, the projection of a point $P(X, Y, Z)$ of the scene in the image plane of a camera is obtained by:

$$u = \frac{m_{11}X + m_{12}Y + m_{13}Z + m_{14}}{m_{13}X + m_{13}Y + m_{13}Z + m_{13}}$$

$$v = \frac{m_{21}X + m_{22}Y + m_{23}Z + m_{24}}{m_{23}X + m_{23}Y + m_{23}Z + m_{23}}$$

$$u' = \frac{m_{21}X + m_{22}Y + m_{23}Z + m_{24}}{m_{23}X + m_{23}Y + m_{23}Z + m_{23}}$$

$$v' = \frac{m_{21}X + m_{22}Y + m_{23}Z + m_{24}}{m_{23}X + m_{23}Y + m_{23}Z + m_{23}}$$

$$M_1 = \begin{bmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{12} & m_{12} & m_{13} & m_{14} \\ m_{13} & m_{13} & m_{13} & m_{14} \end{bmatrix}$$

$$M_2 = \begin{bmatrix} m_{21} & m_{22} & m_{23} & m_{24} \\ m_{22} & m_{22} & m_{23} & m_{24} \\ m_{23} & m_{23} & m_{23} & m_{24} \end{bmatrix}$$

With:

$M_1$ represents the matrix of perspective projection of the left camera and $M_2$, the matrix of projection of the right camera.

The equations can be written in the following way (2):

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = A \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} + B$$

With:

$$A = \begin{bmatrix} m_{11} - u.m_{13} & m_{12} - u.m_{13} & m_{13} - u.m_{13} \\ m_{21} - v.m_{23} & m_{22} - v.m_{23} & m_{23} - v.m_{23} \\ m_{21} - u'.m_{23} & m_{22} - u'.m_{23} & m_{23} - u'.m_{23} \end{bmatrix}$$

$$B = \begin{bmatrix} m_{13}u - m_{14} \\ m_{13}v - m_{12} \\ m_{21}u' - m_{24} \\ m_{23}v' - m_{24} \end{bmatrix}$$

The calculation of the point $P(X, Y, Z)$, is obtained by using the Least Squares Method, solving the following linear system of 4 equations and 3 unknown factors $(X,Y,Z)$ (3).

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = (A^t.A)^{-1}.A^t.B$$

9. Results

Usually, the points obtained after triangulation are not very stable. Indeed, a shifting of a pixel on the image can make the 3D coordinates vary. For instance, if a marker is put on a static object, every calculation of the 3D co-ordinates will result in slight different values of the coordinates. This difference is due to the fact that luminous markers are used. A small variation of light affects the calculations which results in an unstable object. To avoid that effect, a Kalman’s filter can be used to stabilize the system (Figure 7).
Virtual Reality

Fig. 7. Stabilization with kalman’s filter (top without kalman’s filter, down with kalman’s filter).

10. 3D interaction techniques

The user interaction with a VE is regarded today as a major axis of virtual reality (VR). The traditional methods of 2D interactions (mouse and menus) reach their limits quickly. It is thus necessary to find new forms of interaction in VE.

The 3D interaction can be defined like a form of interaction between the human and the machine in a three-dimensional context. 3D interaction does not necessarily mean that 3D input devices are used. For example in the same application, if the user clicks on a target object to navigate to the object, then the 2D mouse input has been directly translated into 3D location and thus 3D interaction occurs (Bowman et al., 2004).

According to Coquillart, 3D interaction techniques can be considered as methods for the use of hardware interface allowing the user to carry out a precise task in a VE (Coquillart et al., 2006). An interaction technique includes both hardware (input/output devices) and software components. The software component of the interaction technique is responsible for
mapping the information from the input devices into some action within the system, and for mapping the output of the system to a form that can be displayed by the output devices. Bowman decomposes the 3D interaction process in four virtual behavioral primitives (Bowman et al., 1999):

- Navigation
- Selection
- Manipulation
- Control application

Using this decomposition, interaction metaphors can be developed to enable the interaction process VE. These metaphors necessitate the use of our 3D tracking system. Applications that use these techniques of interactions will be presented in the next section.

11. Applications

Lots of 3D interaction applications require the use of a 3D tracking system. For example, a 3D low cost tracking system can be used in the context of collaborative virtual environment. Such applications require the use of 3D interaction techniques which use 3D tracking systems. Figure 8 shows an operation of selection and manipulation using a device tracked in space by a 3D low cost tracking system.

Fig. 8. The selection/manipulation application with the 3D tracking.

Figure 9 shows an application of 3D navigation in a virtual environment.
Fig. 9. Navigation application with the 3D tracking.

The figure 10 shows a Virtual ball game for rehabilitation in this application the examiner indicates which objects the patient must catch and which objects should avoid. For example it may define the patient will have to catch the blue balls and abandon the red balls.

Fig. 10. Rehabilitation application with the 3d tracking system.
12. Conclusion

3D tracking systems are required in numerous 3D interaction applications in various domains. With the spread of virtual reality applications in different domains, 3D low cost tracking systems are required. The presented 3D tracking system uses common technologies and known image processing. This technology has to be coupled with efficient interaction metaphors for a better realistic interaction and natural perception of the user on his interaction.

13. References


