SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

April 2016, Vol. 59 042412:1–042412:12
doi: 10.1007/s11432-015-5516-2

A low-complexity sensor fusion algorithm based on a fiber-optic gyroscope aided camera pose estimation system

Zhongwei TAN[1], Chuanchuan YANG[1*], Yuliang LI[1], Yan YAN[2], Changhong HE[1], Xinyue WANG[3] & Ziyu WANG[1]

[1]*State Key Laboratory of Advanced Optical Communication Systems and Networks, School of Electronics Engineering and Computer Science, Peking University, Beijing* 100871*, China;*
[2]*School of Information Science and Technology, Xiamen University, Xiamen* 361005*, China;*
[3]*College of Information and Electrical Engineering, China Agricultural University, Beijing* 100083*, China*

Received September 16, 2015; accepted October 20, 2015; published online February 1, 2016

Abstract Visual tracking, as a popular computer vision technique, has a wide range of applications, such as camera pose estimation. Conventional methods for it are mostly based on vision only, which are complex for image processing due to the use of only one sensor. This paper proposes a novel sensor fusion algorithm fusing the data from the camera and the fiber-optic gyroscope. In this system, the camera acquires images and detects the object directly at the beginning of each tracking stage; while the relative motion between the camera and the object measured by the fiber-optic gyroscope can track the object coordinate so that it can improve the effectiveness of visual tracking. Therefore, the sensor fusion algorithm presented based on the tracking system can overcome the drawbacks of the two sensors and take advantage of the sensor fusion to track the object accurately. In addition, the computational complexity of our proposed algorithm is obviously lower compared with the existing approaches (86% reducing for a 0.5 min visual tracking). Experiment results show that this visual tracking system reduces the tracking error by 6.15% comparing with the conventional vision-only tracking scheme (edge detection), and our proposed sensor fusion algorithm can achieve a long-term tracking with the help of bias drift suppression calibration.

Keywords visual tracking, camera pose estimation, fiber-optic gyroscope, low-complexity, sensor fusion

Citation Tan Z W, Yang C C, Li Y L, et al. A low-complexity sensor fusion algorithm based on a fiber-optic gyroscope aided camera pose estimation system. Sci China Inf Sci, 2016, 59(4): 042412, doi: 10.1007/s11432-015-5516-2

1 Introduction

Visual tracking can present a real-time camera pose estimation related to its environment or other solid objects [1], which has a wide range of applications. For example, Ref. [2] implemented an augmented reality (AR) application, which uses several sensors to track the features in a large room and draw visual overlays. In contrast to other kinds of tracking, such as acoustic or magnetic tracking [3,4], visual tracking has the advantages that it is capable of tracking a larger range and it is not prone to interference [1].

* Corresponding author (email: yangchuanchuan@pku.edu.cn)

Visual tracking is able to estimate the position of objects or cameras from the image observed directly as well. However, the conventional vision-based tracking methods often suffer from some shortcomings, such as the lack of robustness and high computational cost. They usually fail to track a rapid or jumpy motion because the general mathematical model of the object is not appropriate for characterizing such motion. Several methods have been proposed to overcome the disadvantages of the conventional visual tracking methods. Refs. [5, 6] used swarm intelligence based searching method to handle the abrupt motion, which improves both the accuracy and efficiency of the visual tracking. Ref. [7] put forward a vision-only tracking algorithm, which presents a temporal probabilistic combination of discriminative observers according to different life spans, thus improves the accuracy of tracking when there exists an abrupt motion, but the rapid motion still significantly degrades the quality of the video image. Generally, the conventional vision-only tracking methods use cameras to capture the images all the time and do image processing, which has an acknowledged problem of involving the high computational complexity [8], to orientate the object in each image. Meanwhile, the conventional methods largely depend on the quality of every image.

Inertial sensors, such as rate gyroscopes and linear accelerometers, which provide the measurements for the rotation velocity and the linear acceleration, are widely used in motion tracking [9–12]. Inertial sensors can be sampled at a relatively high frequency (100 Hz in this paper, compared with 30 Hz or so for the general video rates) and with a low latency, thus they are able to provide more amounts of information for the pose estimation than the cameras at the same time. The signals of inertial sensors are integrated to predict the position, since the inertial sensors are so robust to track an abrupt motion that they are ideal for complementing the visual tracking. However, there is still a problem worthy of concern that the interference of the drift and noise in an inertial sensor should be reduced or corrected.

The fusion of vision data and inertial measurements has been the focus of substantial research [13], which attempts to overcome the drawback of using only one sensor. Ref. [9] provided a full analysis of different methods in fusing accelerometer and gyroscope data to camera measurements with extended Kalman filters (EKF). In addition, Project Glass is a fresh research and development program by Google to develop an augmented reality optical head-mounted display (OHMD) in this area [14]. Before the development of the Project Glass, there is a wide use of sensor fusion for visual tracking. For instance, in [15], the authors fused the data from cameras and inertial sensors with two EKFs, one of which estimates the motion and the other estimates the 3D locations of the points in the scene. One advantage of the method in [15] is that there is no need of prior knowledge about the scene during tracking. To correct the slow drift caused by the gyro data integration, Ref. [10] employed inclinometers as well as a compass, and designed a Kalman filter to integrate the data from visual sensors and inertial sensors to achieve the tracking of three dimensions of freedom (3DOF) orientations. Ref. [16] constructed a two independent channel motion-filter structure, where one channel is used for the low-frequency visual sensor while the other is used for the high-frequency inertial gyroscope, thus the fusion is performed by using an EKF to track 6DOF orientations. Besides implementing the fusion with the extended Kalman filters, Ref. [1] provided a method to combine visual sensors and the gyroscope information to improve the stability and robustness of tracking when there is motion blur during tracking. In [1], rotation measurements are used not only to predict the pose but also to modify the operation of edge detection with the estimation by the exposure time of the camera and a new edge detection kernel.

The fiber-optic gyroscope is a ring interferometer based on the Sagnac effect, which can produce an output voltage which linearly varies with the rotation velocity within a limit range [17]. Its performance could allow strategic applications, and its appropriate size makes it convenient to be used in different environments [18]. Thus, fiber-optic gyroscopes are applicable in the visual tracking system. Compared with the MEMS gyroscopes, which are usually used in the previous visual tracking research, fiber-optic gyroscopes are more accurate, especially for their bias drift (The bias drift stability of the fiber-optic gyroscope is about 0.01–0.001°/h, while the MEMS gyroscope is about 1°/h [18]). For the visual tracking system based on the fusion of the camera and the MEMS gyroscopes, it is usually needed to provide a complex algorithm to accomplish the sensor fusion work, since the MEMS gyroscopes could not be employed to locate the object without the help of image processing in a long term due to their inaccuracy,

as a result that the frequency for image processing is still high. If utilizing the fiber-optic gyroscopes instead of the MEMS gyroscopes, the fusion system becomes more reliable and does not need to rely on the complex algorithm to switch frequently between cameras and gyroscopes as the method given in [16].

In this paper, we innovatively propose a sensor fusion algorithm based on the combination of cameras and fiber-optic gyroscopes to design a camera pose estimation system. We just conduct the image processing at the beginning of each tracking stage, and then make full use of the accurate data from the fiber-optic gyroscopes during the tracking stage to implement the control over the position and the orientation of the object. Since the fiber-optic gyroscopes are "blind" to locate the object in the image at the beginning, the cameras are needed. After capturing the initial object coordinates, only the fiber-optic gyroscope is enough to finish the estimation due to its accuracy. During the whole tracking process, we only use a few frames from camera which have a relatively long interval time between each other but derive a very efficient camera estimation system with the aid of the fiber-optic gyroscope. Our proposed algorithm can considerably reduce the complexity of the whole visual tracking system. In this paper, we assume that the rotation is a dominant camera motion since the camera rotation includes more rapid and serious image deformations, thus we do not contain the accelerometer in our sensor fusion algorithm as [19].

In addition, a long-term tracking experiment is conducted based on this camera pose estimation system. As for a long-term tracking, the calibration of the tracking results is necessary, since the accumulation of the bias drift in the gyroscope cannot be neglected. With the help of our previous work, our proposed algorithm can easily achieve a long-term tracking effectively after the calibration of bias drift suppression.

Generally, our contribution in this paper can be summarized as follows:

(1) We propose a novel fiber-optic gyroscope aided sensor fusion algorithm to achieve camera pose estimation, which reduces the computational complexity but earns an effective tracking result.

(2) Our proposed algorithm can achieve a long-term tracking with the help of bias drift suppression calibration.

The structure of this paper is as follows. Section 2 describes the proposed camera pose estimation system. Section 3 shows the fusion experiment, simulation, as well as the results. Section 4 draws the conclusion.

2 The proposed camera pose estimation system

Figure 1 illustrates the fusion system we proposed. The camera provides the image of the scenes, while the gyroscope provides rotation measurements for tracking. The signal processing module detects the object and implements the tracking. The camera and the gyroscope synchronously work during the visual tracking. Images from the camera and data from the gyroscope are transmitted to the signal processing module in real time. At the same time, a method for suppressing the gyroscope bias drift and a Kalman filter can be chosen to be employed to improve the performance of the system.

Figure 2 illustrates the strategy of the fusion system we proposed. The whole tracking process includes N tracking stages. For each stage, at the discrete time j=0, initialization is performed and then the tracking stage follows from discrete time j=1 to j=n. The initialization is realized by the image processing with the data from the camera, and then the coordinates of object are tracked with the data from the gyroscope.

Initialization of the visual tracking system is described by a mathematical model in detail in Subsection 2.1. The tracking processes of the object coordinates in the tracking system are given in Subsection 2.2. Subsection 2.3 provides the summary and complexity discussion of the system.

2.1 Initialization of the tracking system ($j = 0$)

The gyroscopes are "blind" to capture the image when the object enters the scene. However, the camera provides the image so that the system is able to learn about the object by object detection and initialize the tracking system.

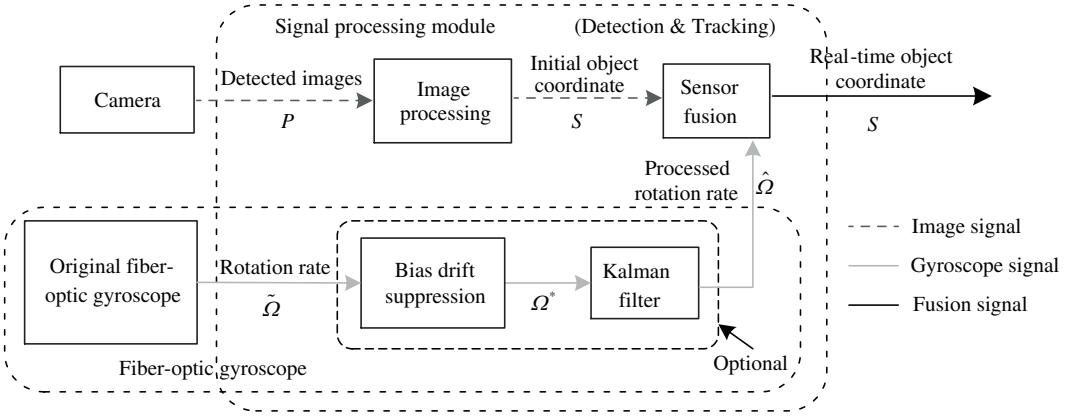

Figure 1 The proposed sensor fusion system.

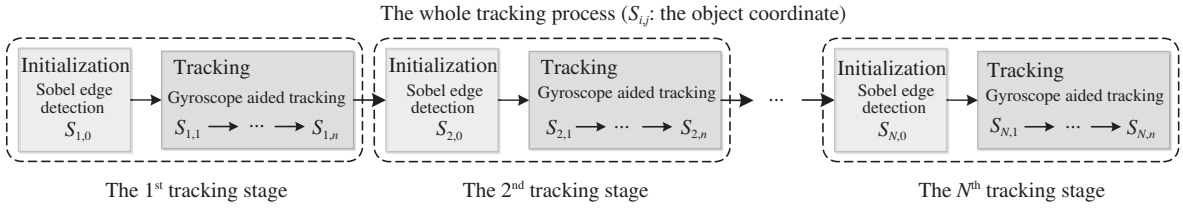

Figure 2 The proposed visual tracking strategy.

We can detect the first frame $\boldsymbol{P}_{i,0}$ of each fusion stage through the camera. Except for some complex objects such as human faces, which are only detected by using face detection or other complex methods as in [7, 20], general cases, such as simple marks and real scene images, can be easily detected by using edge detection.

From $\boldsymbol{P}_{i,0}$, we could locate the object in the image and find out a set of points $\boldsymbol{S}_{i,0}$ to mark the object edges in the pixel coordinate by the Sobel edge detection [21]. The number M of the points we used to describe the object edges depends on the characteristics of the object. The matrix $\boldsymbol{S}_{i,0}$ takes the form

$$\boldsymbol{S}_{i,0} = \mathscr{S}(\boldsymbol{P}_{i,0}) = \begin{bmatrix} u_1^{(i,0)} & u_2^{(i,0)} & \cdots & u_M^{(i,0)} \\ v_1^{(i,0)} & v_2^{(i,0)} & \cdots & v_M^{(i,0)} \end{bmatrix}, \tag{1}$$

where $\mathscr{S}$ means the Sobel edge detection function, $u_p^{(i,0)}$ is the coordinate value in the u-direction, and $v_p^{(i,0)}$ is the coordinate value in the v-direction. The superscript $(i, 0)$ means that the coordinates are obtained in the initialization of the i-th tracking stage, and the subscript p refers to the p-th point to describe the object, and $p = 1, 2, \ldots, M$.

We define the distance vector $\boldsymbol{D}_{i,0}$, which is essential for initializing the tracking system, as

$$\boldsymbol{D}_{i,0} = \begin{bmatrix} d_1^{(i,0)} & d_2^{(i,0)} & \cdots & d_M^{(i,0)} \end{bmatrix}. \tag{2}$$

The component $d_p^{(i,0)}$ is defined as

$$d_p^{(i,0)} = \sqrt{\left(x_p^{(i,0)}\right)^2 + \left(y_p^{(i,0)}\right)^2 + \left(z_p^{(i,0)}\right)^2}, \tag{3}$$

where $\left(x_p^{(i,0)}, y_p^{(i,0)}, z_p^{(i,0)}\right)$ is the position of the object in the camera coordinate. The distance vector $\boldsymbol{D}_{i,0}$ is difficult to measure accurately. Fortunately, on most occasions, the size of the object is far less than the distance between the object and the camera, so it is feasible to consider that all the distance values are approximately the same.

In order to obtain the position of the object in the 3-D camera coordinate, we need to use the following

Eq. (4) for each p from 1 to M:

$$\begin{cases} \begin{bmatrix} u_p^{(i,0)} \\ v_p^{(i,0)} \end{bmatrix} = \dfrac{\gamma}{z_p^{(i,0)}} \begin{bmatrix} f_u & 0 \\ 0 & f_v \end{bmatrix} \begin{bmatrix} x_p^{(i,0)} \\ y_p^{(i,0)} \end{bmatrix} + \begin{bmatrix} u_0 \\ v_0 \end{bmatrix}, \\ d^{(i,0)} = \sqrt{\left(x_p^{(i,0)}\right)^2 + \left(y_p^{(i,0)}\right)^2 + \left(z_p^{(i,0)}\right)^2}, \end{cases} \tag{4}$$

where f_u, f_v are the focal lengths in pixel, u_0 and v_0 are the pixel coordinates of the center point in an image, γ is a parameter varying with the coordinate of each point to approximate and correct the substantial radial distortion caused by a wide angle lens camera, and $d^{(i,0)}$ can be derived by measuring the distance between the object and the camera. The first equation in Eq. (4) is described in detail in the Appendix, which gives the transformation of the object between the camera coordinates and the pixel coordinates.

From Eq. (4), we can solve the 3-D camera coordinate of the p-th point as

$$\begin{bmatrix} x_p^{(i,0)} \\ y_p^{(i,0)} \\ z_p^{(i,0)} \end{bmatrix} = \frac{d^{(i,0)}}{\sqrt{\frac{1}{\gamma^2 f_u^2}\left(u_p^{(i,0)} - u_0\right)^2 + \frac{1}{\gamma^2 f_v^2}\left(v_p^{(i,0)} - v_0\right)^2 + 1}} \begin{bmatrix} \frac{1}{\gamma f_u}\left(u_p^{(i,0)} - u_0\right) \\ \frac{1}{\gamma f_v}\left(v_p^{(i,0)} - v_0\right) \\ 1 \end{bmatrix}, \tag{5}$$

and get the location $\boldsymbol{L}_{i,0}$ of the object in the 3-D camera coordinate, which is shown as follows:

$$\boldsymbol{L}_{i,0} = \begin{bmatrix} x_1^{(i,0)} & x_2^{(i,0)} & \cdots & x_M^{(i,0)} \\ y_1^{(i,0)} & y_2^{(i,0)} & \cdots & y_M^{(i,0)} \\ z_1^{(i,0)} & z_2^{(i,0)} & \cdots & z_M^{(i,0)} \end{bmatrix}. \tag{6}$$

2.2 Tracking stages ($j = 1$ to $j = n$)

After the initialization, the coordinates of the object are tracked with the aid of the output data of a fiber-optic gyroscope.

As shown in Figure 1, the relative rotation rate between the camera and the object is measured by the fiber-optic gyroscope and processed by the signal processing algorithms.

2.2.1 *The motion model*

We can get the rotation angle by

$$\boldsymbol{\Theta}_{i,j} = \begin{bmatrix} \theta_{i,j} \\ \phi_{i,j} \\ \delta_{i,j} \end{bmatrix} = \hat{\boldsymbol{\Omega}}_{i,j} \cdot \Delta t, \tag{7}$$

where $\hat{\boldsymbol{\Omega}}_{i,j}$ is the processed output of the fiber-optic gyroscope, Δt is the time interval between two discrete times, and θ, ϕ, δ are Euler Angles, which are shown in Figure 3.

In order to describe the whole related motion model between the camera and the object, we define the transform matrix $\boldsymbol{E}_{i,j}$ as

$$\boldsymbol{E}_{i,j} = \begin{bmatrix} \boldsymbol{R}_{i,j} & \boldsymbol{t}_{i,j} \\ \boldsymbol{0}_{1\times 3} & 1 \end{bmatrix}, \tag{8}$$

where $\boldsymbol{R}_{i,j}$ is the rotation matrix during the interval between the discrete time $j-1$ and the discrete time j in the i-th tracking stage. $\boldsymbol{t}_{i,j}$ is the translation vector, as is shown in Figure 3, which needs a linear accelerometer to get.

Note that a general motion model should be measured by both gyroscopes and accelerometers. However, the visual tracking precision affected by the fiber-optic gyroscope is concerned more in this paper, thus the detailed description for the accelerometer is omitted.

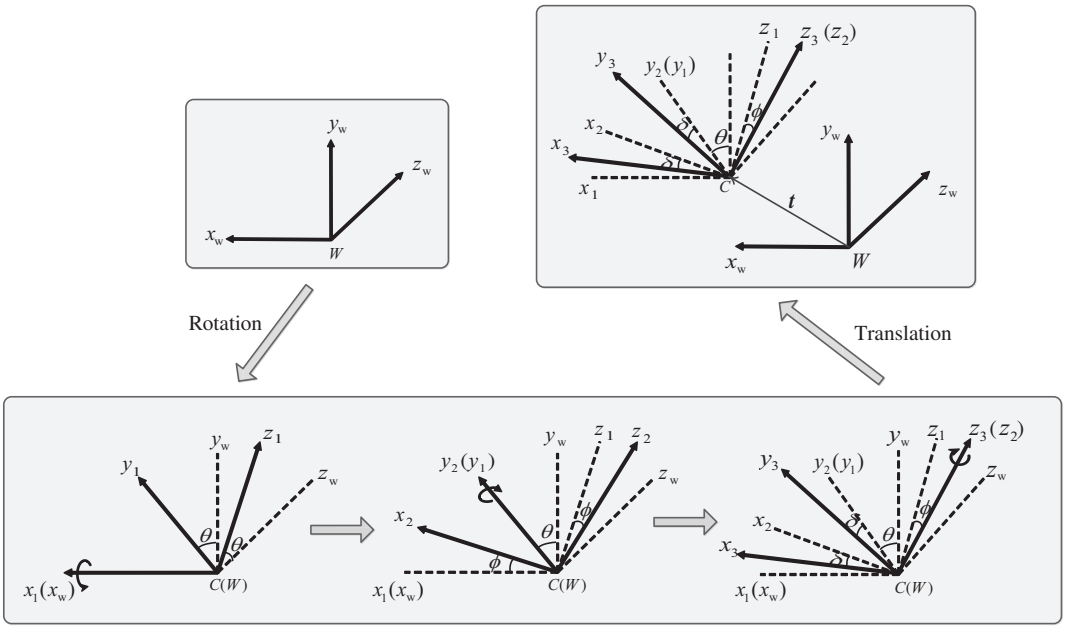

Figure 3 Motion model and related coordinates."W" means world-coordinate, and "C" means camera-coordinate.

2.2.2 *Sensor fusion*

In this system, the sensor fusion is conducted as the following process. At the discrete time j in the i-th tracking stage, the location of the object $\boldsymbol{L}_{i,j}$ in the camera coordinate is able to be calculated by

$$\begin{bmatrix} \boldsymbol{L}_{i,j} \\ \mathbf{1}_{1\times M} \end{bmatrix} = \boldsymbol{E}_{i,j} \cdot \begin{bmatrix} \boldsymbol{L}_{i,j-1} \\ \mathbf{1}_{1\times M} \end{bmatrix}, \tag{9}$$

where $\mathbf{1}_{1\times M}$ is the $1 \times M$ matrix of all. Thus, we can get the concrete form of $\boldsymbol{L}_{i,j}$:

$$\boldsymbol{L}_{i,j} = \begin{bmatrix} x_1^{(i,j)} & x_2^{(i,j)} & \cdots & x_M^{(i,j)} \\ y_1^{(i,j)} & y_2^{(i,j)} & \cdots & y_M^{(i,j)} \\ z_1^{(i,j)} & z_2^{(i,j)} & \cdots & z_M^{(i,j)} \end{bmatrix}, \tag{10}$$

where each column in this matrix gives the coordinate value of each point in the camera coordinate at the discrete time j in the i-th tracking stage.

The position of every point to mark the object in the pixel coordinate can be calculated as Eq. (11) for each p from 1 to M:

$$\begin{bmatrix} u_p^{(i,j)} \\ v_p^{(i,j)} \end{bmatrix} = \frac{\gamma}{z_p^{(i,j)}} \begin{bmatrix} f_u & 0 \\ 0 & f_v \end{bmatrix} \begin{bmatrix} x_p^{(i,j)} \\ y_p^{(i,j)} \end{bmatrix} + \begin{bmatrix} u_0 \\ v_0 \end{bmatrix}, \tag{11}$$

where the parameters $(f_u, f_v, u_0, v_0, \gamma)$ have the same definition as them in Eq. (4). Last, we get $\boldsymbol{S}_{i,j}$ to mark the object in the pixel coordinate at the discrete time j in the i-th tracking stage, which is defined as

$$\boldsymbol{S}_{i,j} = \begin{bmatrix} u_1^{(i,j)} & u_2^{(i,j)} & \cdots & u_M^{(i,j)} \\ v_1^{(i,j)} & v_2^{(i,j)} & \cdots & v_M^{(i,j)} \end{bmatrix}. \tag{12}$$

2.3 Summary and complexity discussion

The proposed algorithm presented above is summarized in Algorithm 1. Our proposed method in this paper provides an innovative approach. During the whole visual tracking process, which is shown in Figure 2, the proposed system only does the image processing problem at the initialization. Obviously, the initializations occupy a small part among the whole tracking process. In the other stage, the tracking

system takes advantage of the fiber-optic gyroscope to conduct the tracking and avoids complex image processing problems. Thus, our system has a much lower complexity, compared with the conventional vision-only tracking method.

Algorithm 1 Algorithm of the proposed visual tracking system

Require: image $\boldsymbol{P}_{i,0}$ captured by the camera, rotation rates $\hat{\boldsymbol{\Omega}}_{i,j}$, translation vector $\boldsymbol{t}_{i,j}$, time interval Δt, the distance between the camera and the object $d^{(i,0)}$;

for $i = 1, i \leqslant N, i = i + 1$ **do**
  Initialization of the tracking system:
  . Use $\boldsymbol{P}_{i,0}$ to obtain the pixel coordinates matrix $\boldsymbol{S}_{i,0}$ of the object by Sobel edge detection;
  . Use $d^{(i,0)}$, $\boldsymbol{S}_{i,0}$ to calculate the camera coordinates matrix $\boldsymbol{L}_{i,0}$ by Eq. (5);
  for $j = 1, j \leqslant n, j = j + 1$ **do**
    Tracking stage:
    . Use $\hat{\boldsymbol{\Omega}}_{i,j}$ and $\boldsymbol{t}_{i,j}$ to obtain the transform matrix $\boldsymbol{E}_{i,j}$ by Eq. (7) and Eq. (8);
    . Use $\boldsymbol{E}_{i,j}$ and $\boldsymbol{L}_{i,j-1}$ to calculate $\boldsymbol{L}_{i,j}$ by Eq. (9);
    . Use $\boldsymbol{L}_{i,j}$ to calculate $\boldsymbol{S}_{i,j}$ by Eq. (10);
  end for
end for

The comparison of the computational complexity between our proposed method and the conventional vision-only tracking method is briefly analyzed here. The image processing parts are both employed Sobel edge detection. We set the frequency of cameras as f_{c} (typical value is 30 Hz) and gyroscopes as f_{g} (typical value is 100 Hz), respectively. In order to guarantee the fairness, the tracking time should be set as the same for the two methods, thus we know that the conventional vision-only method would conduct $Nnf_{\mathrm{c}}/f_{\mathrm{g}}$ image processing. As is known, the computational complexity of the Sobel edge detection for an image of $p \times q$ is $O(pq)$. The computational complexity of the gyroscope tracking in each discrete time is $O(M)$. From the above, the computational complexity for the conventional vision-only method is $O(Nnpq)$, while our proposed method is $O\left(N(pq + nM)\right)$. In reality, $pq + nM \ll npqf_{\mathrm{c}}/f_{\mathrm{g}}$, thus our proposed method has an obviously low computational complexity. Generally speaking, our tracking system conducts much fewer image processing problems, as a result of lower complexity.

In addition, the memories occupied in the system for storing the image are much larger than the memories for storing the gyroscope data. Since we just capture the image at discrete time $j = 0$, as a contrast that the conventional vision-only method stores the captured images at each discrete time, our proposed algorithm has a clear advantage.

3 Experiments and results

3.1 Platform setup and parameter setting

We set up a demo to verify our proposed camera pose estimation algorithm, which is shown in Figure 4. The experiment in this paper was conducted on a high-precise, single-axis, variable velocity turntable "AVIC", model 901CF. The object in this experiment was chosen as a landmark which consisted of six black squares and was adhered on the wall in front of the camera, thus the relative motion between the camera and the object depended on the camera. The camera and the gyroscope were fixed together so that the relative motion between the camera and the object was able to be detected during all the tracking process. The two sensors transmit their data to a computer, which processes the sensor fusion part in this experiment. The gyroscope employed in this experiment is a single-axis open-loop fiber-optic gyroscope, whose bias drift stability is 0.02°/h and the angle random walk is $0.002^\circ/\sqrt{\mathrm{h}}$. In addition, we set N=1 in this experiment.

A CMOS camera with the squared pixels, which has a lens with a focal length of 4.8 mm, is used. The captured images are RGB and have a resolution of 1280 × 720. The fiber-optic gyroscope has a sampling rate of 100 Hz, and its measurement range is $\pm 30^\circ$/s. The synchronization of two sensors is controlled by the computer system with a synchronization program. Since the format of the images captured is

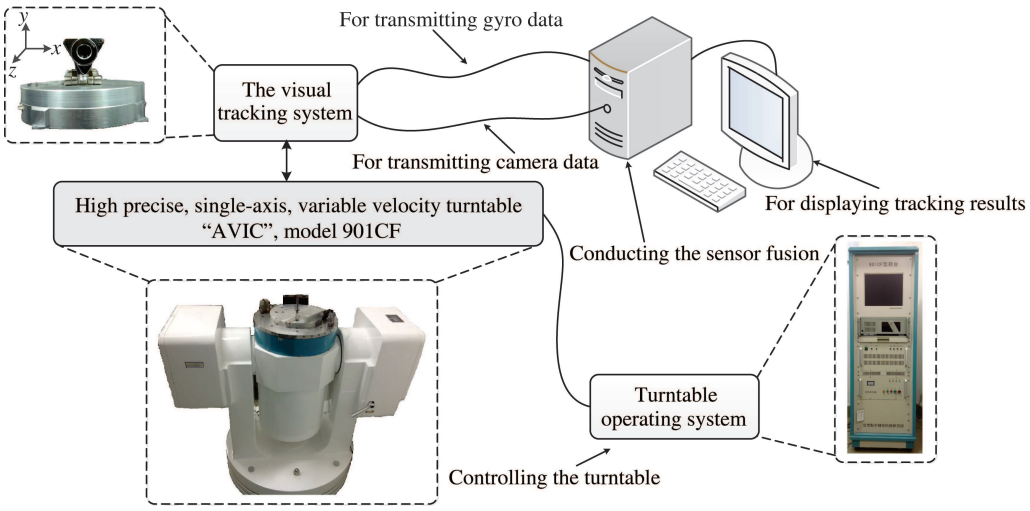

Figure 4 The experiment setup in this paper.

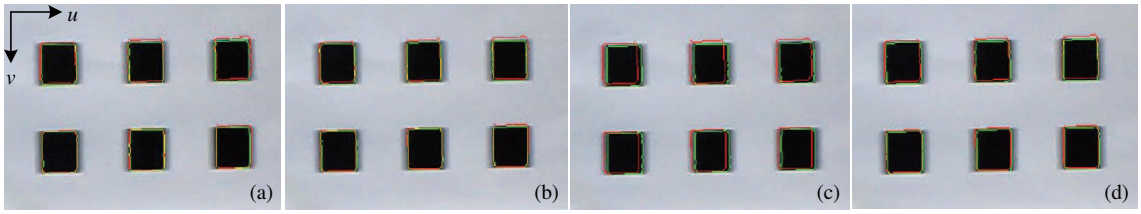

Figure 5 The edge detection results (green) and the proposed sensor fusion tracking results (red) (rate = 15°/s). (a) (b) (c) (d) are 4 tracking results with time interval of 1/3 s.

RGB, we need to convert them to gray-scale in the initialization stage so that the edge detection is easy to implement.

3.2 Test of the visual tracking system

In order to test the efficiency of the algorithm, the camera undergoes a uniform rotation in different speeds. Figure 5 shows the tracking results, when the camera is rotating around the y axis with a rate of 15°/s. The time interval between each scene is 1/3 s and the initialization is conducted only once during the tracking process, i.e. N=1. The green rims are the results by using the conventional edge detection, which is vision-only, while the red rims are the results by using the proposed fiber-optic gyroscope aided fusion method. The main error of the edge detection method is caused by the streaking effect, while the main error of the sensor fusion method is from the inaccuracy of the fiber-optic gyroscope and the image distortion from the camera.

We use the parameter $\mathscr{D}$ to describe the tracking error, which is presented as follows:

$$\mathscr{D} = \sqrt{\frac{\sum_{p=1}^{M}\left[(\hat{u}_p - u_p)^2 + (\hat{v}_p - v_p)^2\right]}{M}}, \tag{13}$$

where u_p and v_p are the real object coordinates, $\hat{u}_p$ and $\hat{v}_p$ are the tracked object coordinates.

In this experiment, since a single-axis gyroscope is used, we can only track the object in the direction of u. Figure 6 shows the comparison of the tracking errors between the conventional vision-only method and the proposed sensor fusion method. From the result, we find that the sensor fusion method has a better performance than the conventional vision-only method. The proposed system can reduce the tracking error by 6.15% in average in this experiment. During experiments, there are some joggles for the camera in the direction of v and some distortion of the images caused by the wide angle lens camera. Based on the error analysis, the performance can be better if the experiment platform is much better.

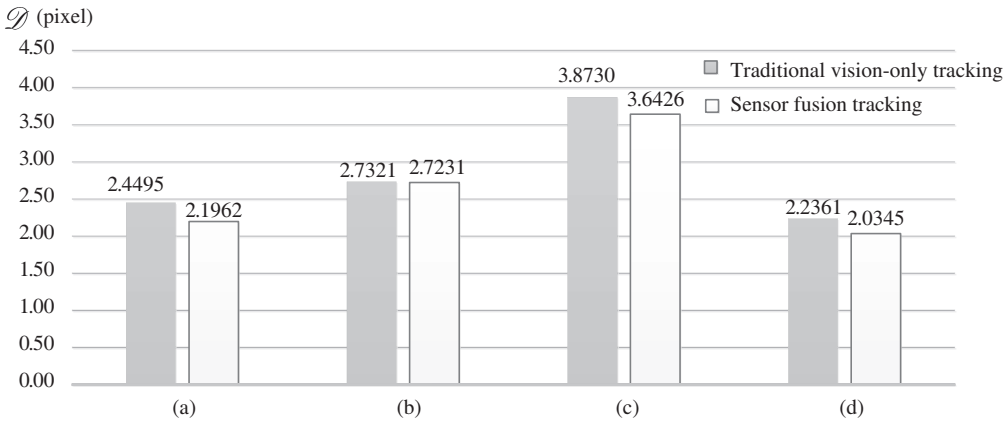

Figure 6 The comparison of the tracking errors between the edge detection tracking and the proposed sensor fusion tracking for the (a) (b) (c) (d) in Figure 5.

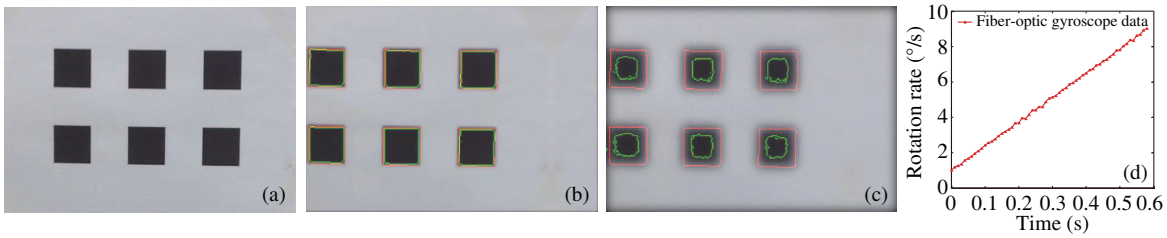

Figure 7 The simulation results between sensor fusion method (red) and the edge detection method (green) in variable rotation rates. (a) Original; (b) variable rotation; (c) variable rotation & camera shaking; (d) the rotation rate.

In addition, we give a simulation result of the tracking in a variable rotation rate. Figure 7(a) is the original object scene. Figure 7(b) is the tracking result after a variable rotation, where Figure 7(d) gives the fiber-optic gyroscope data in a 0.59 s tracking motion. Figure 7(c) shows the same tracking process as Figure 7(b), but has a serious camera shaking. From the simulation results, we know that the sensor fusion method can always get a good tracking result. However, the edge detection method greatly depends on the quality of the image.

We also test the computational complexity of our proposed tracking system. Since the computational complexity of the algorithm is hard to characterize in numerical results, we employ the program running time to describe it. Most of our tracking work on the computer is conducted through Matlab, thus we compare the program running time in Matlab between the two methods. The CPU for our computer is Intel i7-4770, and the Matlab version is R2012a. A 0.5 min visual tracking is implemented, where $f_{\rm c} = 30$ Hz, $f_{\rm g} = 100$ Hz, and $N = 1$. The result is shown in Figure 8, from which we learn that the conventional vision-only method has fewer tracking steps, each of which, however, takes a relatively longer running time, on account of the image processing at every step. As a contrast, our proposed method contains more tracking steps, and the whole tracking stage only takes much less running time, except for the initialization part resulting in a sheer waste of time. In general, the running time of the program, that is, the computational complexity, of our proposed method is 86% lower than the conventional vision-only method in a 0.5 min visual tracking process.

3.3 Test of calibration and long-term tracking performance

In practical application, the long-term tracking performance is more important, which means that the tracking system could track the object successfully for a relatively long period of time. The conventional vision-only tracking method can achieve the long-term visual tracking all the time, but it needs to conduct the quantities of the image processing problems, which has a high computational complexity. In order to guarantee the long-term tracking performance of our proposed system, the calibration for the fiber-optic gyroscope is necessary. For the tracking experiment in Subsection 3.2, none of the calibration is

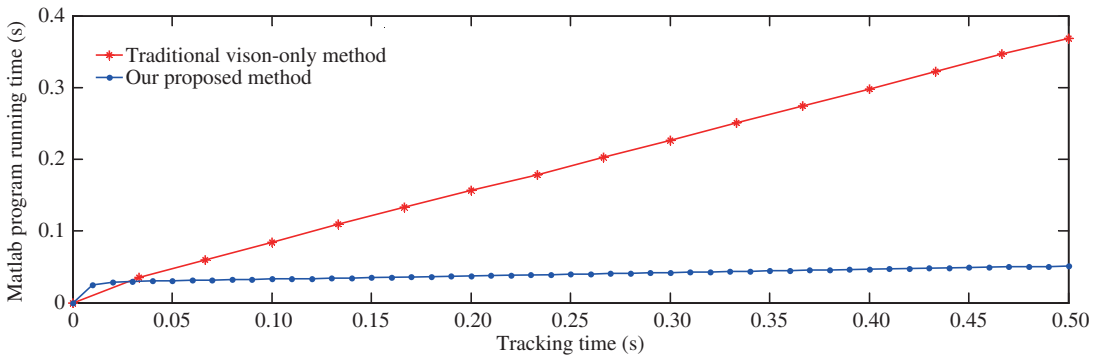

Figure 8 The comparison of Matlab program running time for a 0.5 min visual tracking between two methods.

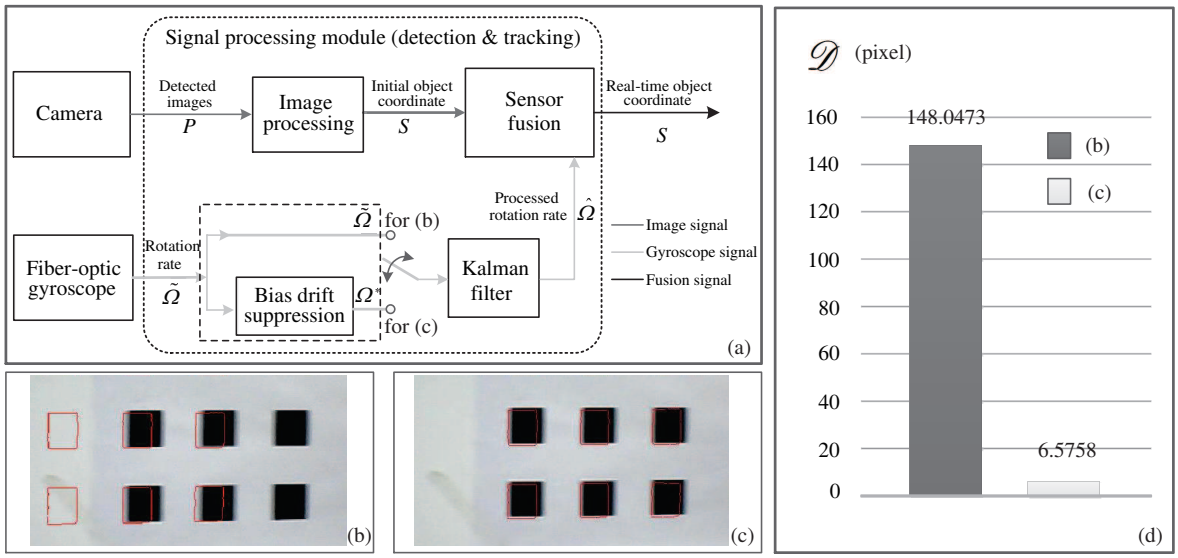

Figure 9 Long-term tracking performance test (tracking time = 20 min, rate = 15°/s). (a) The two tracking schemes; (b) the tracking result without suppressing the bias drift; (c) the tracking result with suppressing the bias drift; (d) the comparison of the tracking errors for the two schemes.

employed on account of the high accuracy of the fiber-optic gyroscope in the short-term tracking. The main error from the gyroscope in a long-term tracking is the accumulated bias drift, which cannot be neglected. Based on the above, our previous work which proposed a signal processing method to suppress the bias drift of the fiber-optic gyroscope can be used to calibrate the bias drift [22]. This method uses the reversing of the piezoelectric modulator to suppress the bias drift, which is a hardware method. In addition, we employ a novel Kalman filter method, which fuses an adaptive Kalman filter and finite impulse response filter in a DSP chip, to eliminate the angle random walk noise to promote the tracking performance in the high rotation velocity, based on our another previous work in [23]. Figure 9(a) shows the two tracking schemes without and with the calibration of the bias drift suppression. Figure 9(b) and (c) show the 20 min tracking results for the two schemes, respectively.

From Figure 9(b) and (c), we find that using the bias drift suppression method to calibrate the gyroscope can help the proposed sensor fusion system realize the superior long-term visual tracking in spite of the errors of a few pixels. Figure 9(d) shows the comparison of the tracking errors in Figure 9(b) and (c).

It is worth mentioning that our proposed system only conducts the initialization once during the 20 min visual tracking. A remarkable advantage for this visual tracking system is its long-term tracking performance, which tremendously reduces the frequency of the image processing and the computational complexity compared with the vision-only tracking method.

4 Conclusion

An efficient fiber-optic gyroscope aided camera pose estimation algorithm based on the sensor fusion is proposed. The angular velocity provided by the gyroscope helps the camera to track the object accurately and the tracking results are better than the vision-only (edge detection) tracking results (reducing the tracking error by 6.15%) by employing the proposed novel sensor fusion algorithm. It is worth mentioning that our proposed algorithm has a relatively lower computational complexity (reducing the program running time by 86% for a 0.5 min visual tracking experiment). Furthermore, the system can track the object in a long term with the help of the bias drift suppression calibration.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant No. 61275005), International S&T Cooperation Program of China (Grant No. 2015DFG12520), National Natural Science Foundation of China (Grant No. 61571379), and State Key Laboratory of Advanced Optical Communication Systems and Networks, China.

Conflict of interest The authors declare that they have no conflict of interest.

References

1 Klein G S W, Drummond T W. Tightly integrated sensor fusion for robust visual tracking. Image Vision Comput, 2004, 22: 769–776
2 Hol J D. Pose Estimation and Calibration Algorithms for Vision and Inertial Sensors. Sweden Linköping: Linköping Univ Press, 2008. 7–23
3 Song S, Qiao W, Li B, et al. An efficient magnetic tracking method using uniaxial sensing coil. IEEE Trans Magn, 2014, 50: 4003707
4 Wang Q, Chen W P, Zheng R, et al. Acoustic target tracking using tiny wireless sensor devices. Inf Process Sens Netw, 2003, 2634: 642–657
5 Zhang X, Hu W, Xie N, et al. A robust tracking system for low frame rate video. Int J Comput Vis, 2015, 115: 279–304
6 Zhang X, Hu W, Maybank S, et al. Sequential particle swarm optimization for visual tracking. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, Anchorage, 2008. 1–8
7 Li Y, Ai H, Yamashita T, et al. Tracking in low frame rate video: a cascade particle filter with discriminative observers of different life spans. IEEE Trans Anal, 2008, 30: 1728–1740
8 Park I K, Singhal N, Lee M H, et al. Design and performance evaluation of image processing algorithms on GPUs. IEEE Trans Parall Distr, 2011, 22: 91–104
9 Erdem A T, Ercan A O. Fusing inertial sensor data in an extended Kalman filter for 3D camera tracking. IEEE Trans Image Process, 2015, 24: 538–548
10 Foxlin E. Inertial head-tracker sensor fusion by a complementary separate-bias Kalman filter. In: Proceedings of IEEE Virtual Reality Annual International Symposium, Santa Clara, 1996. 185–194
11 Hol J D, Dijkstra F, Luinge H, et al. Tightly coupled UWB/IMU pose estimation. In: Proceedings of IEEE International Conference on Ultra-Wideband, Vancouver, 2009. 688–692
12 He P, Cardou P, Desbiens A, et al. Estimating the orientation of a rigid body moving in space using inertial sensors. Multibody Syst Dyn, 2014, 35: 1–27
13 Corke P, Lobo J, Dias J. An introduction to inertial and visual sensing. Int J Robot Res, 2007, 26: 519–535
14 Starner T. Project glass: an extension of the self. IEEE Pervas Comput, 2013, 12: 14–16
15 Chai L, Nguyen K, Hoff B, et al. An adaptive estimator for registration in augmented reality. In: Proceedings of 2nd IEEE and ACM International Workshop on Augmented Reality, San Francisco, 1999. 23–32
16 You S, Neumann U. Fusion of vision and gyro tracking for robust augmented reality registration. In: Proceedings of IEEE Conference on Virtual Reality, Yokohama, 2001. 71–78
17 Zhang G. The Principles and Technologies of Fiber-Optic Gyroscope. Beijing: National Defense Industry Press, 2008. 1–25
18 Barbour N, Schmidt G. Inertial sensor technology trends. IEEE Sens J, 2001, 1: 332–339
19 Hwangbo M, Kim J S, Kanade T. Gyro-aided feature tracking for a moving camera: fusion, auto-calibration and GPU implementation. Int J Robot Res, 2011, 30: 1755–1774
20 Yu J, Wang Z F. 3D facial motion tracking by combining online appearance model and cylinder head model in particle filtering. Sci China Inf Sci, 2014, 57: 029101
21 Gonzalez R C, Woods R E, Eddins S L. Digital Image Processing Using Matlab. Beijing: Publishing House of Electronics Industry, 2014. 205–229
22 Wang X, He C, Wang Z. Method for suppressing the bias drift of interferometric all-fiber optic gyroscopes. Opt Lett, 2011, 36: 1191–1193

23 He C, Yang C, Wang Z. Fusion of finite impulse response filter and adaptive Kalman filter to suppress angle random walk of fiber optic gyroscopes. Opt Eng, 2012, 51: 124401

Appendix A The principle of image formation in camera

This Section exhibits the principle of image formation in camera, which is the detailed derivation of Eq. (4).

Video images are acquired from the video capture hardware which connected with the camera. The pin-hole model is commonly used to describe the images acquired by cameras. Figure A1 shows the model, where the camera acquires digital images on the image plane, which is described by the pixel coordinate system and the retina coordinate system. The unit length of the former is measured by pixels and that of the latter is measured by physical length such like millimeter. The origin of the pixel coordinate system O_p usually set at the vertex of the image and u-axis is parallel to x-axis in the camera coordinate while v-axis is parallel to y-axis, just as Figure A2 shows. According to the model, the origin of the retina coordinate system O_r is in the center of image $(u_0, v_0)^{\mathrm{T}}$, and it is on the z-axis of the camera coordinate system. There is a matrix that transfers the camera coordinates $(x_{\mathrm{c}}, y_{\mathrm{c}}, z_{\mathrm{c}})^{\mathrm{T}}$ to the retina coordinates $(x_{\mathrm{r}}, y_{\mathrm{r}})^{\mathrm{T}}$, just as follows:

$$\begin{bmatrix} x_{\mathrm{r}} \\ y_{\mathrm{r}} \end{bmatrix} = \frac{1}{z_{\mathrm{c}}} \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \end{bmatrix} \begin{bmatrix} x_{\mathrm{c}} \\ y_{\mathrm{c}} \\ z_{\mathrm{c}} \end{bmatrix}, \tag{A1}$$

where f is the focal length of the camera. The retina coordinate system as well as the pixel coordinate system is a 2-D coordinate system. They are unable to acquire the depth. The relation between the retina and pixel coordinate system is as follows:

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} 1/du & 0 & u_0 \\ 0 & 1/dv & v_0 \end{bmatrix} \begin{bmatrix} x_{\mathrm{r}} \\ y_{\mathrm{r}} \\ 1 \end{bmatrix}, \tag{A2}$$

where du and dv are the unit physical length of pixels in the direction of u-axis and v-axis.

Therefore, by using Eqs. (A1) and (A2), the positions of edges are given by pixel coordinates $(u, v)^{\mathrm{T}}$ and there is a matrix to transfer them to camera coordinates:

$$\begin{bmatrix} u \\ v \end{bmatrix} = \frac{r'}{r} \cdot \frac{1}{z_{\mathrm{c}}} \begin{bmatrix} f_u & 0 \\ 0 & f_v \end{bmatrix} \begin{bmatrix} x_{\mathrm{c}} \\ y_{\mathrm{c}} \end{bmatrix} + \begin{bmatrix} u_0 \\ v_0 \end{bmatrix}, \tag{A3}$$

where f_u and f_v are the pixel length of the focal length, which are defined by

$$f_u = f/du, f_v = f/dv. \tag{A4}$$

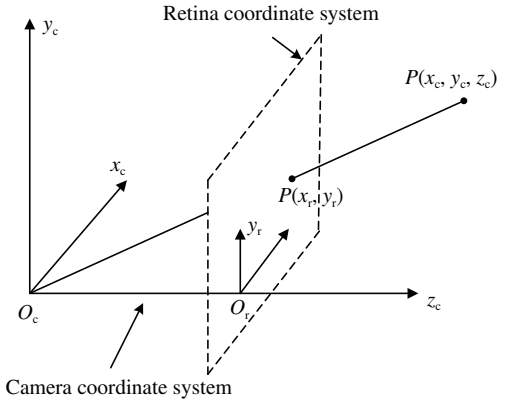

Figure A1 The relationship between the camera coordinate system and retina coordinate system.

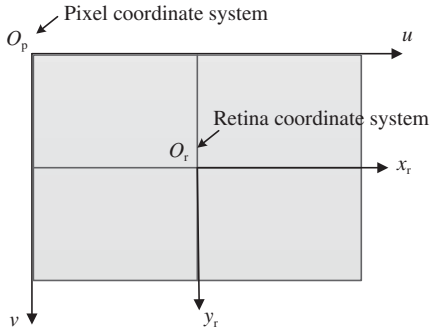

Figure A2 The relationship between the retina coordinate system and pixel coordinate system.

Since the images which are acquired by wide angle lens camera exhibit substantial radial distortion, a standard pin-hole model cannot be used directly. The distort needs to be approximated and corrected by normalizing the radius in camera coordinate [1]:

$$r' = r + \alpha r^3 + \beta r^5, \tag{A5}$$

where

$$r = \sqrt{\left(\frac{x_{\mathrm{c}}}{z_{\mathrm{c}}}\right)^2 + \left(\frac{y_{\mathrm{c}}}{z_{\mathrm{c}}}\right)^2}. \tag{A6}$$

In order to simplify the Eq. (A3), we define $\gamma = r'/r$ in this paper. For this visual tracking system, the f_u, f_v, u_0, v_0, α and β are the camera parameters and they can be calibrated on-line. In addition, $\alpha = -0.3$, β=0.06 in the tracking system proposed.