A Comparative Study on Pose Estimation for Monocular Vision and Binocular Vision Without Modeling

Lei He
Graduate School of the Chinese Academy of Science
Shenyang Institute of Automation, Chinese Academy of Science
Shenyang City, Liaoning Province, China

Feng Zhu, Yingming Hao
Shenyang Institute of Automation, Chinese Academy of Science
Shenyang City, Liaoning Province, China

Abstract - Pose estimation algorithms with regard to both monocular vision and binocular vision have been studied respectively since about twenty years ago. However, few researchers considered comparing two representative methods for the sake of actual application. In this paper, considering the actual model, we have done a comparative study about these two methods quantitatively. The statistical result of calculated error based on simulation experiments denotes that the model-based monocular vision behaves more outstanding than the binocular vision on the computational accuracy and robustness. Inasmuch the monocular vision introduces the known object model in the process of computing the coordinates of feature points. The simulated result has certain guided significance for choosing a suitable pose estimation algorithm in actual engineering projects.

Keywords: pose estimation monocular vision binocular vision model-based.

1 Introduction

In actual project application, measuring the position and attitude between two objects, commonly named pose estimation, usually should be achieved. The mathematical depict of pose estimation is to measure six transformation variants which contain three position parameters and three rotational ones between two reference frames. Theoretically, once the coordinate of at least three feature points respectively in two reference systems have been known, the unique transformation can be obtained by only effortless matrix operation. Therefore, the pivotal problem is how to obtain the coordinates of features points in above two reference frames.

Pose estimation in computer vision, which developed as a typical non-touched metrical method, has been applied in many fields such as industry, medical treatment and aviation. The simplest pose estimation process can be implemented as follows. In the beginning, setting an array of feature points in the object (means target that will be observed) and obtaining the precise 3D coordinates of these points in the objective reference frame. Then, equipping one or several cameras on another object (what is called observer) and getting the coordinates of imaging points in image reference frame by imaging process appropriately. Finally, acquiring the 3D coordinates of feature points in the reference system that can be called metrical reference system by means in computer vision and realizing the measurement of relative position and attitude between objective reference system and metrical reference one.

In order to obtain the coordinates of feature points in observer reference system, two representative methods exist. One is monocular vision that means using only one camera for imaging and combining the inherent geometrical constraints related to objective feature points. This method was called PNP problem, which is an active research facet in this domain. The other more representative method is binocular vision, called stereo vision commonly. In this method, a couple of cameras is supposed to be adopted, that is to say, the coordinates of feature points in metrical reference can be calculated by triangulation. The model based monocular vision only use one camera, so the measurement system should be simpler, but the geometrical constraint must be known. On the other hand, the binocular vision need not to obtain the geometrical constraint and can calculate the coordinate of objective points directly but the system is more complex. In the domain of pose estimation, the coordinates of at least three points at observed object must be known so as to realizing the calculation of position and attitude, therefore, the priority of stereo vision in this facet should not exist. When considering how to select the metrical methods, what we definitely should focus on the metrical precision and robustness of diverse methods. If these methods perform tolerably, maybe the model based monocular vision method should be preponderant.

The metrical precision of measurement system is mainly determined by two aspects that are the error of input parameters and how the system transfers this error. In view of above two methods, their input parameters
consist of three categories as follows. One is the 3D coordinates of features points in model reference frame which also be called objective reference frame thereafter, the other is the 2D imaging coordinates of feature points in image reference system and another is model parameters of camera. The precision analysis of the measurement system is to analyze the relative position and attitude corresponding to the error of these input parameters. In addition, the robustness of the system denotes the sensitivity of metrical result corresponding to the error of input parameters, it is correlative with the design of the measurement system and the metrical methods directly.

Presently, the binocular stereo vision is dominant and the relative research concentrates on the relationship between metrical error of feature points and system parameters. Jeffrey et al.\textsuperscript{[3]}and Chiencuang Chang et al.\textsuperscript{[4]} studied the relationship between calculated precision and measurement system parameters with regard to the quantization error. Liu et al.\textsuperscript{[5]} considered the diverse origins of error and mainly analyzed the relationship between the metrical error of 3D coordinates and system design. With regard to the model based monocular vision, the correlative research mainly focus on calculative methods. Owing to the complexity of solution, the simulation experiments are leading analytical means. Larry Davis\textsuperscript{[6]} gave the simulation results of pose estimation for Marker-based Tracking. For the comparative study, only Hassan Mostafavi\textsuperscript{[7]} got the conclusion based on the rigid body that the precision of monocular vision is comparative with the binocular stereo vision except the offset error in the same direction with optical axis.

In this paper, with actual application of pose estimation as the backdrop, we did the comparative study of model base monocular vision and binocular stereo vision under some given limitative premises. The given premises are accord with the accustomed disposal method that will be introduced next.

We begin with a review of previous work on pose estimation of both monocular and binocular vision in early vision. In section 2, we give our calculative pose estimation which are adopted for simulation. In section 3, we set out our analytical method and simulation condition. In section 4, final experimental results and relevant analysis are presented. The conclusion is completed in last section.

2 Calculative methods of pose estimation

For the sake of obtaining the relationship between two reference frames, the coordinates of more than three points that are not on a line completely should be acquired, then we can calculate the relative position and attitude by using the rotation theory.

Model base monocular vision and binocular stereo vision adopt different methods for calculating the 3D coordinates in metrical reference system.

2.1 Model based monocular vision

Model base monocular vision system only equips one camera for image acquisition, the constraint among feature points of objective model, concerned in last section, must be considered for calculating 3D coordinates in metrical reference system, hence, the calculative result satisfies the objective model exactly.

In this paper, using the optimal dichotomy which has been presented in paper\textsuperscript{[8]} and combining the imaging coordinates and known model constraint, we can calculate the 3D coordinates of three feature points in metrical reference system in good time.

2.2 Binocular stereo vision

Using binocular stereo vision method, we can calculate the 3D coordinates of feature points at first hand with the known constraint of two cameras. If we consider the model constraint simultaneously, the redundant information, which means at least three more equations, should exist. When the quantization error of imaging points is considered, it can be ensured that no exact solution can be derived, in this condition, we commonly pursue the optimal solution under a certain optimization function that makes sense considerably. Utilized appropriately, the redundancy of acquired information can enhance the robustness of system definitely. For instance, easily implementing the P3P method for two cameras respectively and taking the average of two results as the final one, we can obviously improve the precision statistically. The comparative study between model based binocular vision optimization algorithm and monocular vision method has been done in our paper\textsuperscript{[9]}.

In this paper, we mainly concentrate on metrical precision of binocular stereo vision that means we cannot introduce the objective model into the calculation of 3D coordinates in metrical reference frame. To begin with, in conjunction with the imaging coordinates of three objective feature points, the 3D coordinates of feature points $P_{vl} = (x_{vl}, y_{vl}, z_{vl})$ in metrical reference system can be obtained by triangulation, which was known as a typical stereo vision method. Considering the quantization error of imaging coordinates, the coordinates calculated by triangulation surely dissatisfy the model constraint so that the relative position and attitude cannot be solved by rotation theory. Therefore, the obtained 3D coordinates
must be modified to satisfy the objective model constraint. In this paper, with the objective model as the constraint and the square of distance between \( r'_w = (x'_w, y'_w, z'_w) \) and \( r_w \) as the optimization function, the modified points \( r'_w \) which satisfy can be obtained by least square method.

### 3 Analytical method and simulation

#### 3.1 Model based monocular vision

In the calculative process of both model based monocular vision and stereo vision, iterative computation and nonlinear optimization exist definitely. The metrical results are hard to be analytically expressed with the error of input parameters, therefore, the theoretical analysis are difficult to realize. In this paper, experimental method has been adopted, when acquiring the simulated result, we give an intuitionistic explanation about the comparative study. As the metrical error is correlative with both input parameters and the relative position and attitude between object and camera, the comprehensively comparative study can be accomplished statistically. The statistical experiment needs a lot of experimental data, which is nearly impossible to be obtained by real experiment. Accordingly, simulated experiment was considered in this paper.

#### 3.2 Experimental condition

The calculative precision of monocular vision relates to diverse facets that contain not only quantization error of imaging points but also the geometrical relationship of objective points, the current position and attitude of object, intrinsic and extrinsic parameters of camera and so on. With a view to binocular stereo vision, besides above mentioned factors, the calculative precision should be influenced by length of baseline, the longer baseline means higher precision. Two methods are hard to be compared impartially, for general case, the comparison has been completed under a array of given basic items.

![Image of simulated model of stereo vision](figure 1. simulated model of stereo vision)

Figure 1 denotes the simulated model of binocular stereo vision. Supposing that the view angle of camera be known, the objective model, composed with three feature points, should exist right in the common view area of two cameras through precise calculation. The metrical reference system is \( O, X_w, Y_w, Z_w \), the baseline are the same as axis \( Z_w \) and the length is 0.4m, resolving power is 512×512, the intrinsic parameters are as follows. The principle point is \( u_0 = v_0 = 256 \), focus ratio is \( f \). The monocular vision simulated condition is similar with this, the only difference is the number of cameras is one, so the rationality and equity can be guaranteed.

Combining the actual application model, three objective feature points consist of a triangle ABC as shown in figure 2 denotes \( AB=0.8m, BC=0.5m, AC=0.5m \). The midpoint of \( AB \), namely point \( O \) denotes the origin of objective reference system, the line \( AB \) is in the same direction with axis \( Z \) while the line \( OC \) is in the same direction with axis \( X \). The coordinates of three feature points in objective reference system are \( A(0,0,0.4), B(0,0,0.4), C(0.3,0,0) \).

For simplicity, only the quantization error of input parameters is considered. Due to the same condition for the comparison, our hypothesis, which can be considered appropriately, cannot influence the basic conclusion. In the simulation, we set 0.25 pixels as the quantization standard.

![Image of objective model of three feature points](figure 2. objective model of three feature points)

In the simulation, through stimulant disposal of quantization error, the coordinates of imaging points in image reference system can be obtained. Using above mentioned two representative methods, we can get the coordinates in objective reference frame and metrical reference system respectively. The moving range along the optical axis orientation is \( 4m - 20m \) according with the actual application, then the moving range along \( Y \) and \( Z \) axis can be calculated based on the view angle of camera.
we suppose the variational range of attitude angle be 
\(-6 \sim 6^\circ\). In above defined area, through changing the 
position and attitude of objective reference system, we 
simulated the process of pose estimation in different 
condition and the statistical experiment can be realized.

4 Experimental results and analysis

Figure 3 gives the comparative results of model based 
monocular vision and binocular vision without modeling, 
the square sign denotes the binocular stereo vision while 
the arris sign denotes the monocular vision. The x-axis 
shows the distance from object to camera while y-axis 
denotes the statistical mean square root error of six degrees 
of freedom (EOF).

It can be seen from the figure 3 that under the 
condition that presented in this paper, the metrical 
precision of monocular vision perform more outstanding 
than binocular stereo vision mainly because to the latter, 
the objective model constraint was not introduced in the 
calculative process of the coordinates of feature points in 
metrical reference system. Taking the attitude error into 
count, the binocular stereo vision method perform too pool 
to be applied in the some project application especially 
when the distance of camera and object increases, the 
result reflects the inherent malpractice that exists in stereo 
vision without introducing model constraint. When 
considering the error of position, in the condition of 
reasonable length of baseline, the monocular vision 
method is comparative with binocular stereo vision, even 
in our simulation, it perform slightly better than the latter. 
Theoretically speaking, when the baseline becomes longer, 
the binocular vision method definitely perform better and 
in certain existed case, it will indeed performs better than 
the monocular vision method. But in actual application, 
the longer baseline means the more spatial cost and less 
common view area of two cameras and it is impossible for 
us to lengthen the baseline without limit.

In addition, the robustness of binocular stereo vision 
perform badly especially the distance from camera to 
object increases because the error increasing speed 
aggrandize rapidly, on the other hand, the monocular 
method performs more steadily.

Figure 3. simulation error of EOF in two different methods

5 conclusion

In this paper, combining the actual model, two 
representative methods of pose estimation have been 
compared in relatively fair condition through massive
simulated experiments. It can be concluded from simulated results that in most actual engineering applications, as the objective model is and must be introduced in the process of calculating the coordinates of feature points, it performs more outstanding than binocular stereo vision on calculative precision and robustness. This conclusion is significant for choosing designing vision based metrical system in actual application especially when the objective model can be set randomly and more concise system are required.

6 Acknowledgement

This work is supported by National Science Foundation of P.R. China (60575024).

7 reference


