The Research of Binocular Stereo Measuring System

Ting SUN, Fan-liang BU and Meng-qi YUAN
Police Information Engineering of People’s Public Security
University of China, Beijing 102623, China

Keywords: Binocular stereo vision, feature point extraction, stereo matching, 3D data acquisition.

Abstract. This paper mainly focuses on the binocular stereo vision system. The system gains a set of pixels corresponding to the spatial scene points through the process of feature point detecting based on SIFT and stereo matching, so as to calculate the three-dimensional coordinate points of spatial scene points and realize the target distance measurement. This paper basically realized the experiment purpose and has achieved good results.

Introduction

Nowadays, binocular stereo vision measuring technology has been the hotspot and difficulty in the field of computer vision research. Since it has the advantages of simple structure and convenient operation, it has been widely used in industrial inspection, robot visual navigation and positioning, target recognition and other fields.

The binocular stereo vision system is based on the analysis of two cameras which have certain geometric relationship. The two cameras shoot the same scene at the same time to get two images of such scene, thus obtaining the 3D geometric information of objects in the space. With the great improvement of hardware processing performance and the further development of computer image processing technology, the accuracy of binocular vision measuring has been significantly improved, the operation of experiment has been simplified and the repeatability has been also improved, so such method has becomes one of the most popular vision measurement methods.

This paper simply introduces the binocular stereo vision system, studies the mathematical model as well as its algorithm, and performs some experiments according to this, thus realizing the basic function.

Binocular Stereo Measuring System

The binocular stereo vision system generally uses two cameras which have completely same structure and performance, and the two cameras have to be put symmetrically. This article uses a camera with tripod head. We realize the effect of two cameras by moving and rotating. Using one camera can fix the relative position of the two, thus making the experiment more repeatable.

The model of binocular stereo vision system based on pinhole imaging is shown in Figure 1. In the figure, \( O_1 \) and \( O_2 \) are respectively the origins of two cameras coordinate systems. Space point \( P \) images on the image planes of two coordinate systems respectively, and its image points are \( P_1 \) and \( P_2 \). The binocular stereo measuring system finds the image points \( P_1 \) and \( P_2 \) of spatial scene points by stereo matching, and calculates the three-dimensional coordinates of \( P \) through a series of image processing operations, so as to realize the 3D spatial distance measurement.
The binocular vision measuring system is mainly composed of four parts: camera calibration, feature point extraction, image stereo matching and three-dimensional coordinate acquisition.

**Camera Calibration**

Camera calibration studies the pinhole imaging principle and considers the distortion existing in the process of imaging by making use of the checkerboard calibration board to obtain the performance parameters of the camera as well as the position parameters of cameras before and after moving, namely the camera internal and external parameters, which provides strong data support for the subsequent 3D coordinate acquisition.

**Feature Point Extraction**

**Feature Point Detection:** In SIFT algorithm, we can perform the convolution operation by using Gaussian function of variable scale \( G(x, y, \sigma) \) and the image \( I(x, y) \) to obtain the image scale space. The following function \( L(x, y, \sigma) \) represents the scale space, namely

\[
L(x, y, \sigma) = G(x, y, \sigma) \otimes I(x, y) \tag{1}
\]

SIFT algorithm firstly performs the subtraction operation on the image which is adjacent to the Gaussian scale space, thus will get a response image \( D(x, y, \sigma) \) of DoG (Difference of Gaussians). And then the image is performed local maximum search to determine the specific location of feature point in position space and scale space \([1]\). Among them,

\[
D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) \otimes I(x, y) \tag{2}
\]

In the formula, \( k \) is constant, which represents the ratio between adjacent scale spaces \([2]\).

**SIFT Feature Descriptor and Feature Vector**

SIFT feature descriptor is used for image feature detection. Through the description of image local characteristics, they can be processed quantitatively, so as to objectively reflect the image feature distribution. We can get a good stability direction in the vicinity of feature points by using the histogram of image gradient to realize the image rotation invariance. The detected scale value according to the feature points is determined, and the Gaussian image which is the most close to such value can be calculated by the following formula

\[
L(x, y) = G(x, y, \sigma) \ast I(x, y) \tag{3}
\]

We draw a circle taking feature point as the center and \( 3 \times 1.5\sigma \) as the radius, calculate the direction and size of the image gradient in circular area, and then obtain the gradient direction and size of the image in the neighborhood area by using histogram. The 360° of gradient direction is equally divided within the scope of 8 directions, then the gradient size within each scope is accumulated, and the direction with the maximum amplitude is chosen as the principal direction of the gradient. In order to enhance the system robustness, the direction corresponding to the peak
whose energy value equals to the 80% of the main peak value is chosen as the auxiliary direction of the feature point, thus using the above two directions to jointly describe the feature points.

SIFT descriptor can be expressed in a vector form to describe the statistical results of Gaussian image gradient which is in the neighbor domain of feature points. As is in the neighborhood 4 \times 4 area of feature point, the gradient direction is divided into 8 scopes, so the feature vector is 128-dimensional, which directly determines the running speed of the system, namely close to real-time processing.

**Image Feature Points Matching**

If feature vectors of the two images corresponding to the same scene are close to each other, then we consider it as the two feature points which are corresponding to the same spatial point; on the contrary, we consider it in different positions. Feature point matching technology is to find the nearest two feature points from feature points sets in the two images which corresponding to the same scene, so as to realize the feature points one-to-one matching.

The paper adopts the kd-tree algorithm to match the feature points. Data query in kd-tree is to retrieve the data points which are closest to query data in kd-tree data set. The search is according to the binary tree structure, along the search path, so as to find the nearest neighbor approximation point, namely the leaf node which concludes the query point. However, the leaf node found in this way is not necessary the nearest neighbor point, for the nearest neighbor point must be closer to the query point, that is to say, the nearest neighbor point must be in the circular domain which is centered in the query point and through the leaf node as the radius. This needs us to perform backtracking operation along the search path to see whether there is a data point which is closer to the query point. If there exists such a point, then we will replace the leaf node by the newly found data point; otherwise, we consider the found leaf node be the nearest neighbor point, completing the feature point matching.

**Stereo Matching**

Stereo matching[3] is the process of finding pixel points in the image to be matched which corresponding to any pixels in the reference image, namely finding the pixel points which are corresponding to any spatial scene point in two images. These pixel points could be the feature points that have been detected, and they could also be the common pixel points. As the feature points detected using SIFT algorithm are invariable and are in a relatively stable area in the image, therefore we can establish a relationship between the pixel points to be matched and the surrounding feature points to realize the matching between pixels.

For stereo matching, in addition with the help of feature points matching, it also has some constraint conditions itself: 1) Uniqueness constraints. Each reference pixel exists and only exists one pixel to be matched; 2) Sequential consistency constraints. The two pixels on the same scanning line are in the same position order in reference image and the image to be matched: 3) Consistency constraints. In two images, the corresponding pixels have the same parallax for the same unobstructed scene. We can further optimize the stereo matched points pair by using the stereo matching conditions, so as to achieve better matching effect.

**3D Data Acquisition**

The binocular vision measuring system involves three coordinate systems: the left camera coordinate system, the right camera coordinate system and the world coordinate system. To improve the portability of binocular data measuring system, we make the left camera coordinate system and the world coordinate system coincide. According to the principle of ideal lens imaging, the relationship between spatial point \( P(X, Y, Z) \) and the corresponding image point \((x,y)\) can be represented as:
\[
\begin{bmatrix}
  x \\
  y \\
  z
\end{bmatrix} = 
\begin{bmatrix}
  f & f & 1 \\
  r_1 & r_2 & r_3 & t_1 \\
  r_4 & r_5 & r_6 & t_2 \\
  r_7 & r_8 & r_9 & t_3
\end{bmatrix}
\begin{bmatrix}
  X \\
  Y \\
  Z \\
  1
\end{bmatrix}
\] (4)

In the formula, \( f \) is the focal length of camera, and \( c \) represents the internal parameter of camera. \( r = [r_1, r_2, r_3, r_4, r_5, r_6, r_7, r_8, r_9] \) is the rotation matrix of world coordinate system to right camera coordinate system, and it is also the external parameter of camera. In this paper, the camera planes before and after rotation and lifting movement are approximately in the same plane, therefore the above matrix is similar to the identity matrix. \( t = [t_1, t_2, t_3] \) is the translation matrix of world coordinate system to right camera coordinate system. There only exists the change in vertical direction before and after moving in theory. \( c \) is a nonzero constant.

One scene point corresponds to two image points, namely there exist two pairs of image point coordinates. In a similar way, we can get the corresponding equations. By solving the simultaneous equations and using the least square method, we can obtain the spatial coordinate of the scene point.

**Experimental Results and Analysis**

In this paper, we use some experimental equipment: a laptop computer, a Nikon camera whose facial length is 85mm, an intelligent tripod with a pan-tilt.

Fix the camera on the intelligent tripod, and set up the planning parameters to make the tripod do fixed length of 20cm up-and-down movement. The camera shoot the two images in the same scene one of the two images is the scene before moving, the other is the scene after moving.

![Figure 3. Two images of the same scene.](image)

We apply the above algorithm to do a series of image processing on the obtained images. Observing Figure 4, it’s to find that the black spot is the feature points detected, the white line is the connecting line of the matching feature points and the red line is the connecting line of the target points to be measured. From the Figure 4, we can observe that the algorithm of feature points detected a lot of corners and the inflection points whose distributions were centered. In the aspect of matching feature points, we also correctly realize the matches of most of feature points. The point manually selected can recognize the corresponding point automatically in the other image, but the accuracy of the matching needed to be improved.
Form the Table 1, the absolute difference between the true value and experimental results and the relative error are acceptable, we can believe that the experimental results are reliable. With the distance between the camera and the object to be measured becoming farther and farther, the relative error is increasing.

In the experiment, because of the friction and some other reasons during the processing of movement, the position of the experimental equipment hardly tallied with the expected, that is to say, the two image planes are not in the same plane, that also can be found from the horizontal ordinate of the reference pixel and that of the pixel to be matched, which makes the results of error. In addition, some other factors also influence the accuracy of the results, such as the limitations of the device, the imaging distortion of camera, the error in the stereo matching, the error in the least squares method. There are a lot of things to decrease the relative error.

Table 1. The results of experiments.

<table>
<thead>
<tr>
<th>Tests</th>
<th>Coordinates of reference pixel</th>
<th>Coordinates of matching pixel</th>
<th>Expert-me</th>
<th>True value</th>
<th>Absolute differences (cm)</th>
<th>Relative differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>first group</td>
<td>(320.75, 347.75)</td>
<td>(319.8, 539.62)</td>
<td>26.66</td>
<td>26.85</td>
<td>0.19</td>
<td>0.71%</td>
</tr>
<tr>
<td>second group</td>
<td>(317.75, 386.75)</td>
<td>(316.84, 580.97)</td>
<td>27.48</td>
<td>27.29</td>
<td>0.19</td>
<td>0.69%</td>
</tr>
<tr>
<td>third group</td>
<td>(337.25, 311.75)</td>
<td>(329.83, 511.90)</td>
<td>24.15</td>
<td>23.85</td>
<td>0.3</td>
<td>1.25%</td>
</tr>
<tr>
<td>fourth group</td>
<td>(511.25, 115.25)</td>
<td>(514.81, 240.94)</td>
<td>18.58</td>
<td>18.30</td>
<td>0.28</td>
<td>1.53%</td>
</tr>
</tbody>
</table>

**Conclusion**

In the criminal scene, there are a large number of the artificial objects, whose features are obvious. This paper introduces the SIFT, studies the methods of detecting feature points, stereo matching and calculating 3D measurement data and so on to achieve the targets of the experiments. The results of the experiments are acceptable. But there have been imperfections in this paper, the precision of the results are to be improved.

**References**
