A Fast Stereovision Measurement Algorithm Based on SIFT Keypoints for Mobile Robot

Haitao Song, Han Xiao, Wenhao He, Feng Wen and Kui Yuan

Institute of Automation
Chinese Academy of Sciences
95 Zhongguancun East Road, Beijing, China
haitao.song@ia.ac.cn

Abstract - When a robot moves independently, it needs to localize itself and avoid obstacles. In order to enable the robot to obtain the depths of surrounding objects, a fast stereovision measurement algorithm based on SIFT keypoints is proposed and implemented on an embedded image processing board based on DSP and FPGA. The calculation of the keypoint descriptor is simplified by directly using the pixel values in the block neighborhood of a keypoint, resulting in an improvement in real-time performance. In order to guarantee the accuracy of searching for the matching points, a two-pass searching strategy is employed, that is, to choose a point in the left image and search the candidate point in the right image, then determine the correspondence by searching the matching point in the left image again. Moreover, for improving the precision of the measurement results, a quadratic polynomial is introduced to compensate the measurement results. Measurement results and comparison of different methods demonstrate the effectiveness of the proposed algorithm. In addition, the measurement system is applied to a mobile robot platform, and the experiment results validate that the system can satisfy the demand of robot applications.

Index Terms - Robot, Stereovision, SIFT, Block descriptor.

I. INTRODUCTION

Localization and obstacle avoidance require a robot to sense the environment. With the advantages of low cost and real-time performance, stereovision measurement method which can acquire the distance between a robot and an object accurately is widely used in the robotics domain. In such systems, how to determine the correspondence between points from two cameras is a key issue.

Reference [1, 2] proposed a method to extract local features, which is named SIFT (Scale Invariant Feature Transform). The algorithm searches keypoints in the image scale space and extracts locations that are invariant with respect to image translation, scaling, and rotation. The SIFT features are individually discriminating and are highly robust to perspective and illumination changes. Besides, the feature can remain invariant in messy scenes. All of the above advantages are good for matching. On account of large computation, the SIFT is difficult to be applied in the system where the real-time performance is critical. Therefore, a variety of improvements of SIFT have been proposed. Reference [3] proposed an algorithm which uses PCA (Principal Components Analysis) to generate the keypoint descriptor. Reference [4] proposed an algorithm based on a hybrid of Harris operator and SIFT descriptor. Reference [5] employed GPU to compute the SIFT descriptor, so the processing speed is significantly improved. Reference [6] improved the processing speed by implementing the SIFT on an embedded system which mainly contains a DSP and a FPGA.

SIFT is widely applied by many researchers for stereovision in various situations [7-10]. However, in the stereovision system, there is no obvious rotation and viewpoint changes between two images captured by two cameras, so we propose a SIFT-based fast stereovision measurement algorithm which generates descriptors by using the pixel values in the block neighbourhood of the SIFT keypoints.

This paper is organized as follows. Section 2 simply introduces our measurement system. Section 3 explains the calibration of the system. Our SIFT-based fast stereovision measurement algorithm is introduced in detail in Section 4. Section 5 provides specific experimental results. Finally, we draw a conclusion of this paper in Section 6.

II. SYSTEM SUMMARY

The experimental system shown in the left plane of Fig. 1 is composed of two CCD cameras. The two cameras capture two images of the same scene respectively, and then the three-dimensional information is obtained based on triangulation by the matching keypoints between two images.

Fig. 1 Left: the system of stereovision measurement. Right: physical diagram of the embedded image processing board.

Image processing, keypoints detection and 3D coordinate calculation are realized on the embedded image processing board. The right plane of Fig. 1 shows the physical diagram of the image processing board, it mainly contains FPGA, DSP, SRAM, FLASH and two video inputs, one video output, and various interfaces to upper-computer. FPGA collects images and detects SIFT keypoints, while DSP extracts descriptors of
the keypoints and calculates the 3D coordinates of space points. The results can be sent to the control system of the robot via HPI bus or Ethernet. The size of the board is 158x123 mm².

III. SYSTEM CALIBRATION

Before using, the system needs to be calibrated. The calibration process contains two parts: the first is the calibration of cameras, and the second is to calibrate a quadratic polynomial to compensate the measurement results.

Fig. 2 illustrates the basic measuring model of the stereovision. According to the pin-hole imaging model of a camera, a point in the 3D space can be transformed to a 2D point in the image through the projection transformation. So we have:

$$\begin{bmatrix}
    u_l \\
    v_l \\
    1
\end{bmatrix} = \frac{1}{Z_l} \begin{bmatrix}
    f_{sl} & 0 & u_{0l} \\
    0 & f_{sl} & v_{0l} \\
    0 & 0 & 1
\end{bmatrix} \begin{bmatrix}
    X_l \\
    Y_l \\
    Z_l
\end{bmatrix} \quad (1)
$$

$$\begin{bmatrix}
    u_r \\
    v_r \\
    1
\end{bmatrix} = \frac{1}{Z_r} \begin{bmatrix}
    f_{sr} & 0 & u_{0r} \\
    0 & f_{sr} & v_{0r} \\
    0 & 0 & 1
\end{bmatrix} \begin{bmatrix}
    X_r \\
    Y_r \\
    Z_r
\end{bmatrix} \quad (2)
$$

Where $f_{sl}, f_{sr}, u_{0l}, v_{0l}$ and $f_{sr}, f_{sr}, u_{0r}, v_{0r}$ are the intrinsic parameters of the left and right cameras respectively. $(u_l, v_l)$ and $(u_r, v_r)$ are the pixel image coordinates of left and right images, $(X_l, Y_l, Z_l)$ and $(X_r, Y_r, Z_r)$ are the Euclidean coordinates of the space point in the left and right camera coordinate system.

The distortion model of the camera is introduced to acquire more accurate results. Let $(u_l, v_l)$ and $(u_r, v_r)$ be the ideal (nonobservable distortion-free) pixel image coordinates of left and right images, and let $(\bar{u}_l, \bar{v}_l)$ and $(\bar{u}_r, \bar{v}_r)$ be the corresponding real observed image coordinates. Similarly, $(x_l, y_l)$ and $(\bar{x}_l, \bar{y}_l)$, $(x_r, y_r)$ and $(\bar{x}_r, \bar{y}_r)$ are the ideal (distortion-free) and real (distorted) normalized image coordinates of left and right images. Hence, the distortion models of cameras are:

$$\bar{x}_l = x_l + x_l \left[ k_{1l} \left( x_l^2 + y_l^2 \right) + k_{2l} \left( x_l^2 + y_l^2 \right)^2 \right] + 2 p_{1l} y_l + p_{2l} \left( x_l^2 + y_l^2 \right) + 2 x_l \left( x_l^2 + y_l^2 \right) \quad (3)
$$

$$\bar{y}_l = y_l + y_l \left[ k_{1l} \left( x_l^2 + y_l^2 \right) + k_{2l} \left( x_l^2 + y_l^2 \right)^2 \right] + p_{1l} x_l + p_{2l} \left( x_l^2 + y_l^2 \right) + 2 y_l \left( x_l^2 + y_l^2 \right) \quad (4)
$$

Before using, the system needs to be calibrated. The calibration process contains two parts: the first is the calibration of cameras, and the second is to calibrate a quadratic polynomial to compensate the measurement results.

Fig. 2 illustrates the basic measuring model of the stereovision. According to the pin-hole imaging model of a camera, a point in the 3D space can be transformed to a 2D point in the image through the projection transformation. So we have:

$$\bar{u}_l = u_l + (u_l - u_{0l}) \left[ k_{1l} \left( x_l^2 + y_l^2 \right) + k_{2l} \left( x_l^2 + y_l^2 \right)^2 \right] + 2 p_{1l} y_l + p_{2l} \left( x_l^2 + y_l^2 \right) + 2 x_l \left( x_l^2 + y_l^2 \right) \quad (5)
$$

$$\bar{v}_l = v_l + (v_l - v_{0l}) \left[ k_{1l} \left( x_l^2 + y_l^2 \right) + k_{2l} \left( x_l^2 + y_l^2 \right)^2 \right] + p_{1l} x_l + p_{2l} \left( x_l^2 + y_l^2 \right) + 2 y_l \left( x_l^2 + y_l^2 \right) \quad (6)
$$

Besides, the spatial relationship between the two cameras is:

$$\begin{bmatrix}
    X_r \\
    Y_r \\
    Z_r
\end{bmatrix} = \begin{bmatrix}
    R & T \\
    0 & 1
\end{bmatrix} \begin{bmatrix}
    X_l \\
    Y_l \\
    Z_l \\
    1
\end{bmatrix} \quad (7)
$$

Where $R$ is the rotation matrix between the two cameras, $T$ is the translation vector.

Refer to the equation above, in the case of that the intrinsic parameters, distortion coefficients and the structure parameters of cameras are known, the 3D coordinates can be
obtained. We use the Bouguet method implemented in the Caltech Camera Calibration Toolbox [11] to calibrate all the parameters above.

3D coordinate of a space point can be acquired by using stereovision measurement method, but we are interested in the $Z$ coordinate in the robot applications. The $Z$ coordinate calculated by the system is the distance that relative to the optical center of the left camera. In practice, it is difficult to determine the location of the optical center precisely. Accordingly, a reference plane paralleling to the optical plane is introduced. The real measurement results are the distances between the reference plane and the object to be measured. Since the stereovision measurement is based on triangulation, the measurement error is increasing as the distance increases. Therefore, a quadratic polynomial is introduced to compensate the results in this paper. Let

$$e = p_1Z_C^2 + p_2Z_C + p_3$$  (8)

Where $Z_C$ is the distance between the plane to be measured and the reference plane, $e$ is the absolute error between the measurement result and the actual distance, $p_1, p_2, p_3$ are the quadratic polynomial coefficients.

Therefore, the final measurement results after compensation can be derived from:

$$Z_S - Z = e = p_1Z^2 + p_2Z + p_3$$  (9)

$$Z_S = -(p_2 - 1) - \sqrt{(p_2 - 1)^2 - 4p_1(p_3 + Z)}$$  (10)

Where $Z$ is the depth calculated by the stereovision, and $Z_S$ is the final result after compensation.

IV. ALGORITHM REALISATION

The embedded image processing board is mainly to complete three tasks: keypoint detecting, generating the descriptors and calculating the 3D coordinates in space. The FPGA extracts keypoints, while the DSP is responsible for generating descriptors and calculating the 3D coordinates. The flowchart of the proposed algorithm is shown in Fig. 3.

A. Detecting Keypoints

Reference [6] improved the SIFT algorithm to make it suitable for parallel implementation on hardware. This paper employs the SIFT implemented on FPGA to detect keypoints from left and right images respectively.

B. Generating Descriptors

The computation of SIFT algorithm is mainly concentrated in the calculation process of high dimension descriptors. Since the rotation, viewpoint, scale and illumination between the two images obtained by the stereo system are not changed a lot, there is no need for the complex descriptor.

The proposed algorithm simplifies the calculation of descriptors by removing the orientation information and reducing the dimension of the descriptor. First, choose a 9x9 neighborhood with the keypoint as the center, in Fig. 4, the yellow pentagram is the location of the present keypoint, every grid denotes a pixel within the neighborhood of the keypoint. Then use a 3x3 median filter (red window in Fig. 4) to filter the neighborhood. Finally, use the filtered pixel values within the neighborhood to constitute an 81 dimensions descriptor, and then normalize the descriptor to remove the effect of light.

C. Keypoints Matching

The matching between keypoints from two images decides the accuracy of 3D coordinates. As the descriptor is simplified, the information included is much less than the original algorithm. In order to keep the accuracy of matching, the proposed algorithm performs a two-pass matching to search the matching points between the left and right images.

Firstly, choose a keypoint in the left image, and find a keypoint in the right image which makes the distance between them minimum to constitute a pair of candidate matching points. Then search for a point in the left image closest to the candidate point found in the right image. If the searched point in the left image is the one originally chose, a pair of matching points comes out, otherwise the matching fails. Fig. 5 illustrates the matching process with a simple example. And Fig. 6 shows a pair of matching results of actual scene.
V. EXPERIMENTAL RESULTS

A. Comparison of Different Algorithm

Three algorithms are compared in aspects of processing speed and matching accuracy. The first method used in [6] is to implement the SIFT algorithm on the hardware, recorded as SIFT. Second one is to employ the way proposed in this paper to generate the keypoint descriptors, and search matching points from left image to right image only, recorded as SIFT-Block. The third one is the very method proposed in this paper with bidirectional searching for matching points, recorded as Bi-SIFT-Block. The comparative results are shown in Table I. According to the table, the method proposed in this paper can effectively improve the processing speed without any reduction of accuracy.

<table>
<thead>
<tr>
<th>Method</th>
<th>Matching numbers</th>
<th>Matching accuracy</th>
<th>Max process time(s)</th>
<th>Average process time(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIFT</td>
<td>25</td>
<td>92%</td>
<td>0.04652</td>
<td>0.04365</td>
</tr>
<tr>
<td>SIFT-Block</td>
<td>54</td>
<td>83.33%</td>
<td>0.02055</td>
<td>0.01795</td>
</tr>
<tr>
<td>Bi-SIFT-Block</td>
<td>44</td>
<td>95.46%</td>
<td>0.02887</td>
<td>0.02691</td>
</tr>
</tbody>
</table>

B. Parameters Calibration

A 2D chessboard is used as the calibration object. Capture 20 images of the calibration object from different viewpoints by left and right cameras respectively. One pair of the calibration images is shown in Fig. 7. The intrinsic parameters, distortion coefficients of the left and right cameras and the structure parameters between them which are calibrated by using MATLAB toolbox are concluded as follows:

\[
f_{xl} = 885.53389, f_{xr} = 967.03482
\]

\[
u_{xl} = 176.57115, v_{xl} = 162.90934
\]

\[
k_{ll} = -0.06010, k_{2l} = -2.53966
\]

\[
p_{ll} = 0.00366, p_{2l} = 0.00186
\]

\[
u_{lr} = 884.27382, v_{lr} = 966.59921
\]

\[
k_{lr} = -1.4079, k_{2r} = 1.06427
\]

\[
p_{lr} = 0.00516, p_{2r} = 0.00605
\]

\[
R = \begin{bmatrix} 0.99967 & 0.02267 & -0.01210 \\ -0.02284 & 0.99964 & -0.01422 \\ 0.01177 & 0.01449 & 0.99983 \end{bmatrix}
\]

\[
T = \begin{bmatrix} -111.03039 \\ 1.92323 \\ 3.31505 \end{bmatrix}^T
\]

The quadratic polynomial parameters are calibrated after the calibration of cameras. The distances between the plane to be measured and the reference plane are measured every other 200 mm in the range of 600 mm to 3000 mm, and then the absolute errors of each measurement are taken as the calibration data. The parameters can be obtained by fitting the quadratic polynomial curve. The calibration data are shown in Table II.

<table>
<thead>
<tr>
<th>Actual Distance(mm)</th>
<th>Absolute Error(mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>600</td>
<td>16.38354</td>
</tr>
<tr>
<td>800</td>
<td>72.13081</td>
</tr>
<tr>
<td>1000</td>
<td>138.8259</td>
</tr>
<tr>
<td>1200</td>
<td>212.4229</td>
</tr>
<tr>
<td>1400</td>
<td>294.5479</td>
</tr>
<tr>
<td>1600</td>
<td>373.7654</td>
</tr>
<tr>
<td>1800</td>
<td>459.3270</td>
</tr>
<tr>
<td>2000</td>
<td>570.5966</td>
</tr>
<tr>
<td>2200</td>
<td>685.6576</td>
</tr>
<tr>
<td>2400</td>
<td>769.3647</td>
</tr>
<tr>
<td>2600</td>
<td>909.1868</td>
</tr>
<tr>
<td>2800</td>
<td>1034.5262</td>
</tr>
<tr>
<td>3000</td>
<td>1170.6520</td>
</tr>
</tbody>
</table>
The data in TABLE II are the results calculated by the triangulation without compensation. We cannot ensure the axes of the camera being perpendicular to the plane to be measured, therefore the measurement errors become larger as the distance increase. Using these data to fit a polynomial curve, the parameters of the polynomial curve are:

\[ p_1 = 8.698 \times 10^{-5} \]
\[ p_2 = 0.1668 \]
\[ p_3 = -114.8 \]

and the fitting result is shown in Fig. 8.

From the table, the maximum measurement error after compensation is 54.7581 mm, and the average error is 11.6207 mm.

C. Measurement Experiment

The real measurement experiment is conducted with the calibrated parameters. In the range of 600 mm to 3000 mm, the distances from the plane to be measured to the reference plane are measured every other 100 mm. Fig. 9 shows a pair of experiment images. The average value of Z coordinates of all the matching points in the plane to be measured is taken as the measurement result. Some of the results are shown in Table III.

<table>
<thead>
<tr>
<th>Real distance(mm)</th>
<th>Measurement result(mm)</th>
<th>Absolute error(mm)</th>
<th>Relative error</th>
</tr>
</thead>
<tbody>
<tr>
<td>600</td>
<td>600.8134</td>
<td>0.8134</td>
<td>0.136%</td>
</tr>
<tr>
<td>700</td>
<td>706.4431</td>
<td>6.4431</td>
<td>0.920%</td>
</tr>
<tr>
<td>800</td>
<td>803.3257</td>
<td>3.3257</td>
<td>0.416%</td>
</tr>
<tr>
<td>900</td>
<td>902.6224</td>
<td>2.6224</td>
<td>0.291%</td>
</tr>
<tr>
<td>1000</td>
<td>999.9697</td>
<td>0.0303</td>
<td>0.003%</td>
</tr>
<tr>
<td>2600</td>
<td>2591.784</td>
<td>8.2165</td>
<td>0.316%</td>
</tr>
<tr>
<td>2700</td>
<td>2754.758</td>
<td>54.7581</td>
<td>2.028%</td>
</tr>
<tr>
<td>2800</td>
<td>2812.065</td>
<td>12.0651</td>
<td>0.431%</td>
</tr>
<tr>
<td>2900</td>
<td>2901.284</td>
<td>1.2836</td>
<td>0.044%</td>
</tr>
<tr>
<td>3000</td>
<td>2994.932</td>
<td>5.0680</td>
<td>0.169%</td>
</tr>
</tbody>
</table>

From the table, the maximum measurement error after compensation is 54.7581 mm, and the average error is 11.6207 mm.

D. Obstacle Avoidance of Mobile Robot

The measurement system is installed in a mobile robot platform. We use it to complete obstacle avoidance under an indoor environment. Fig. 10 illustrates the whole process with a simple example.

E. Results Analysis

Since the system employs two cameras, the target must appear in two camera image planes at the same time. If cameras are too close to the target, it is impossible to get the depth information of the target. Therefore, the proposed method has a range limitation.

The proposed algorithm is compared with other two methods and real measurement experiment is conducted. We also apply the stereovision system to a mobile robot to
complete obstacle avoidance task. Experimental results show that the proposed method can meet the need of localization and obstacle avoidance tasks of a robot in aspects of speed and precision.

VI. CONCLUSION

Stereovision measurement method with the advantages of low cost and real-time performance is widely applied in the field of robot. Searching for the matching points of two images is one of its key issues. This paper proposes a fast SIFT feature based stereovision measurement algorithm and realizes it on the embedded image processing board. The system has the advantages of small size, light weight and low power consumption. This method simplifies the descriptor without loss of matching precision, so the computational speed is improved. Furthermore, we also use a quadratic polynomial to compensate the measurement results in order to improve the accuracy.

For mobile robots, experimental results and comparison of different methods validate that the algorithm is effective.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (Grant No.61273360), National Science Fund for Distinguished Young Scholars (Grant No.61203328) and Key Project of S&T Plan of Beijing Municipal Commission of Education (No.KZZ201210005001).

REFERENCES