Recover 3D Information of the Moving Object from Video Streams

Yu-tong Zheng[1] and Ming Li and Fang Liao

Abstract Perception of the moving object in 3D from video streams has been one hot topic in computer vision. We present a fast method to reconstruct 3D information of the moving object from binocular video streams. System is assembled as two pipelines, technica are used to excavate the potential parallelism. With the corresponding points searching confined to very limited and credible region, the mismatching errors and time-consumed computation are reduced considerably. At the last, sparse depth map is calculated and then 3D contours and location of the object are estimated. The system is implemented and tested with outdoor and indoor moving object perception on 640×480 frame. Results show that the proposed method is improved in speed and stability. It can be used as a reference for autonomous navigation of mobile robot and object tracking.

Keywords Binocular, corresponding points, epipolar constraints, sparse depth map, autonomous navigation

1 Introduction

Making manmade machine capable to perceive the moving object in depth and co ntour from videos has received considerable attention in computer vision. Since th e projective transformation from 3D to image (2D) is intrinsic ambiguity, it is hard to recover 3D information from a single image [1]. We need more than one image to reconstruct 3D, so correspondence points matching inevitably play an importan t role in all the performance metrics. Along with achievement in intelligent compu ting, implementation technique and neurobiology relevant to vision [2, 3], many n ovel algorithms appear [4, 5, 6, 7], but further improvement is still wanted in univ ersal, stability, timeliness and accuracy. In many areas, such as human motion anal ysis, traffic monitoring, tracking, autonomous navigation, we want the system to work better like humans.

In this paper binocular are used to capture video streams synchronously, their two optical axis are parallel to each other and baseline between the two optical

[1] Yu-tong Zheng (✉)
Information Engineering College, Minzu University of China, Beijing, 100081
e-mail: zhengyutong68@aliyun.com
Ming Li
Information Engineering College, Minzu University of China, Beijing, 100081
Fang Liao
Information Engineering College, Minzu University of China, Beijing, 100081

centre hold invariant. When the moving object appears in the field, its 3D contour and location are reconstructed and updated in time. Section 2 describe the architecture of the whole system, section 3 describe the detail of the implementation, section 4 shows the development environment and tested results, and finally section 5 gives the conclusion.

2 Design Scheme

Simulating the human vision, the hardware of the system consists of two CMOS c ameras and one PC. Two cameras with identical parameters, fixed baseline and pa ralleled optical axis grab the same scene from two different positions synchronous ly; the left camera coordinate is regard as world coordinate. The intrinsic paramete rs and the poses of cameras are calibrated and rectified in advance precisely, so dis tortion are eliminated and calculation of subsequent process will reduce. The desig n scheme is shown in Fig.1. The whole system is assembled as two pipelines, whil e early in the program the information coming from the two cameras are processed in a parallel way.

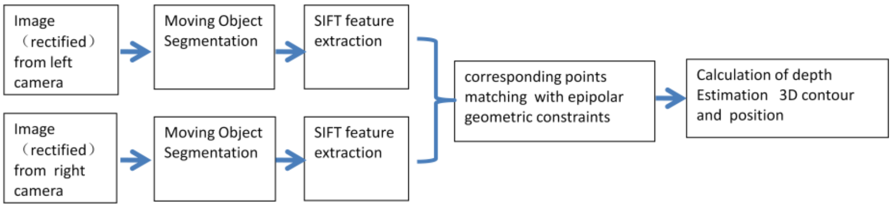
Image （rectified） from left camera
Moving Object Segmentation
SIFT feature extraction
Image （rectified） from right camera
Moving Object Segmentation
SIFT feature extraction
corresponding points matching with epipolar geometric constraints
Calculation of depth Estimation 3D contour and position

Fig.1 design scheme

Segmentation and the subsequent SIFT feature [8] extraction are processed syn chronously in two pipelines. The next step merged together is searching for corres ponding points under epipolar geometry constraints [1]. This step has key effect o n the ultimate real-time, accuracy and stability, we adopt an improved algorithm w hich will be described in detail .Triangulation is used for the calculation of depth f rom disparity. Along with the objects moving initiatively, the sparse depth maps ar e updated and show on the screen. Finally, the 3D outline of the moving object an d its location are estimated.

The test results show that the proposed method is reasonable improved in speed. The why behind are some tricks as follows: the first is two pipeline used in the early stage; the second is the limited and valid matching area; the third reason is the improved stereo correspondence algorithm. The test results also show that the improvement in robustness, the reason in nature comes from the advantages of SIFT itself, SIFT feature is invariant to image scaling, rotation, illumination, viewpoint and well localized in spatial, so it will reduce the probability of disruption by occlusion, clutter, or noise [8].If the relative position of binocular camera remains unchanged, the system can be used for perception of the moving object in 3D with no need for calibration and rectification in operative mode and can be expanded for multi-target tracking and 3D modeling in real-time.

3 The details of the implementation

In this section, we describe the details step by step and the improved corresponding point match algorithm is proposed in the latter part.

3.1 Calibration and Stereo-rectification

Usually the lens has certain distortion. In many scenarios, eliminating the image distortion is the chief problem of image pre-processing. By calibration[9,11], we can get the camera intrinsic parameters, including focal length, the optical centre, distortion coefficient and relative orientation of the cameras .Images can be undistorted in a pre-processing step using distortion coefficients obtained during calibration.

Image stereo-rectification [10, 11] is the process by which two images of the same solid scene undergo homographic transforms, so that their corresponding epipolar lines coincide and become parallel to the x-axis of image. So, given x in the left image from the pair of stereo-rectified images, the search domain for x' in the right image corresponding to x is restricted on the same horizontal scan line.

Bouquet's Algorithm [12] is exploited for rectifying calibrated cameras. In OPENCV [13, 14], function void stereorectify ( ) computes rectification transforms for each head of calibrated stereo camera.

3.2 Moving Object Segmentation

Experiment shows that when tracking objects moving very fast, the target missing will happen occasionally if algorithm such as Mean-shift and SAD template applied. Segmentation [15] based on differential of sequential frame might have weakness such as slur and cavity, but missing target is not likely to happen. We exploit an improved method based on differential image of three consecutive frames. Details are as follows: first, two differential image according to the frame K-1 and the followed frame K, frame K and the followed K+1 are acquired; second, spot with pixel value greater than the threshold is regarded as the target area, opposite as the background; finally intersection of the target areas is calculated, the result is accepted as rough estimate of moving objects.

3.3 Feature Extraction

David Lowe put forward the SIFT features to describe image in 1989 [8]. It is a local feature, invariant to image scaling, rotation and robust to illumination, viewpoint, occlusion, clutter, or noise. It has been validated in image retrieval, tracking, image fusion and many other applications. The reason why we select it is stability and straightforward use for disparity calculation. SIFT feature point $s \in \Re^{133}$ is a high dimension data including scale, location, orientation, key point descriptor (128 dimension).By feature extraction two images are represented by two SIFT feature vector respectively.

3.4 Corresponding Point Match with Epipolar Geometric Constraints

Corresponding point match is an important and difficult task, its aim is to find the corresponding point if the source image point has been given.

Many matching algorithms [16] have been explored these years which can be divided into two kinds: sparse matching and dense matching. Dense matching matches all the pixels in the image. Region around the pixel is selected, according to the gray scale distribution or some other feature of the region, the comparability or relativity of the region in another image be obtained. This method has the problem of huge computation and mismatching errors .Usually texture images show a better performance, but if the region is too large then blocking effect appears. The advantage is that dense depth map is straightforward. Sparse matching selects the strong feature points for image matching, so feature extraction need be done first and then decides matching point in a candidates queue. Sparse depth map is obtained directly; the interpolation or affine transformation is required for dense depth map. But owing the blindness of the interpolation, it shows shortage in describe 3D structure in details. Due to no need to search in the whole image, the computation is reduced considerably. But the feature points are required to have definite characteristics in order to obtain high quality signals and stability.

In consideration of the needs of the autonomous navigation of moving robot, we exploit a sparse matching algorithm using epipolar geometric constraints which is the most fundamental and reliable geometric constraints in stereo vision. The projection X in the image plane of the point in the scene is on the corresponding epipolar line of the X' in another image plane [1]. If the parameters and the poses of cameras are calibrated beforehand precisely and rectified to make the two image planes parallel, then baseline intersects the image plane at infinity, and Epipolar lines are parallel to X axis, so searching will be on the same horizontal level. The pseudo code of the algorithm is as follows:

Assume:

Input:

$$L = \{V_k | V_k \in \text{FeaturePoint from } Left\,\mathrm{Im}age\}$$

$$R = \{V_k' | V_k' \in \text{FeaturePoint from } Right\,\mathrm{Im}age\}$$

Output:

$$Match_{l,r} = \{\langle P_l , P_r \rangle | P_l \in V_k , P_r \in V_k' \text{ and } \in \text{corresponding points of P}\}$$

Step1: let

$$L = \bigcup_{m=1}^{M} \{P_m | P_m \in \mathrm{V_k} \text{ have the identical Y coordinate}\}$$

$$R = \bigcup_{m=1}^{M} \{P_m' | P_m' \in \mathrm{V_k}' \text{ have the identical Y coordinate}\}$$

$$M \in 1,2,3......Y_{\max}$$

And denote subsets as:

$$P_m = \bigcup_{j=1}^{w} \left\{ V_j \middle| \mathrm{V_j} \in \Re^{133}, \quad \mathrm{j} \in 1,2,3......,\mathrm{w} \right\}$$

$$P_m' = \bigcup_{n=1}^{q} \left\{ W_n \middle| \mathrm{W_n} \in \Re^{133}, \quad \mathrm{n} \in 1,2,3......,\mathrm{q} \right\}$$

```
Step2: initialize
$\mathrm{MIN} = 1000, \mathrm{Match}_{\mathrm{l,r}} = \phi$
$\mathrm{k} = 0, \mathrm{k} \in 0,1,2,......,\mathrm{Y}_{\max}; \mathrm{j} = 1, \mathrm{j} \in 0,1,2,......,\mathrm{w}; \mathrm{n} = 1, \mathrm{n} \in 0,1,2,......,\mathrm{q};$
    step21: $\mathrm{k} = \mathrm{k} + 1;$
        step22: $\mathrm{j} = \mathrm{j} + 1$
            step23: $\mathrm{n} = \mathrm{n} + 1$
                Calculate similarity $\mathrm{S}_{\mathrm{j,n}}$ by formula:
$\mathrm{S}_{\mathrm{j,n}} = \lambda_1 \times \sqrt{\sum_{\alpha=1}^{2} [P_j^{\alpha} - P_n^{\alpha}]^2} + \lambda_2 \times \sqrt{\sum_{\beta=1}^{128} [P_j^{\beta} - P_n^{\beta}]^2}$     (3.1)
$P_j \in L, P_n \in \mathrm{R};$
$P_j^{\alpha}$ , $P_n^{\alpha}$ : $\alpha$ denotes the subdomain orientation ;
$P_j^{\beta}$ , $P_n^{\beta}$ : $\beta$ denotes the subdomain describtor;
$\lambda_1$ denotes weight coefficient of subdomain orientation;
$\lambda_2$ denotes weight coefficient of subdomain describtor.
                If $\mathrm{S}_{\mathrm{j,n}} < \mathrm{MIN}$ then $\mathrm{MIN} = \mathrm{S}_{\mathrm{j,n}}$;
                GOTO step23 until n=q;
            Put MIN into $\mathrm{Match}_{\mathrm{l,r}}$;
        GOTO step22 until j=w;
    GOTO step21 until $\mathrm{k} = \mathrm{Y}_{\max}$.
Step3: swap L and R, prepare $\mathrm{Match}_{\mathrm{r,l}}$ to store the output then GOTO Step2
Step4: $output = Match_{l,r} = Match_{r,l} \cap Match_{l,r}$
```

Finally, mismatch point need to be removed. Study of Visual science show that human have depth concept by integrating images from two eyes limited to a certain range of disparity gradient, this reflect the continuity constraint on the surface of the object to some extent. It is reasonable that we presume change of disparity from the adjacent points on the surface of the object is in a certain range. The corresponding points with difference of X values greater than threshold are removed.

3.5 Calculation of depth and estimation 3D contours and position

Feature point P （X, Y, Z）in the scene is projected on the left image plane as $\mathrm{P_l}$ $(\mathrm{x_l}, \mathrm{y_l})$ and right as Pr（$\mathrm{x_r}, \mathrm{y_r}$）, 3 D information in the world coordinate can be calculated based on triangulation. Formulas are:

$$Z = \frac{fB}{| x_l - x_r |} \tag{3.2}$$

$$X = \frac{Bx_l}{| x_l - x_r |} \tag{3.3}$$

$$Y = \frac{By_l}{| x_l - x_r |} \tag{3.4}$$

B denotes the distance between the optic centers; it is 150mm in this system, f denote camera focal length and it is 820 pixels. Sparse depth map can be obtained directly from sparse matching. Fig.2 is sparse depth map of the object moving parallel to the camera plane, horizontal ordinate is frame, the unit is frame, vertical coordinate is the depth, unit is millimeters, for convenience of observation, and unit in the figure is set to the decimeter. Fig.3 is depth map of the object moving relative vertical to the camera plane.

Fig.4 shows the 3D of contour and location of the moving object, respectively 22, 36, 42 frames, the unit is millimeters. For facilitate observation, unit in the figure is set to the decimeter.

In some application such as 3D reconstruction where detail information about the structure of the object surface is needful, interpolation or affine transformation will be exploited to calculate dense depth map. In the application such as autonomous navigation of mobile robot, extreme are used to estimate the external contour. We use the arithmetic mean P_i (X, Y, Z) (subscript i denote all the feature points) estimate object's center of mass P_c. The estimate of the contourΛ is a cuboids, the formula is as follows:

$$\Lambda = \Omega_1 \cap \Omega_2 \cap \Omega_3 \cap \Omega_4 \cap \Omega_5 \cap \Omega_6 \tag{3.5}$$

Ω_1 and Ω_2 are two planes parallel to the camera plane XY, the distance from the XY are Z_{min} and Z_{max}; Ω_3 and Ω_4 are two planes parallel to the plane XZ, the distance from the optic center of the left the camera are Y_{min} and Y_{max}; Ω_5 and Ω_6 are two planes parallel to the plane YZ, the distance from the optic center of the left the camera are X_{min} and X_{max}.

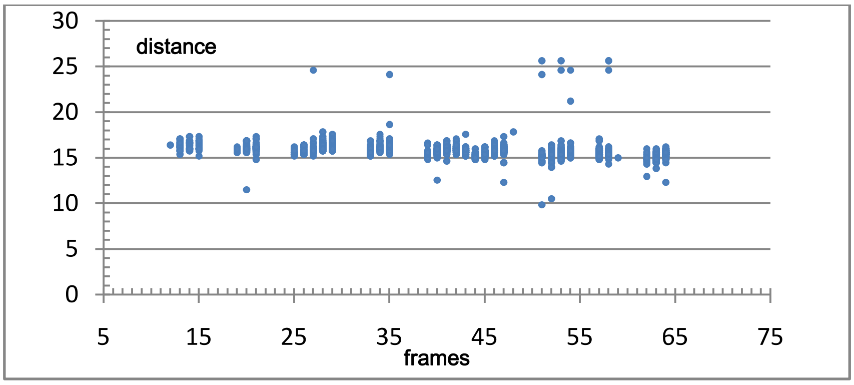

Fig. 2 sparse depth map of the object moving parallel to the camera plane

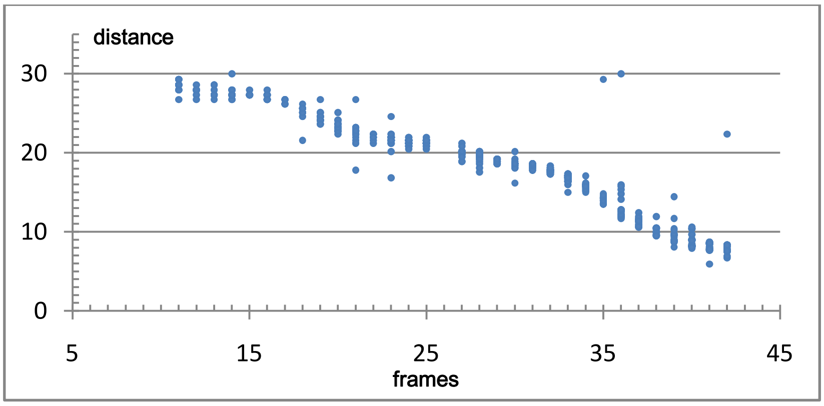

Fig. 3 sparse depth map of the object moving vertical to the camera plane

4 Results

We achieved a binocular system built on PC platform (Core i3-2120 3.30GHz, RA M 2G) + two CMOS cameras in the form of two pipeline, it can be used for recov er 3D information of the moving object from video stream. The programming envi ronment is OPENCV+VISUAL STUDIO 2008, programming language is C and C ++. The system has a satisfied speed and stability, where continuous image output without obvious dithering phenomenon. Fig. 4 shows that the shape composed by feature points in each frame remains generally stable, therefore the steady distribut ion of SIFT points is a valuable suggestion deserved further excavation for reconst ructing the object 3D outline.

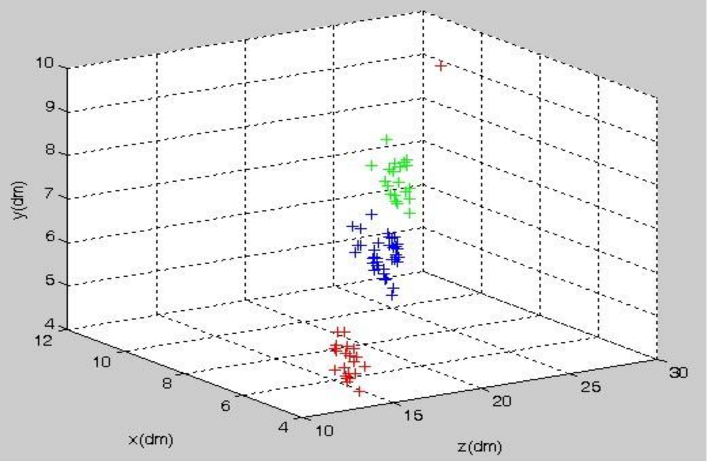

Fig. 4 3D display of contour and location of the moving object

5 Conclusions and Future Work

With the PC platform, the system receives data from outside two CMOS digital cameras, reconstructs sparse 3D image of the moving object with two peculiar pipelines, shows the result on VGA monitor in time. Owing to its good performance in real-time and robustness, it can be used as a reference for autonomous navigation of mobile robot and object tracking.

The main feature of the system is as follows:

Multi-pipeline and Multi-threading technology improve the system parallelism to the greatest extent; Frame differential method is used to estimate the moving object, it improves the speed of the subsequent stages by reducing calculation limited to a particular area; Improved stereo correspondence algorithm greatly reduces the search range. All these measures make the whole task showed a good performance in real-time.

SIFT features show robustness and powerful resistance to noise in many application, it is a guarantee for the correctness of the corresponding match.

In the case of multi-targets, the targets can still be estimated by segmentation and Multi-pipeline and Multi-threading technology still works, so the system can be adjusted to multi- target tracking conveniently.

There is no need for calibration and rectification in operative mode. If the relative position of binocular camera remains unchanged, the system still works even if the binocular cameras are in motion.

Time analysis shows that feature extraction is bottleneck in pipeline. If we do something such as substitute SURF for SIFT to optimize, we can further improve the real-time performance of the system.

6 References

1. R. Hartley and A. Zisserman. (2005). Multiple View Geometry in Computer Vision, 2nded.. Cambridge University Press, Cambridge
2. D. Leopold and N. K. Logothetis. (1996).Activity changes in early visual cortex reflect monkeys' precepts during binocular rivalry. Nature:379, 549-553
3. Takanori Uka and Gregory C. DeAngelis. (2006). Linking Neural Representation to Function in Stereoscopic Depth Perception: Roles of the Middle Temporal Area in coarse versus Fine Disparity Discrimination. The Journal of Neuroscience, June 21, 26(25): 6791– 6802
4. Guan, S., Klette, R. (2006).Belief Propagation on edge image for stereo analysis of image sequences. In Proceedings Robot Vision. LNCS 4931, p. 291 - 302.
5. Y Boykov, O Veksler, R Zabih, Fast approximate energy minimization via graph cuts, IEEE Trans PAMI, 2001,23(11): 1222-1239
6. D. Scharstein and R. Szeliski. (2002). Taxonomy and evaluation of dense two-frame stereo correspondence algorithms.International Journal of Computer Vision, April-June 47(1/2/3):7-42
7. Guo Longyuan,Xia Yongquan,Yang Jingyu.(2008).Adaptive Search Region Fast Area_Based Stereo Correspondence.IEEE Trans.Image and Signal Processing,2:27-30
8. Lowe D G. (2004).Distinctive image features from scale-invariant key points IJCV, 60(2): 91-110
9. J.-Y.Bouguet.(2001).Camera calibration toolbox for Matlab, http://www.vision.caltech.edu/ bouquet/calib_doc/
10. Luca Lucchese.(2005).Geometric calibration of digital cameras through multi-view rectification, Image and Vision Computing ,23 ,517–53
11. Shimizu, M., Okutomi, M. (2008).Calibration and rectification for reflection stereo. In IEEE Conference on Computer Vision and Pattern Recognition CVPR. Anchorage (USA), p. 1 - 8.
12. http://www.vision.caltech.edu/bouguetj/calib doc/index.html. 2004.
13. http://sourceforge.net/projects/opencvlibrary/
14. G. Bradsky and A. Kaehler. (2008).Learning OpenCV: Computer Vision with the OpenCV library. O'Reilly, Sebastopol, CA
15. Lukac, P., Hudec, R., Benco, M., Kamencay, P., Dubcova, Z., Zachariasova, M. (2011).Simple comparison of image segmentation algorithms based on evaluation criterion. In Proceedings of 21st International Conference Radio electronics. Brno (Czech Republic), p. 233 - 236.
16. Kuhl, A. (2005).Comparison of stereo matching algorithms for mobile robots. Centre for Intelligent Information Processing System. University of Western Australia, p. 4-24.