A Robust Two Feature Points Based Depth Estimation Method[1)]

ZHONG Zhi-Guang YI Jian-Qiang ZHAO Dong-Bin

(*Laboratory of Complex Systems and Intelligence Sciences, Institute of Automation, Chinese Academy of Sciences, Beijing* 100080)

(E-mail: {zhiguang.zhong, jianqiang.yi, dongbin.zhao}@mail.ia.ac.cn)

Abstract This paper presents a novel depth estimation method based on feature points. Two points are selected arbitrarily from an object and their distance in the space is assumed to be known. The proposed technique can estimate simultaneously their depths according to two images taken before and after a camera moves and the motion parameters of the camera may be unknown. In addition, this paper analyzes the ways to enhance the precision of the estimated depths and presents a feature point image coordinates search algorithm to increase the robustness of the proposed method. The search algorithm can find automatically more accurate image coordinates of the feature points based on their detected image coordinates. Experimental results demonstrate the efficiency of the presented method.

Key words Depth estimation, image sequence, motion vision, mobile robot navigation

1 Introduction

The focus of this paper is to discuss the depth from motion problem, which has many important applications such as mobile robot navigation. Reported methods for this problem mainly are optical-flow based, gradient-based and feature-based. Among them, the optical-flow based method[1,2] is very sensitive to errors in optical flow computation since the estimation of optical flow is an ill-posed problem solvable only by introduction of additional constraints (*e.g.*, smoothness). The gradient-based method[3,4], also called direct method, is difficult to recover the scene depth in the vicinity of focus-of-expansion[5]. Further, reported direct method generally is based on the brightness change constraint equation using least-squares method as an estimator, which is sensitive to the residual errors that are not Gaussian distributed[4]. The above two groups of methods generally require image motion be small, usually less than several pixels or even one pixel[6].

The feature-based method has been extensively studied because it requires fewer constraints on a motion model or scene depth. Selected features can be points, lines, curves, etc. Only feature points based method is considered in this paper. The number of selected feature points can be two or more. For example, Or[7] estimated depths according to a set of feature points whose three-dimensional coordinates in the space are known before the camera moves. Oberkampf[8] presented a method for depth estimation using four or more coplanar feature points. Fujii[9] assumed that a camera moves along its focal axis and there are two feature points on its object with the same depths in the camera coordinate system, then estimated their depths according to their distance difference in two consecutive images. Huber[10] introduced a similar technique. However, instead of assuming, it directly selects two feature points from an object with the same depth in the camera coordinate system.

This paper proposes a novel depth estimation method based on two feature points. Different from Fujii[9] and Huber[10], the proposed method selects feature points arbitrarily. It does not require a camera move along its focal axis. In addition, it only needs the distance in the space between the two feature points is known while the camera motion can be unknown. According to two consecutive images taken when a camera approaches its object in an arbitrary direction, the presented technique can estimate simultaneously the depths of the two feature points before and after the camera moves.

Further, since detected image coordinates of the feature points may not be very accurate due to many factors such as image noise, this paper also analyzes the ways to enhance the estimation precision and presents a feature point image coordinates search algorithm to increase the robustness of this

1) Supported by the National "973" Plan (2003CB517106) and National Natural Science Foundation of P. R. China (60475030)

Received January 6, 2004; in revised form July 6, 2005

Copyright © 2005 by Editorial Office of Acta Automatica Sinica. All rights reserved.

technique. The search algorithm can find automatically more accurate image coordinates of the feature points based on their detected image coordinates.

2 Depth estimation

As shown in Fig. 1, points A and B are the selected feature points. l is their distance in the space. Points C_1 and C_2 represent the projection centers of a camera before and after it moves. It is assumed that the camera can rotate before and after it moves while it only translates during the motion. l_{a1} *etc.* denote the length of line segment C_1A *etc.*

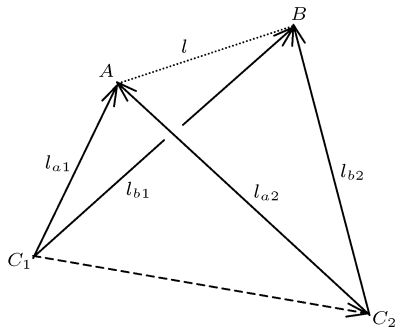

Fig. 1 Depth estimation principle

Assume that the intrinsic parameters of the camera and the image coordinates of the feature points are known. Then according to Fig. 1, we can get

$$l_{a1}\boldsymbol{m}_{a1} - l_{a2}\boldsymbol{m}_{a2} = l_{b1}\boldsymbol{m}_{b1} - l_{b2}\boldsymbol{m}_{b2} \tag{1}$$

where $\boldsymbol{m}_{ai}$ and $\boldsymbol{m}_{bi}(i = 1, 2)$ are the normalized vectors of A and B on the two consecutive image planes, respectively. As we have assumed above, they are all known.

According to (1), after some dot product and cross product operations, the following equation can be obtained:

$$l_{a2}(\boldsymbol{m}_{a2}, \boldsymbol{m}_{a1}, \boldsymbol{m}_{b1}) = l_{b2}(\boldsymbol{m}_{b2}, \boldsymbol{m}_{a1}, \boldsymbol{m}_{b1}) \tag{2}$$

Here, $(\bullet)$ denotes scalar triple product.

If $(\boldsymbol{m}_{a2}, \boldsymbol{m}_{a1}, \boldsymbol{m}_{b1})$ is not zero, that is, $\boldsymbol{m}_{a2}, \boldsymbol{m}_{a1}$ and $\boldsymbol{m}_{b1}$ are not coplanar, it is easy to get

$$l_{a2}/l_{b2} = (\boldsymbol{m}_{b2}, \boldsymbol{m}_{a1}, \boldsymbol{m}_{b1})/(\boldsymbol{m}_{a2}, \boldsymbol{m}_{a1}, \boldsymbol{m}_{b1}) = \mu_2 \tag{3}$$

Since

$$l^2 = l_{a2}^2 + b_{b2}^2 - 2l_{a2}l_{b2}(\boldsymbol{m}_{a2}, \boldsymbol{m}_{b2}) \tag{4}$$

where $(\boldsymbol{m}_{a2}, \boldsymbol{m}_{b2})$ represents dot product, the above two equations lead to

$$\begin{cases} l_{a2} = \mu_2 l_{b2} \\ l_{b2} = l/\sqrt{1 + \mu_2^2 - 2\mu_2(\boldsymbol{m}_{a2}, \boldsymbol{m}_{b2})} \end{cases} \tag{5}$$

Similarly, we can get

$$\begin{cases} l_{a1} = \mu_1 l_{b1} \\ l_{b1} = l/\sqrt{1 + \mu_1^2 - 2\mu_1(\boldsymbol{m}_{a1}, \boldsymbol{m}_{b1})} \end{cases} \tag{6}$$

where

$$(\boldsymbol{m}_{b1}, \boldsymbol{m}_{a2}, \boldsymbol{m}_{b2})/(\boldsymbol{m}_{a1}, \boldsymbol{m}_{a2}, \boldsymbol{m}_{b2}) = \mu_1 \tag{7}$$

Therefore, the depths of feature points A and B in the original and consecutive camera coordinate systems can be computed as

$$\begin{cases} d_{a1} = l_{ai}\boldsymbol{m}_{ai}(3) \\ d_{bi} = l_{bi}\boldsymbol{m}_{bi}(3) \end{cases}, \quad i = 1, 2 \tag{8}$$

where $\boldsymbol{m}_{ai}(3)$ and $\boldsymbol{m}_{bi}(3)$ represent the third element (*i.e.*, Z element) of vectors $\boldsymbol{m}_{ai}$ and $\boldsymbol{m}_{bi}$. If $(\boldsymbol{m}_{a2}, \boldsymbol{m}_{a1}, \boldsymbol{m}_{b1})$ is zero, then points C_1, C_2, A and B are coplanar. For this case, the depth estimation principle is similar and not discussed in this paper.

From the above depth estimation process, one can see that the proposed method is simple and visualized. It needs less computation time and is applicable to real-time application. However, the accuracy of the estimated depths depends heavily on that of the detected image coordinates. To enhance the precision of the estimated results and the robustness of the proposed technique, it is needed to optimize the computed depths according to different given conditions. For example, if n feature points can be selected arbitrarily from an object and their distances in the space are all known, then to enhance the depth estimation precision of a feature point P, we can compute its $n-1$ estimated depths according to (8) from the $n-1$ feature point pairs constituted by P and the other $n-1$ feature points, and then obtain the optimum estimation depth of P by various optimization algorithms such as calculating their mean value.

In the following section, a search algorithm is presented that requires the selected two feature points have the same height from the ground plane. It can enhance the accuracy of the detected image coordinates and then increase the depth estimation precision.

3 Search algorithm

For the convenience of discussion, we take mobile robot navigation for example to introduce this search algorithm. Assume a mobile robot is moving to an object, on which there are two feature points with the same height from the floor and their distance in the space is known.

As shown in Fig. 2, points A and B are the two feature points with the same height. Points O_{c1} and O_{c2} are the projection centers of the camera before and after the robot moves. $O_{r1}X_{r1}Y_{r1}Z_{r1}$ is the robot coordinate system and also represents the ground plane containing A and B. Its Z_{r1} axis is parallel to the translation direction of the robot. Points O_{r1} and O_{r2} are the orthogonal projections of points O_{c1} and O_{c2}, respectively. h is the altitude difference between the feature points and the projection center of the camera. It may be zero or not zero and only the latter case is discussed in this paper. l is the length of line segment AB. r is the displacement of the robot.

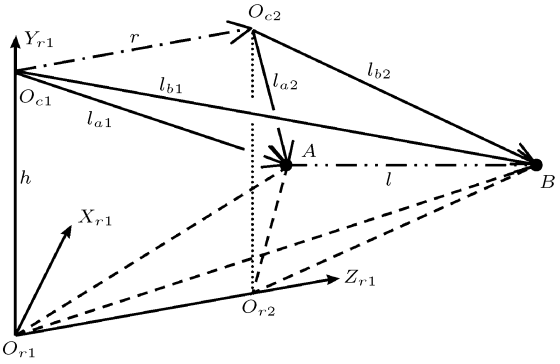

Fig. 2 Geometric relationship for the search algorithm

Obviously, the above depth estimation process is also applicable to this situation. In addition, since the two feature points A and B are at the same height from the floor, we can obtain the constraint equations, *i.e.*,

$$\begin{cases} \overrightarrow{O_{r1}A} = [0, h, 0] + l_{a1}(R \cdot \boldsymbol{m}_{a1}) \\ \overrightarrow{O_{r1}B} = [0, h, 0] + l_{b1}(R \cdot \boldsymbol{m}_{b1}) \\ \overrightarrow{O_{r2}A} = [0, h, 0] + l_{a2}(R \cdot \boldsymbol{m}_{a2}) \\ \overrightarrow{O_{r2}B} = [0, h, 0] + l_{b2}(R \cdot \boldsymbol{m}_{b2}) \end{cases} \tag{9}$$

Here, R is the rotational matrix of the robot coordinate system relative to the camera coordinate system and can be computed according to the pan and tilt angles of the camera.

These constraint equations determine the geometric relationships, which the image coordinates of the two feature points in the two consecutive images must satisfy. Therefore, if the detected image coordinates are not very accurate, it is possible to find their more accurate image coordinates based on their detected ones and then obtain more accurate depth estimation results.

Assume that the detected image coordinates in the two consecutive images are (u_{a1}, v_{a1}), (u_{b1}, v_{b1}), (u_{a2}, v_{a2}) and (u_{b2}, v_{b2}), respectively. We first keep u_{a1}, u_{b1}, u_{a2} and u_{b2} constant and search the N pixels neighborhoods of v_{a1}, v_{b1}, v_{a2} and v_{b2}. The specific process is explained as follows: 1) for each quadruple $[v_{a1}, v_{b1}, v_{a2}, v_{b2}]$ constituted by the coordinates within the neighborhoods of v_{a1}, v_{b1}, v_{a2} and v_{b2}, $\overrightarrow{O_{r1}A}$, $\overrightarrow{O_{r1}B}$, $\overrightarrow{O_{r2}A}$ and $\overrightarrow{O_{r2}B}$ can be computed from (9). If $|\overrightarrow{O_{r1}B}(2) - \overrightarrow{O_{r1}A}(2)|$ (the difference between the Y elements of vectors $\overrightarrow{O_{r1}B}$ and $\overrightarrow{O_{r1}A}$) and $|\overrightarrow{O_{r2}B}(2) - \overrightarrow{O_{r2}A}(2)|$ are both less than the threshold value v, then this quadruple is considered as relatively accurate image coordinates. The reason is if $[v_{a1}, v_{b1}, v_{a2}, v_{b2}]$ are correct, then computed $\overrightarrow{O_{r1}A}$, $\overrightarrow{O_{r1}B}$, $\overrightarrow{O_{r2}A}$ and $\overrightarrow{O_{r2}B}$ should lie in the ground plane containing points A and B, that is, their Y_{r1} elements should be zero. Such quadruple is recorded until the four neighborhoods are completely searched. 2) If the number of the above satisfying quadruples is more than 1, then we compute $V = |v_{b1} - v_{a1}| + |v_{b2} - v_{a2}|$ for each recorded quadruple. The one that makes V the least is considered as more accurate image coordinates. The reason is that generally h is greatly less than the distance between the feature points and the camera. 3) If the quadruples making V the least are more than one, then the one that makes $|\overrightarrow{O_{r1}A}(2)|$ and $\overrightarrow{O_{r2}A}(2)$ the least is selected as the optimum image coordinates. The reason is the same as that in (1).

It is noted that more accurate image coordinates may not be found when this search step is over. The main cause is the detected image coordinates deviate heavily from their actual values. Then, we can expand the search range N or increase the threshold value v. Obviously, the former leads to more computation time while the latter results in a lower estimation precision. Therefore, it is needed to trade off the two ways in applications.

After the above search step, the depths of the feature points can generally be estimated accurately. To obtain better estimation results, we keep the found optimum quadruple $[v_{a1}, v_{b1}, v_{a2}, v_{b2}]$ constant and search the M (M is the set value of the search range) pixels neighborhoods of u_{a1}, u_{b1}, u_{a2} and u_{b2}. For each quadruple $[u_{a1}, u_{b1}, u_{a2}, u_{b2}]$ constituted by the coordinates within the neighborhoods of u_{a1}, u_{b1}, u_{a2} and u_{b2}, $\overrightarrow{O_{r1}A}$, $\overrightarrow{O_{r1}B}$, $\overrightarrow{O_{r2}A}$ and $\overrightarrow{O_{r2}B}$ can be computed from (9). If $\overrightarrow{O_{r1}A}(2)$ and $\overrightarrow{O_{r2}A}(2)$ have opposite signs, then this quadruple is considered relatively inaccurate. Otherwise, we compute $U = |\overrightarrow{O_{r1}A}(2)| + |\overrightarrow{O_{r2}A}(2)| + |\overrightarrow{O_{r1}B}(2) - \overrightarrow{O_{r1}A}(2)| + |\overrightarrow{O_{r2}B}(2) - \overrightarrow{O_{r2}B}(2) - \overrightarrow{O_{r2}A}(2)|$. The quadruple $[u_{a1}, u_{b1}, u_{a2}, u_{b2}]$ that makes U the least is considered as the most accurate image coordinates. The two constraint conditions are to make the second elements (*i.e.*, Y elements) of $\overrightarrow{O_{r1}A}$, $\overrightarrow{O_{r1}B}$, $\overrightarrow{O_{r2}A}$ and $\overrightarrow{O_{r2}B}$ as the same as possible.

Since the presented depth estimation method needs only two feature points, the constraint conditions that can be used for the proposed search algorithm is very limited. Therefore, the search results are not always the best. However, they always get much better than before.

4 Experiments and analysis

Some experiments are performed on real images to test the efficiency of the proposed depth estimation method and the search algorithm. To measure conveniently and accurately the actual depths of feature points, a calibration reference is selected as an object and a camera is forced to translate on a desk with 4×4 grids. Fig. 3 is a typical image for these experiments. Points E and F, G and H, A and B are 3 pairs of feature points with the same height from the floor. Their distances l in space are all 16.0 cm while the height differences h are 13.1cm, 5.1 cm and 0.0 cm (this case is not discussed in this paper), respectively. A part of experimental results are listed in Table 1, Table 2 and Table 3. All these results are obtained from the above two steps search algorithm.

Since the focus of this paper is not image processing, the image coordinates of the feature points are detected by hand. The search algorithm is operated in the 1 pixel neighborhoods of the detected image coordinates. The threshold value v is set to 0.0005. In all these tables, (d_1, d_2) (unit: cm)

represent the depths of the first feature point before and after the camera moves. The estimation results of the second feature point are similar to those of the first feature point and they are not listed here.

The data in Table 1 and Table 2 are obtained when E and F are selected as feature points. The depth estimation results of the first feature point are shown in Table 1 when the tilt angle is not zero before the camera moves and the camera translates 4cm each time along the Z-axis direction of the object coordinate system (*i.e.*, the world coordinate system). The estimation results shown in the first two lines of Table 2 are obtained when the pan angle is 45 degrees while the tilt angle is 8 degrees or 12 degrees and the camera translates simultaneously each time 4cm along the X-axis and Z-axis. The estimated depths listed in the last two lines of Table 2 are obtained when the pan angle is 63.4 degrees while the tilt angles is 8 degrees or 12 degrees and the camera translates each time 4cm and 2cm along the X-axis and Z-axis, respectively. The estimated depths listed in Table 3 are obtained when both the pan angle and the tilt angle are zero while height difference h varies and the camera translates 4cm each time along the Z-axis. When h is equal to 5.1cm, points G and H are selected as feature points. If l varies, then the depth estimation results are similar to those obtained when h varies. Therefore, they are not listed here.

Fig. 3 A typical image for real experiment

Table 1 Estimated depths when the tilt angle varies

Tilt angle	(d_1, d_2)				
4.0	(89.0, 91.1)	(93.2, 95.5)	(97.9, 100.4)	(103.0, 104.4)	(107.2, 108.7)
8.0	(90.0, 91.6)	(93.8, 96.1)	(98.5, 100.4)	(103.0, 105.1)	(107.9, 109.4)
12.0	(90.0, 92.2)	(94.4, 96.1)	(98.5, 101.0)	(103.7, 105.1)	(107.9, 108.7)
16.0	(90.6, 92.2)	(94.4, 96.1)	(98.5, 100.4)	(103.0, 104.4)	(107.9, 108.7)
Actual values	(88.6, 92.6)	(92.6, 96.6)	(96.6, 100.6)	(100.6, 104.6)	(104.6, 108.6)

Table 2 Estimated depths when the pan angle and the tilt angle vary

Pan angle	Tilt angle	(d_1, d_2)			
45.0	8.0	(89.5, 92.7)	(93.2, 96.7)	(98.5, 102.3)	(103.0, 107.2)
45.0	12.0	(89.0, 92.7)	(93.8, 97.3)	(99.1, 102.3)	(102.3, 105.1)
Actual values		(88.6, 92.6)	(92.6, 96.6)	(96.6, 100.6)	(100.6, 104.6)
63.4	8.0	(89.5, 90.6)	(91.6, 93.2)	(93.8, 96.7)	(97.3, 99.1)
63.4	12.0	(89.5, 91.1)	(92.2, 93.8)	(94.9, 96.1)	(96.7, 97.9)
Actual values		(88.6, 90.6)	(90.6, 92.6)	(92.6, 94.6)	(94.6, 96.6)

Table 3 Estimated depths when varies

h	(d_1, d_2)				
13.1	(88.0, 92.2)	(92.2, 96.1)	(96.7, 100.4)	(101.0, 103.7)	(105.1, 107.9)
5.1	(88.5, 91.1)	(91.6, 96.7)	(97.3, 99.1)	(101.7, 103.7)	(105.7, 107.9)
Actual values	(88.6, 92.6)	(92.6, 96.6)	(96.6, 100.6)	(100.6, 104.6)	(104.6, 108.6)

According to Table 1, Table 2, and Table 3, it is clear that the estimated depths obtained from the proposed method have fairly high accuracy. Under different experimental conditions, the relative errors of most estimation results are less than 2%. The sources of the estimation errors mainly are measurement errors and image distortion, which is not rectified in these experiments. If the search ranges are expanded or the threshold value is increased, then the estimation precision can be further enhanced.

5 Conclusions

This paper proposes a novel depth estimation method based on two feature points and two images. It can estimate simultaneously the depths of two feature points before and after a camera moves and the camera motion can be unknown. In addition, the presented search algorithm can find automatically more accurate image coordinates of the feature points based on their detected ones. Experimental results indicate that it can enhance greatly the precision of the estimation results. The proposed approach requires less computation time and is very applicable to real-time applications such as mobile robot navigation.

References

1 Toshiharu M, Noboru O. Motion and structure from perspectively projected optical flow by solving linear simultaneous equations. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robot and Systems, Picataway, NJ, USA: IEEE, 1997. 740∼745

2 Tagawa N, Inagaki A, Minagawa A. Parametric estimation of optical flow from two perspective views. *IEICE Transanction on Information and Systems*, 2001, **E84-D**(4): 485∼494

3 Hung Y S, Ho H T. A Kalman filter approach to direct depth estimation incorporating surface structure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1999, **21**(6): 570∼575

4 Park S K, Kweon I S. Robust and direct estimation of 3-D motion and scene depth from stereo image sequences. *Pattern Recognition*, 2001, **34**(9): 1713∼1728

5 Hanna K J, Okamoto N E. Combining stereo and motion analysis for direct estimation of scene structure. In: Proceedings of IEEE 4th International Conference on Computer Vision, Los Alamitos, CA, USA: IEEE, 1993. 357∼365

6 Bergen J R, Anandan P, Hanna K J, Hingorani J. Hierarchical model-based motion estimation. In: Proceedings of European Conference on Computer Vision, London, UK: Spring-Verlag, 1992. 237∼252

7 Or S H, Luk W S, Wong K H, etc. An effective iterative pose estimation algorithm. *Image and Vision Computing*, 1998, **16**(5): 353∼362

8 Oberkampf D, Dementhon D F, Davis L S. Iterative pose estimation using coplanar feature points. *Computer Vision and Image Understanding*, 1998, **63**(3): 495∼511

9 Fujii Y, Wehe D K, Weymouth T E. Robust monocular depth perception using feature pairs and approximate motion. In: Proceedings of the IEEE International Conference on Robotics and Automation, Picataway, NJ, USA: IEEE, 1992. 33∼39

10 Huber J, Graefe V. Motion stereo for mobile robots. *IEEE Transactions on Industrial Electronics*, 1994, **41**(4): 378∼383

ZHONG Zhi-Guang Received his bachelor and master degrees from Hunan University in 1998 and Institute of Automation, Chinese Academy of Sciences in 2002, respectively. Now he is a Ph. D. candidate at Institute of Automation, Chinese Academy of Sciences. His research interests include intelligent control, mobile robot, and computer vision.

YI Jian-Qiang Received his Ph. D. degree in 1992 from Kyushu Institute of Technology, Japan, and currently he is a professor at Institute of Automation, Chinese Academy of Sciences. His research interests include intelligent control, robotics, and mechatronics.

ZHAO Dong-Bin Received his Ph. D. degree from Harbin University of Technology in 2000. He is now an associate professor at Institute of Automation, Chinese Academy of Sciences. His research interests include intelligent control, robotics, and mechatronics.