REDUCING ACCUMULATED ERRORS IN EGO-MOTION ESTIMATION USING LOCAL BUNDLE ADJUSTMENT

Akihiro Sugimoto
National Institute of Informatics
Chiyoda, Tokyo 101-8430, Japan
Email: sugimoto@nii.ac.jp

Tomohiko Ikeda
Chiba University
Inage, Chiba 263-8522, Japan

Keywords: Accumulated errors, ego-motion estimation, bundle adjustment, fixation control, computer vision.

Abstract: Incremental motion estimation methods involve a problem that estimation accuracy gradually becomes worse as the motion trajectory becomes longer and longer. This is due to accumulation of estimation errors incurred in each estimation step. To keep estimation accuracy stable even for a long trajectory, we propose to locally apply the bundle adjustment to each estimated motion so that the modified estimation becomes geometrically consistent with time-series frames acquired so far. To demonstrate the effectiveness of this approach, we employ an ego-motion estimation method using the binocular fixation control, and show that (i) our modification of estimation is statistically significant; (ii) in order to reduce estimation errors most effectively, three frames are optimal for applying the bundle adjustment; (iii) the proposed method is effective in the real situation, demonstrating drastic improvement of accuracy in estimation for a long motion trajectory.

1 INTRODUCTION

In the wearable computer environment (Clarkson et al., 2000) understanding where a person was and where the person is/was going is a key issue (Aoki et al., 2000; Davison et al., 2003) for just-in-time teaching, namely, for providing useful information at teachable moment. In the robot vision, on the other hand, the SLAM (Simultaneous Localization and Mapping) problem (Dissanayake et al., 2001; Guivant and Nebot, 2001; Thrun, 2002), in particular, mobile robot navigation and docking require the robot localization, the process of determining and tracking the position (location) of mobile robots relative to their environments (Davison and Murray, 1998; DeSouza and Kak, 2002; Nakagawa et al., 2004; Sumi et al., 2004; Werman et al., 1999). Computing three-dimensional camera motion from image measurements is, therefore, one of the fundamental problems in computer vision and robot vision.

Most successful vision-based approaches to estimating the position and motion of a moving robot usually employ the stereo vision framework. For example, Davison-Murray (Davison and Murray, 1998) assumed planar motions and proposed a method that reconstructs 3D points as landmarks and that imposes on the points geometrical constraints derived from planar motions. Gonçalves–Araújo (Gonçalves and Araújo, 2002) proposed a method for estimating motions that uses information obtained from optical flows and a 3D point reconstructed by stereo vision. Molton–Brady (Molton and Brady, 2003), on the other hand, proposed a method that reconstructs 3D points and uses their correspondences before and after a motion for the motion estimation.

When we employ the stereo vision framework, however, we have to make two cameras share the common field of view and, moreover, establish feature correspondences across the images captured by two cameras. This kind of processing has difficulty in its stability. In addition, keeping the baseline distance wide is hard when we mount cameras on a robot or wear cameras. Therefore, accuracy of motion estimation is limited if we employ the stereo vision framework.

To overcome such problems, Sugimoto *et al.* (Sugimoto et al., 2004; Sugimoto and Ikeda, 2004) introduced *the binocular independent fixation control* (Sugimoto et al., 2004) (Fig. 1) to two active cameras, and proposed a method for incrementally estimating camera motion that ensures estimation accuracy independent of the baseline distance of the two cameras. In the method, the correspondence of the fixation point over last two frames obtained through the camera control together with line correspondences (Sugimoto et al., 2004) or optical flows (Sugimoto and Ikeda, 2004) plays a key role in motion estimation. This method, however, has a problem that estimation accuracy gradually becomes worse as the motion trajectory becomes longer and longer. This is because estimation errors incurred in each estimation step are accumulated and accumulated errors cannot be ignored in the case of a long trajectory. This problem is commonly involved in incremental estimation of parameters.

In the area of camera calibration, in particular calibration of multiple cameras, on the other hand, parameter estimation is usually formulated as a nonlinear least squares problem whose cost function is defined in terms of reprojection errors. Two steps are then iterated until convergence is observed: searching for parameters minimizing the cost function and reconstructing 3D feature coordinates using estimated parameters to compute reprojection errors for a cost function. This approach appears long ago in the photogrammetry and geodesy literatures and is referred to the bundle adjustment (Hartley and Zisserman, 2000; Triggs et al., 2000).

In this paper, we propose a method that incorporates the bundle adjustment to ego-motion estimation method using the binocular independent fixation control in order to reduce estimation errors in each step and, as a result, to reduce accumulated errors in estimation. The bundle adjustment is originally applied to all estimated parameters after all estimation steps are finished to modify all the parameters simultaneously. In this sense, the bundle adjustment is a batch process and not suitable for an incremental process like ego-motion estimation. To solve this contradiction, we here employ an approach of the bundle adjustment application to parameters estimated within local computation (Zhang and Shan, 2001; Zhang and Shan, 2003). We locally apply the bundle adjustment to each estimated motion so that the modified estimation becomes geometrically consistent with time-series images captured so far. This modification keeps the estimation method incremental and, at the same time, reduces estimation errors incurred in each estimation step and drastically reduces accumulated errors. The contributions of this paper to ego-motion estimation are summarized in three important respects: (i) we show that our modification of estimation is statistically significant; (ii) we show that in order to reduce estimation errors most effectively, three frames are optimal for applying the bundle adjustment; (iii) we demonstrate the effectiveness of the proposed method in the real situation, showing drastic improvement of accuracy in estimation for a long motion trajectory.

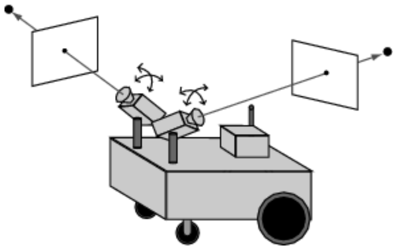


Figure 1: Binocular independent fixation control.

2 BINOCULAR INDEPENDENT FIXATION CONTROL

We here review the ego-motion estimation method that uses the binocular fixation control (Sugimoto and Ikeda, 2004). We note that in the binocular fixation control, each camera independently and automatically fixates its optical axis to its own fixation point in 3D and two fixation points are not necessarily the same.

Between a right camera and a left camera, we set the right camera is the base. Moreover, for simplicity, we assume that the orientation of the camera coordinate system does not change even though we change pan and tilt of the camera for the fixation control. This means that only the ego motion causes changes in orientation and translation of the camera coordinate systems. We also assume that the ego motion is identical with the motion of the right-camera coordinate system. We let the translation vector and the rotation matrix to make the left-camera coordinate system identical with the right-camera coordinate system be $\boldsymbol{T}_{\rm in}$ and $R_{\rm in}$ respectively. We also assume that the rotation and the translation of the right-camera coordinate system incurred by the ego motion are expressed as rotation matrix R in the right-camera coordinate system at time t and translation vector $\boldsymbol{T}$ in the world coordinate system. We remark that the extrinsic parameters between the two cameras as well as the intrinsic parameters of each camera are assumed to be calibrated in advance.

We denote by $\boldsymbol{v}_{\rm r}^t$ the unit vector from the projection center of the right camera toward the fixation point of the right camera in the right-camera coordinate system at time t. The constraint on ego-motion parameters follows from the fixation correspondence of the right camera:

$$\det\left[R_0\boldsymbol{v}_{\rm r}^t \mid R_0R\boldsymbol{v}_{\rm r}^{t+1} \mid \boldsymbol{T}\right] \quad = \quad 0,$$

where R_0 is the rotation matrix that makes the orientation of the world coordinate system identical with the orientation of the right-camera coordinate system at time t. Similarly, we obtain the following constraint on the ego-motion parameters from the fixation correspondence of the left camera.

$$\det\left[R_0 R_{\mathrm{in}}^\top \boldsymbol{v}_\ell^t \mid R_0 R R_{\mathrm{in}}^\top \boldsymbol{v}_\ell^{t+1} \mid \right.$$
$$\left. \boldsymbol{T} - R_0(R-I)R_{\mathrm{in}}^\top \boldsymbol{T}_{\mathrm{in}}\right] = 0.$$

Here $\boldsymbol{v}_\ell^t$ is defined in the same way as $\boldsymbol{v}_{\mathrm{r}}^t$.

Let $\boldsymbol{q}_{\mathrm{r}}^t$ be the coordinates of a point in the right-camera image at time t, and $\boldsymbol{u}_{\mathrm{r}}^t$ be the flow vector at the point. We then have the constraint on the ego-motion parameters:

$$\det\left[R_0 R(M\boldsymbol{u}_{\mathrm{r}}^t + \widetilde{\boldsymbol{q}}_{\mathrm{r}}^t) \mid R_0 \widetilde{\boldsymbol{q}}_{\mathrm{r}}^t \mid \boldsymbol{T}\right] = 0, \quad (1)$$

where

$$M = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}^\top, \qquad \widetilde{\boldsymbol{q}}_{\mathrm{r}}^t = ((\boldsymbol{q}_{\mathrm{r}}^t)^\top, f_{\mathrm{r}})^\top$$

and f_{r} is the focal length of the right camera. In the similar way, optical flow obtained from the left camera gives us the constraint on the ego motion.

$$\det[R_0 R R_{\mathrm{in}}^\top (M\boldsymbol{u}_\ell^t + \widetilde{\boldsymbol{q}}_\ell^t) \mid R_0 R_{\mathrm{in}}^\top \widetilde{\boldsymbol{q}}_\ell^t \mid$$
$$\boldsymbol{T} - R_0(R-I)R_{\mathrm{in}}^\top \boldsymbol{T}_{\mathrm{in}}] = 0,$$

where $\boldsymbol{q}_\ell^t$ is a point in the left-camera image at time t and $\boldsymbol{u}_\ell^t$ is the flow vector at the point. Note that $\widetilde{\boldsymbol{q}}_\ell^t$ is defined by $\widetilde{\boldsymbol{q}}_\ell^t := \left((\boldsymbol{q}_\ell^t)^\top \;\; f_\ell\right)^\top$ using the focal length f_ℓ of the left camera.

Ego motion has 6 degrees of freedom: 3 for a rotation and 3 for a translation. The number of algebraically independent constraints on the ego motion derived from two fixation correspondences, on the other hand, is two. One algebraically independent constraint is derived from each optical flow vector. We can therefore estimate ego motion if we have optical-flow vectors at more than four points. Namely, we form a simultaneous system of all the nonlinear constraints above and then apply a nonlinear optimization algorithm to solve the system. Parameters optimizing the system give the estimation of ego motion.

3 LOCAL BUNDLE ADJUSTMENT

The method reviewed in the previous section incrementally estimates ego motion using last two frames in each step. We cannot, however, guarantee that the method estimates correct motion parameters due to nonlinear optimization involved in estimation in each step: we may be trapped by a locally optimal solution. Even though the globally optimal solution is obtained, we still have errors incurred by numerical computation and/or observation of fixation point or optical flows. This means that we cannot ignore the accumulation of errors incurred in each estimation step and that estimation accuracy gradually becomes worse as the motion trajectory becomes longer and longer. As a result, geometrical inconsistency arises in the relationship between the fixation point in the current image and that in the other images captured so far. We here modify estimated parameters using the bundle adjustment so that such geometrical inconsistency does not occurs.

We assume that images are captured in discrete time-series and that from time s to time t ($t \geq s+2$), motion parameters between two frames are already estimated. We denote by $R_{i,i+1}, \boldsymbol{T}_{i,i+1}$ the motion parameters between time i and time $i+1$ ($i = s, \ldots, t-1$). Here, $R_{i,i+1}$ and $\boldsymbol{T}_{i,i+1}$ are the rotation matrix and the translation vector representing the ego-motion from time i to time $i+1$. We remark that the fixation control enables us to obtain the correspondences of fixation point images $\boldsymbol{p}_i$ ($i = s, \ldots, t$) from time s to time t.

For an image at time t, we focus on last n images including the image at time t ($3 \leq n \leq t-s+1$). We then apply the bundle adjustment to the n images. Since motion parameters relating two frames are already estimated at each time, we can combine them to obtain motion parameters $R_{j,t}$ and $\boldsymbol{t}_{j,t}$ that relate the images at time j and time t ($t-n+1 \leq j \leq t-1$) (see Fig. 2):

$$R_{j,t} = \prod_{i=j}^{t-1} R_{i,i+1}, \quad \boldsymbol{T}_{j,t} = \sum_{i=j}^{t-1} \boldsymbol{T}_{i,i+1}.$$

These parameters allow us to obtain in the image at time t, the epipolar line corresponding to $\boldsymbol{p}_j$. From the theoretical point of view, this epipolar line passes through $\boldsymbol{p}_t$, however, estimation errors and/or their accumulation cause the problem that the line does not pass through the point $\boldsymbol{p}_t$ as shown in Fig. 3. This indicates that epipolar constraints are broken. In other words, geometrical inconsistency arises in the relationship between the fixation point in the current image and that in the time-series images captured so far.

In the image at time t, the epipolar line determined by $\boldsymbol{p}_j$ and $R_{j,t}, \boldsymbol{T}_{j,t}$ is expressed by

$$\tilde{\boldsymbol{x}}^\top E_{j,t}\, \tilde{\boldsymbol{p}}_j = 0, \quad (2)$$

where $\tilde{\boldsymbol{x}}$ is the homogeneous coordinates of a point in the image at time t, $\tilde{\boldsymbol{p}}_j$ is the homogeneous coordinates of $\boldsymbol{p}_j$ and [1]

$$E_{j,t} := [\boldsymbol{T}_{j,t}]_\times R_{j,t}.$$

[1] $[\boldsymbol{T}_{j,t}]_\times$ is the 3×3 skew-symmetric matrix determined by $\boldsymbol{T}_{j,t}$. Namely, for any 3-dimensional vector $\boldsymbol{y}$, $[\boldsymbol{T}_{j,t}]_\times \boldsymbol{y} = \boldsymbol{T}_{j,t} \times \boldsymbol{y}$ is satisfied.

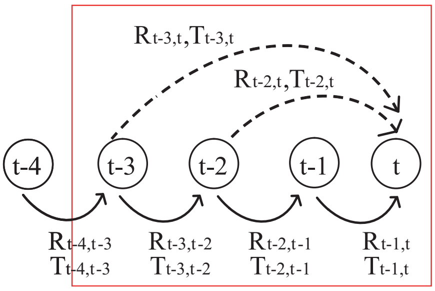


Figure 2: Relationship between motion parameters to which the local bundle adjustment is applied (the case of four frames).

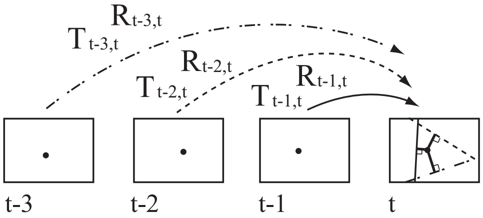


Figure 3: Epipolar lines obtained by the current fixation point and estimated parameters.

We can therefore evaluate errors of $R_{j,t}$ and $\boldsymbol{T}_{j,t}$ using the displacement of $\tilde{\boldsymbol{p}}_t$ from (2). We sum up displacements of $\tilde{\boldsymbol{p}}_t$ from each of the epipolar lines obtained from the last $(n-1)$ images and then modify $R_{t-1,t}$ and $\boldsymbol{T}_{t-1,t}$ so that modified parameters minimize the summation of the displacements over the concerned epipolar lines. Geometrical distance is used to evaluate the displacement of a point from a line (see Fig. 4). The cost function to be minimized is thus

$$\sum_{j=t-n+1}^{t-1} \left(\frac{\tilde{\boldsymbol{p}}_t^\top E_{j,t}\, \tilde{\boldsymbol{p}}_j}{\sqrt{a_{j,t}^2 + b_{j,t}^2}} \right)^2, \tag{3}$$

where $(a_{j,t}, b_{j,t}, c_{j,t})^\top = E_{j,t}\, \tilde{\boldsymbol{p}}_j$.

Modification of parameters $R_{t-1,t}, \boldsymbol{T}_{t-1,t}$ so that modified parameters minimize (3) guarantees that the epipolar constraints derived from the fixation point and the last n frames are more strictly satisfied than before. We remark that our modification is applied only to $R_{t-1,t}, \boldsymbol{T}_{t-1,t}$; the number of parameters to be modified is independent of n. Accordingly, the computational cost required for the modification does not change even if we change the number of last frames to be applied for the bundle adjustment.

As described above, our modification of estimated parameters, i.e., applying the bundle adjustment in each step only to last several frames, keeps geometrical consistency with last frames captured so far. With this modification, errors involved in the parameters before the modification are reduced. This modification thus suppresses the amount of accumulated errors even for incremental estimations and significant improvement of accuracy in estimation for a long trajectory is expected. We finally remark that we minimize the cost function using a nonlinear optimization algorithm because (3) is nonlinear with respect to $R_{j,t}$ and $\boldsymbol{T}_{j,t}$.

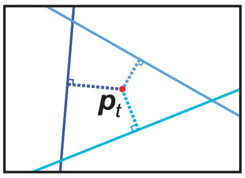


Figure 4: Distance between the fixation point and its corresponding epipolar lines.

4 EXPERIMENTS

We conducted experiments to estimate ego-motion using the proposed method. We first quantitatively evaluate improvement in reducing estimation errors using simulated data. We also evaluate the number of frames to which our local bundle adjustment should be applied to reduce errors most effectively. We finally examine the proposed method using real images.

4.1 Numerical Evaluation Using Simulated Data

We here test the proposed method using simulated data. To see the effectiveness of the proposed method, we implemented the method proposed by Sugimoto–Ikeda (Sugimoto and Ikeda, 2004), called the *comparison method* hereafter, and then evaluated how much estimation errors are reduced by comparing the two methods.

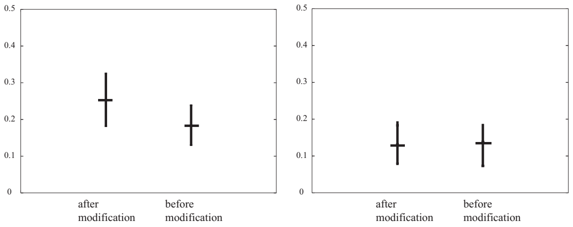
We first test the case of $n = 3$ in the previous section, i.e., the case where the bundle adjustment is locally applied to last three frames and compare errors estimated by the two methods.

The parameters used in the simulation are as follows. Two cameras are set with the baseline distance of 1.0 where each camera is with 21° angle of view and with the focal length[2] of 0.025. The two cameras are in the same pose: the orientation of the camera coordinate systems is the same. The size of images captured by the cameras is 512×512 pixels.

We set a distance between two fixation points to be 10.0 and generated two fixation points in 3D for the two cameras satisfying the distance. We note that the depth of each fixation point was set 10.0 forward from the projection center. We also randomly generated 3 points in 3D nearby each fixation point for the optical-flow computation. The images of the two points were

[2] When the baseline distance is 20cm, then the focal length is 0.5cm, for example.

errors



(a) translations (b) rotations

Figure 5: Estimation errors in the case where the distance between fixation points is 10 baseline.

within the window of 20×20 pixels whose center is the image of the fixation point.

To obtain three time-series images, we generated two steps of motion. For the first step, we randomly generated a translation vector with the length of 0.25 and a rotation matrix where the rotation axis is vertical and the rotation angle is within 1.0 degree. For the second step, on the other hand, we rotated the translation vector generated in the first step around the Z axis of the camera coordinate system where the rotation angle was randomly selected between -1.0^{o} and 1.0^{o}. As for the rotation of the second step, we just added rotation angle randomly selected between -1.0^{o} and 1.0^{o} to the rotation angle of the first step.

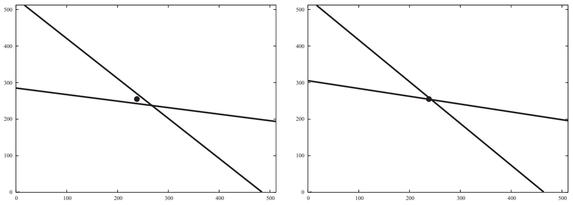
Before and after the motion, we projected all the points generated in 3D onto the image plane to obtain image points that were observed in terms of pixels. To all the image points, we added Gaussian noise. Namely, we perturbed the pixel-based coordinates in the image by independently adding Gaussian noise with the mean of 0.0 pixels and the standard deviation of 2.0 pixels. Next, we computed optical flow of the points generated nearby the fixation points to obtain the flow vectors. We then applied our algorithm to obtain ego-motion estimation: rotation matrix $\widehat{R}$ and translation vector $\widehat{\boldsymbol{T}}$.

To evaluate errors in estimation, we computed rotation axis $\widehat{\boldsymbol{r}}$ ($||\widehat{\boldsymbol{r}}|| = 1$) and rotation angle $\widehat{\theta}$ from $\widehat{R}$. We then defined the evaluation of errors of $\widehat{R}$ and $\widehat{\boldsymbol{T}}$ by

$$\frac{||\hat{\theta}\hat{\boldsymbol{r}} - \theta\boldsymbol{r}||}{|\theta|}, \quad \frac{||\hat{\boldsymbol{T}} - \boldsymbol{T}||}{||\boldsymbol{T}||},$$

where $\boldsymbol{r}$ ($||\boldsymbol{r}|| = 1$) is the rotation axis and θ is the rotation angle of the ground truth, and $\boldsymbol{T}$ is the true translation vector ($||\boldsymbol{T}|| = 0.25$). We iterated the above procedures 200 times and computed the average and the standard deviation of the errors over the 200 iterations.

The result is shown in Fig. 5. We see that estimation errors are actually reduced by the proposed method. From Fig. 5 (b), we observe little difference in estimation errors between the proposed method and the comparison method. As for the rotation estimation, the comparison method is reported to realize highly accurate estimation (Sugimoto and Ikeda, 2004). This is why little improvement is observed. Fig. 5 (a), on the other hand, shows the proposed method actually reduces the average and also the standard deviation of estimation errors. To verify whether or not this difference makes sense from the statistical point of view, we employed Welch's test (Welch, 1938) with significance level of 5%. We then certified that the difference is statistically significant.



(a) before modification (b) after modification

Figure 6: Fixation point and its corresponding epipolar lines before/after parameter modification.

We drew epipolar lines for a case among the 200 cases above, which are illustrated in Fig. 6. Fig. 6 demonstrates that the distance between the epipolar lines and the fixation point in the image becomes smaller after the modification. In fact, we computed the distances between the fixation point and the epipolar lines and found that they are 10 pixels and 0.2 pixels before and after the modification, respectively. We confirmed that in almost all the 200 cases, the distances between the fixation point and epipolar lines are the same degree as this typical case.

We verified that values of the cost function actually become smaller after the modification for all the 200 cases. A smaller value of the cost function, however, does not always mean that more accurate estimation is realized. This is because the case exists where a rotation error and a translation error can cancel each other in terms of the cost function. This implies that however close the cost function achieves zero, estimation errors can exist and that we have a limitation of reducing estimation errors. Considering these factors, we may conclude that our modified estimation becomes geometrically consistent with last two frames captured so far and that our proposed method significantly improves accuracy in estimation, with keeping the method incremental.

Next, we evaluated the number of frames to which the local bundle adjustment should be applied. In other words, we evaluated the relationship between estimation errors and the number of frames for applying the local bundle adjustment and identified the optimal number under the criterion of reducing estimation errors most effectively. For the cases of $3 \leq n \leq 10$, we conducted experiments under the same condition as described above. We generated 10 steps of motion here. We note that since we have al-

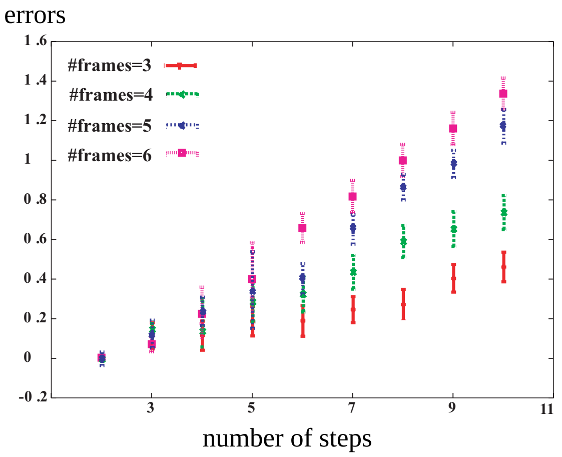
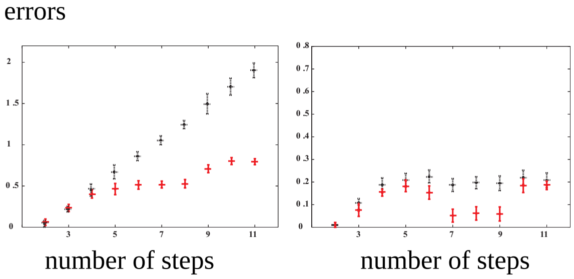


Figure 7: The number of frames to which the local bundle adjustment is applied and estimation errors.



(a) translation vectors (b) rotation matrices

Figure 8: Accumulated errors when the local bundle adjustment with 3 frames is applied (read means with correction and black means without correction).

ready observed little improvement in accuracy of rotation estimation, we here focused on translation estimation only.

The results for the cases of $3 \leq n \leq 6$ are illustrated in Fig. 7. In each case, we did not conduct modification until required number of images are obtained for the bundle adjustment. For example, in the case of $n = 5$, the bundle adjustment was applied only after the 4th step of motion.

Figure 7 shows that estimation accuracy is best in the case of $n = 3$. Though we might expect that estimation accuracy increases with the number of images to which the bundle adjustment is applied, it is not true. This is because when the number of images is large, accumulation of estimation errors incurred by that time already becomes sufficiently large and the constraints imposed for the modification is not reliable any more. We thus observe that the amount of accumulated errors is too large to significantly reduce. This discussion is supported by the fact that the case of $n = 3$ is most effective. Accordingly, we have to apply the bundle adjustment by the time when accumulated errors of estimation does not become large.

We now know that three frames are most effective to apply the bundle adjustment. We next evaluated how accumulated errors increase depending on the number of steps of motion in the case of $n = 3$. We conducted the experiment here under the same condition above. We note that 11 steps of motion were generated here.

The results are shown in Fig. 8. We also conducted the same experiments using the comparison method. As for rotation estimation, we do not observe any significant accumulation of errors for the both methods. On the other hand, we easily see that from Fig. 8 (a), accumulation of estimation errors linearly increases with the number of steps for the comparison method while that for the proposed method does not; we see drastic improvement. We may thus conclude that our method keeps estimation accuracy stable even for a long trajectory.

4.2 Trajectory Estimation Using Real Images

We employed the proposed method to estimate an ego-motion trajectory in the real situation. We also employed the comparison method to see the effectiveness of the proposed method.

We used two off-the-shelf cameras (EVI-G20 from Sony) as active cameras and set up the cameras on the stage of a tripod so that the baseline distance the two cameras is about 20cm and the distance of two fixation points is between 2m and 3m (Fig. 9). We assume that the optical axis of each camera is parallel with the Z axis of its camera coordinate system and the world coordinate system is identical with the right-camera coordinate system at the initial position. We then estimated a motion trajectory of the projection center of the right camera.

We moved the tripod with the cameras in the scene. The trajectory of the right-camera motion is shown in Fig. 10. The length of the trajectory was about 5m. We marked 135 points (20 points along each straight line segment and 45 points along each circular segment) on the trajectory and regarded them as samples during the ego motion. (In other words, 135 points were sampled during the ego motion of about 5m.) We then applied the proposed method and the comparison method respectively only to the samples, i.e., the marked points, to estimate the ego motion. We note that in each method, we selected 3 points from each right image and 3 points from each left image and computed the optical-flow vectors of the 6 points for subsequent computations.

In each image captured by each camera at the starting point of the motion, we manually selected a point to serve as the fixation point. During the estimation, we updated fixation points 7 times for each camera in the case where the current fixation point disappears from the image. This updating was also conducted by

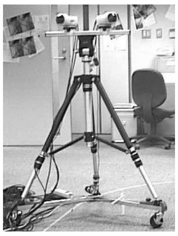


Figure 9: Two camera setup.

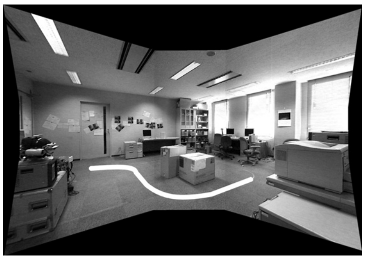


Figure 10: Experimental environment.

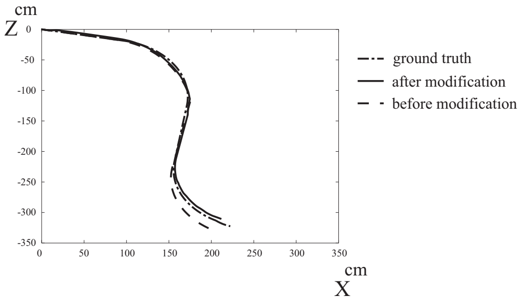


Figure 11: Estimated camera positions.

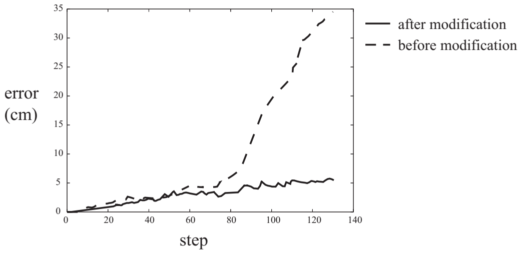


Figure 12: Estimation errors of the camera position.

hand. We computed optical flows within the window of 100×100 pixels whose center is the image of the fixation point. We used three optical-flow vectors for each camera (we thus used six optical-flow vectors in total). The vectors were randomly selected within the windows of 30×30 pixels whose center is the image of the fixation point. In computing optical flows, we used the Kanade-Lucas-Tomasi algorithm (Lucas and Kanade, 1981; Tomasi and Kanade, 1991). In the proposed method, we independently applied the local bundle adjustment to each camera images where last three frames with their fixation points are used.

Under the above conditions, we estimated the right-camera motion at each marked point. The trajectory of the right-camera motion was obtained by concatenating the estimated motions at the marked points. Fig. 11 shows the estimated trajectory that is projected on the XZ-plane of the world coordinate system. Fig. 12 shows errors in position estimation in each step for the proposed method and for the comparison method. In Figs. 11, 12, the solid line and the dotted line respectively indicate the result by the proposed method and that by the comparison method. The broken dotted line in Fig. 11, on the other hand, indicates the ground truth. Table 1 shows the average and the standard deviation of errors in position estimation over the marked 130 points.

Figure 11 shows that the proposed method more accurately estimates the trajectory than the comparison method. We observe that from Fig. 12 the comparison method fails in estimation around the 80th step (the beginning of the second circular part of the trajectory) and begins to have aberration from the actual trajectory. The proposed method, on the other hand, succeeds in reducing estimation errors incurred in the step and, as a result, keeps accurate and stable estimations of the subsequent trajectory. The position error at the terminal marked point was about 35cm for the comparison method and about 4cm for the proposed method. We see that estimation accuracy is greatly improved and that this improvement leads to reducing accumulation of estimation errors. Table 1 also validates the effectiveness of the proposed method; not only the average of estimation errors in each step but also the standard deviation becomes the smaller. We can thus conclude that the proposed method realizes accurate and stable ego-motion estimation even for a long trajectory.

5 CONCLUSION

The ego-motion estimation method using the binocular independent fixation control (Sugimoto et al., 2004; Sugimoto and Ikeda, 2004) involves a problem that estimation accuracy gradually becomes worse as the motion trajectory becomes longer and longer because of accumulation of errors incurred in each estimation step. To avoid falling in such a problem, we proposed a method that incorporates incremental modification of estimated parameters. Namely, the proposed method applies the bundle adjustment in each step only to last several frames so that geometrical consistency, i.e., the epipolar constraint, between the last frames and the fixation points is guar-

Table 1: The average and the standard deviation of estimation errors over 130 steps.

	before modification	after modification
average [cm]	0.268	0.0394
standard deviation [cm]	0.102	0.00739

anteed. With this modification, errors involved in the parameters before the modification are drastically reduced and, therefore, significant improvement of accuracy in estimation for a long trajectory is realized, with keeping the method incremental.

The proposed method focuses on the fixation point for evaluating geometrical consistency, however, using other feature points together may improve estimation accuracy further. An adaptive application of the bundle adjustment, i,e., effectively selecting steps to which the local bundle adjustment is applied, will be promising to reduce computational cost. These investigations are our future direction.

ACKNOWLEDGEMENTS

This work is in part supported by Grant-in-Aid for Scientific Research of the Ministry of Education, Culture, Sports, Science and Technology of Japan under the contract of 13224051 and 18049046.

REFERENCES

H. Aoki, B. Schiele and A. Pentland (2000): *Realtime Personal Positioning System for Wearable Computers*, Vision and Modeling Technical Report, TR-520, Media Lab. MIT.

B. Clarkson, K. Mase and A. Pentland (2000): *Recognizing User's Context from Wearable Sensors: Baseline System*, Vision and Modeling Technical Report, Vismod TR-519, Media Lab. MIT.

A. J. Davison, W. W. Mayol and D. W. Murray (2003): Real-Time Localization and Mapping with Wearable Active Vision, *Proc. of ISMAR*, pp. 18–27.

A. J. Davison and D. W. Murray (1998): Mobile Robot Localization using Active Vision, *Proc. of ECCV*, Vol. 2, pp. 809–825.

G. N. DeSouza and A. C. Kak (2002): Vision for Mobile Robot Navigation: A Survey, *IEEE Trans. on PAMI*, Vol. 24, No. 2, pp. 237–267.

M. W. M. G. Dissanayake, P. Newman, S. Clark, H. F. Durrant-Whyte and M. Csorba (2001): A Solution to the Simultaneous Localization and Map Building (SLAM) Problem, *IEEE Trans. on RA*, Vol. 17, No. 3, pp. 229–241.

N. Gonçalves and H. Araújo (2002): Estimation of 3D Motion from Stereo Images, *Proc. of ICPR*, Vol. I, pp. 335–338.

J. E. Guivant and E. Nebot (2001): Optimization of the Simultaneous Localization and Map-Building Algorithm for Real-Time Implementation, *IEEE Trans. on RA*, Vol. 17, No. 3, pp. 242–257.

R. Hartley and A. Zisserman (2000): *Multiple View Geometry in Computer Vision*, Cambridge Univ. Press.

B. D. Lucas and T. Kanade (1981): An Iterative Image Registration Technique with an Application to Stereo Vision, *Proc. of IJCAI*, pp. 674–679.

N. Molton and M. Brady (2003): Practical Structure and Motion from Stereo When motion is Unconstrained, *Int. J. of Computer Vision*, Vol. 39, No. 1, pp. 5–23.

T. Nakagawa, T. Okatani and K. Deguchi (2004): Active Construction of 3D Map in Robot by Combining Motion and Perceived Images, *Proc. of ACCV*, vol. 1, pp. 563–568.

A. Sugimoto, W. Nagatomo and T. Matsuyama (2004): Estimating Ego Motion by Fixation Control of Mounted Active Cameras, *Proc. of ACCV*, pp. 67–72.

A. Sugimoto and T. Ikeda (2004): Diverging Viewing-Lines in Binocular Vision: A Method for Estimating Ego Motion by Mounted Active Cameras, *Proc. of the 5th Workshop on Omnidirectional Vision, Camera Networks and Non-classical Cameras*, pp. 67–78.

Y. Sumi, Y. Ishiyama and F. Tomita (2004): 3D Localization of Moving Free-Form Objects in Cluttered Environments, *Proc. of ACCV*, vol. 1, pp. 43-48.

S. Thrun (2002): Robotic Mapping: A Survey, *Exploring Artificial Intelligence in the New Millennium*, Morgan Kaufmann.

C. Tomasi and T. Kanade (1991): *Detection and Tracking of Point Features*, CMU Technical Report, CMU-CS-91-132.

B. Triggs, P. McLauchlan, R. Hartley and A. Fitzgibbon (2000): Bundle Adjustment – A Modern Synthesis, *Vision Algorithms: Theory and Practice (B. Triggs, A. Zisserman and R. Szeliski eds.)* LNCS, Vol. 1883, pp. 298–372, Springer.

B. L. Welch (1938): The Significance of the Difference between Two Means when the Population Variances are Unequal, *Biometrika*, Vol. 29, pp. 350–362.

M. Werman, S. Banerjee, S. Dutta Roy and M. Qiu (1999): Robot Localization Using Uncalibrated Camera Invariants, *Proc. of CVPR*, Vol. II, pp. 353–359.

Z. Zhang and Y. Shan (2001): *Incremental Motion Estimation through Local Bundle Adjustment*, Technical Report MSR-TR-01-54, Microsoft Research.

Z. Zhang and Y. Shan (2003): Incremental Motion Estimation through Modified Bundle Adjustment, *Proc. ICIP*, Vol.II, pp.343–346.