An Iterative Method for Improving *Bas-relief* Ambiguity

Sung-Kee Park*, Munsang Kim*, In So Kweon **

*Korea Institute of Science and Technology, Advanced Robotics Research Center
P. O. Box 131 Cheongryang, Seoul 130-650, Korea
skee@kist.re.kr, munsang@kist.re.kr
**Korea Advanced Institute of Science and Technology, Dept. of Electrical Engineering
Kusong-dong 373-1, Yusong-gu, Taejon 305-701, Korea
iskweon@kaist.ac.kr

Abstract

When using small motion model, the problem of "structure from motion" has inevitably the ambiguity between translation and rotation motion. For solving it, conventional methods adopt a single line methodology; first they try to find exact corresponding points with only image brightness, and then motion parameters and scene depths are estimated on the basis of it. But, considering that the previous corresponding methods may have large errors and it is difficult to define its error model, the single line methods cannot improve such ambiguity although they introduce any robust statistical estimator. Therefore, on the assumption that corresponding points and motion estimation have to be iteratively refined, we propose a new method for improving those ambiguities with stereo image sequence.

1. Introduction

For visual reconstruction and robot navigation, structure from motion (SFM) is an active issue in computer vision community and its properties are also becoming well understood. As a drawback in SFM, it is well known that the SFM methods, using small motion model such as optical flow based method and direct method, have inevitably motion ambiguity between translation and rotation, which is called *bas-relief ambiguity*. In this paper we present an iterative method for improving those ambiguities, which is based on the robust direct method using stereo image sequence [1].

As for the bas-relief ambiguity, Weng [2] implemented its error modeling with Cramer-Rao bounds and tested the limits of small motion with various simulations and experiments. Recently, Szeliski and Kang [3] performed its analysis based on the singularities of the Hessian of the information matrix. As tools for overcoming such ambiguities, some methods also have been suggested in computer vision community: they used planar-parallax model [4] and linear transformations [5]. Soatto and Brockett [6] recently suggested a probably convergent algorithm that is robust to large noise, and showed that the bas-relief ambiguity is intrinsic to SFM.

However, when we consider the problem of SFM from practical point of view, it is most important how we obtain exact corresponding feature points. As noted in [6], the previous corresponding methods, such as optical flow or corner finding method, can have large locating errors, and its error distribution cannot be modeled [1]. Therefore, the previous linear methods cannot cope well with its ambiguity. Moreover, when we consider that the shape of cost function under the bas-relief ambiguity can be like narrow diagonal valley [7], we cannot improve the errors with only any non-linear optimization method although it introduces robust statistics. It again means that it is the first settlement for overcoming its ambiguity to achieve exact corresponding feature points as much as possible, of which error distribution cannot be modeled beforehand.

In this paper for improving the estimation error due to bas-relief ambiguity, we basically are to use a robust and direct motion estimation method [1]. It gives a lot of approximated corresponding points. Therefore, based on it, we proposed a new method that can implement stereo and motion correspondence under the sub-pixel accuracy, and then we can refine motion and depth estimates with robust statistical method again.

The next section describes the details of an iterative refinement process and overall algorithm. In section 3, we show the experimental results, and we make conclusions in section 4.

2. A new refinement method

In Fig.1, we show the consequent stereo camera motion and basic notations for this paper. In calibrated stereo camera, its consecutive motion estimation is to estimate parameters **w** and **t**, and it can be achieved by a robust and direct method using the following equation [1]. In this paper, the meaning of *robust* is that estimates are insensitive to the error due to the large departure of small data.

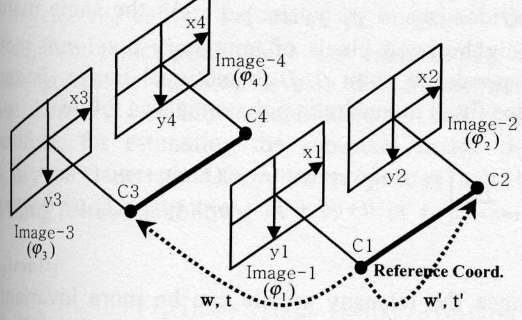


Figure 1: *Consecutive stereo motion*

$$\min_{\mathbf{a}} \sum_{i} \rho(r_i, \sigma), \quad r_i = \mathbf{C}_i \mathbf{a} + d_i \tag{1}$$

$$\mathbf{a} = [t_x \;\; t_y \;\; t_z \;\; w_x \;\; w_y \;\; w_z]^T$$

$$\mathbf{C}_i = [(\nabla I_i^T \mathbf{B}_i \mathbf{w}' + I_{ti}')\nabla I_i^T \mathbf{A}_i \quad -(\nabla I_i^T \mathbf{A}_i \mathbf{t}')\nabla I_i^T \mathbf{B}_i]$$

$$d_i = -(\nabla I_i^T \mathbf{A}_i \mathbf{t}')I_{ti}$$

where i represents each image pixel in reference image (*image-1*) and Lorenztian function [9] is used as estimator function (ρ).

However, because the brightness of all image pixels can't satisfy the *optical flow constraint* that is central assumption of equation (1), it is erroneous to finish the motion estimation process with all image pixels. Therefore, we are to regard above estimation as coarse one, and try to refine the stereo matching and motion correspondence about each pixel in reference image on the basis of its coarse estimation. Fig. 2 shows the graphical representation about finding exact corresponding feature points. As seen in this figure, approximated motion is obtained by the robust and direct method, and those are filtered and refined by following pixel unit matching and sub-pixel unit matching, which is applied to both stereo displacement (between *image-1* and *image-2* in Fig. 1) and motional displacement (between *image-1* and *image-3*). At this time, the exact stereo matching gives more exact depth estimation, and the exact motion correspondences give more precise estimates of motion parameters together with its refined exact depth. Here, the basic equation for those estimates is the following small motion model equation;

$$u = \frac{dx}{dt} = \frac{1}{z}(-t_x + xt_z) + w_z y - w_y(1+x^2) + w_x xy \tag{2}$$

$$v = \frac{dy}{dt} = \frac{1}{z}(-t_y + yt_z) - w_z x + w_x(1+y^2) - w_y xy$$

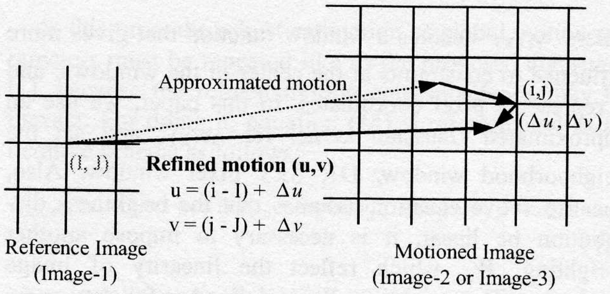


Figure 2: *Graphical expression of refinement process*

2.1 Refinement of stereo matching

When the approximated stereo matching is given by the previous robust technique [1], we can improve the accuracy of stereo matching by the two-step stereo matching; matching in pixel unit and matching in sub-pixel unit.

With respect to stereo matching in pixel unit, we implement it as follows. To begin with, when an approximated depth at any pixel location (I, J) in the reference image is given, we can estimate stereo matching (u, v) by using small motion model of equation (2), and then we can decide a temporary pixel-unit stereo matching point *(i, j)* in *image-2* as follows;

$$(i, j) = \mathit{nearest}\,\mathrm{int}\,(I + u, J + v) \tag{3}$$

However, because the pixel point *(i,j)* is estimated by approximated depth, it may not be true pixel-stereo matching point. Therefore, it is necessary to verify whether any neighborhood pixel is true one. In case of the rectified stereo images, of which extrinsic parameters have not rotational component, it is well known that template matching can be a useful measure for finding stereo matching [8]. Therefore, we use *sum of absolute difference* (SAD) as similarity measure for stereo matching in pixel unit. Then, we can define a true pixel-unit stereo matching point among the neighborhood pixels including temporary pixel-unit stereo matching point, if it satisfies the following two conditions;

$$\min_{(i,j)\in R} SAD(i, j : I, J), \quad SAD(i, j : I, J) \le S_{Threshold} \tag{4}$$

where R represents the size of neighborhood. If a reference image point (I, J) cannot satisfy such matching criteria, it is excluded in refinement process.

When the stereo matching in pixel unit is obtained, we next estimate it in sub-pixel unit. By using region-based matching, we can constitute a weighted least-squares equation for sub-pixel stereo matching as follows;

$$E_{D1}(\Delta u) = \sum_{\mathbf{x}\in D1} \mathbf{W}^2(\mathbf{x})\, \mathbf{W}_P{}^2(\mathbf{x})\,[I_x(\mathbf{x})\cdot \Delta u \;+\; I_t(\mathbf{x})]^2 \tag{5}$$

where $\mathbf{W}(\mathbf{x})$ denotes a window function that gives more influence to constraints at the center of the windows; and $\mathbf{x}$ represents pixel coordinates. In this paper, we use an approximated Gaussian kernel for $\mathbf{W}(\mathbf{x})$, and set the neighborhood window, D1, 3×3 pixel window. Also, because above equation assumes that the brightness distribution be linear, it is necessary to impose another weighting, $\boldsymbol{W_p}$, which reflect the linearity of image brightness. The weighting, $\boldsymbol{W_p}$, is defined as follows;

$$W_P(i,j) = \frac{1}{c_1 \;+\; c_2 \times planeness(i,j)} \tag{6}$$

where, $planeness(i,j) = \dfrac{\mathbf{a}(i,j)\times\mathbf{b}(i,j)}{|\,\mathbf{a}(i,j)\times\mathbf{b}(i,j)\,|}\cdot\dfrac{\mathbf{c}(i,j)}{|\,\mathbf{c}(i,j)\,|}$

$$\mathbf{a}(i,j) = [1 \quad 0 \quad (I_{\varphi1}(i+1,j) - I_{\varphi1}(i,j))]^T$$
$$\mathbf{b}(i,j) = [0 \quad 1 \quad (I_{\varphi1}(i,j+1) - I_{\varphi1}(i,j))]^T$$
$$\mathbf{c}(i,j) = [1 \quad 1 \quad (I_{\varphi1}(i+1,j+1) - I_{\varphi1}(i,j))]^T$$

Therefore, we can estimate an exact stereo matching by using equation (4) and (5) as follows;

$$u = (i - I) \;+\; \Delta u \tag{7}$$

Then, a refined depth, z, can be obtained when equation (7) is applied to equation (2).

2.2 Refinement of motion correspondence

About the pixels in reference image that satisfy above sub-pixel stereo matching, we next implement motion correspondence between *image-1* and *image-3* in Fig.1. At this time, motion correspondences are also executed by two steps such as section 2.1; correspondence in pixel unit and in sub-pixel unit.

With respect to motion correspondence in pixel units, we implement it as follows. To begin with, about the image pixel (I, J) satisfying sec 2.1, we can estimate motion displacement (u, v) by using equation (2). Then, we can decide a temporary pixel-unit motion correspondence point *(i, j)* like as equation (3).

However, because the pixel point *(i, j)* may not be true pixel-unit motion correspondence point about the point (I, J), it is necessary to verify whether any neighborhood pixel is true one like as stereo matching. For this process, we propose a new method for finding true pixel-unit motion correspondence point by fitting image intensities to a quadratic polynomial. That is because region-based matching method used in stereo matching is poor in rotational motion [8] as shown in Fig. 3 of section 3.

For a pixel point (I, J) of reference image, its intensity can be fitted to a quadratic polynomial by using neighborhood pixels as follows;

$$\min_{\mathbf{p}(I,J)} \sum_{(m,n)\in D2} \begin{pmatrix} p_0\cdot(I+m)^2 \;+\; p_1\cdot(I+m)(J+n) \;+\; p_2\cdot(J+n)^2 \\ +\; p_3\cdot(I+m) \;+\; p_4\cdot(J+n) \;+ p_5 \;-\; I_{\varphi1}(I+m,J+n) \end{pmatrix}^2 \tag{8}$$

where the coefficient vector of polynomial are defined by $\mathbf{p}(I,J) = [p_0 \;\; p_1 \;\; p_2 \;\; p_3 \;\; p_4 \;\; p_5]^T$. As the same manner, the neighborhood pixels of temporary pixel-unit motion correspondence point *(i, j)* in motional image (*image-3*) are also fitted to quadratic polynomials as follows;

$$\min_{\mathbf{p}(i,j)} \sum_{(m,n)\in D2} \begin{pmatrix} p_0\cdot(i+m)^2 \;+\; p_1\cdot(i+m)(j+n) \;+\; p_2\cdot(j+n)^2 \\ +\; p_3\cdot(i+m) \;+\; p_4\cdot(j+n) \;+ p_5 \;-\; I_{\varphi3}(i+m,j+n) \end{pmatrix}^2 \tag{9}$$

Since the intensity profile can be more invariant to small rotation of image intensities, we can compare the grouped coefficients of intensity polynomials for similarity criteria as follows;

$$\begin{aligned} PMC_1(i,j\,;I,J) &= \sum_{k=0}^{2} |\, p_{(I,J)}(k) - p_{(i,j)}(k)\,| \\ PMC_2(i,j\,;I,J) &= \sum_{k=3}^{4} |\, p_{(I,J)}(k) - p_{(i,j)}(k)\,| \\ PMC_3(i,j\,;I,J) &= |\, p_{(I,J)}(5) - p_{(i,j)}(5)\,| \end{aligned} \tag{10}$$

Then, we can define a true pixel-unit motion correspondence point *(i,j)* among the neighborhood pixels including a temporary pixel-unit motion correspondence point, if it satisfies the following three conditions;

$$\begin{aligned} &\min_{(i,j)\in R} PMC_1(i,j\,;I,J), \quad \min_{(i,j)\in R} PMC_2(i,j\,;I,J), \\ &PMC_3(i,j\,;I,J) \le M_{Threshold} \end{aligned} \tag{11}$$

where R represents the size of neighborhood (5x5) and the value of $M_{Threshold}$ is set to 2.

After the motion correspondence in pixel unit is obtained, we next estimate it in sub-pixel unit. Here, we propose a new method using a region-based matching, which is based on optical flow constraint equation like equation (5). However, the previous methods [8], such as constant or affine motion model, are inexact fundamentally, which is because such image motion model cannot fully reflect true motion. Especially, the previous methods have severe errors when rotational motion is dominant as shown in Fig. 3 of section 3. Therefore, when we combine the optical flow constraint equation with small motion model on the pixel-unit matching, then we can obtain following equation.

$$I_x\Delta u + I_y\Delta v + co_1 t_z + co_2 w_x + co_3 w_y + co_4 w_z + I_t = 0 \quad (12)$$

$$\text{where,}\quad \begin{aligned} co_1 &\equiv (\Delta x \cdot I_x + \Delta y \cdot I_y)/z \\ co_2 &\equiv (x\Delta y + y\Delta x + \Delta x\Delta y)\cdot I_x + (2y\Delta y + \Delta y\Delta y)\cdot I_y \\ co_3 &\equiv -(x\Delta y + y\Delta x + \Delta x\Delta y)\cdot I_x - (2x\Delta x + \Delta x\Delta x)\cdot I_y \\ co_4 &\equiv \Delta y \cdot I_x - \Delta x \cdot I_y \end{aligned}$$

In above equation, $t_z, w_x, w_y, and\ w_z$ represent global motion parameters by global optimization, and $co_1, co_2, co_3,\ and\ co_4$ are calculated by image coordinates and depth at center pixel which was estimated by stereo matching. Above equation is also under-constraint because there are 6 unknown parameters for one equation. Therefore, for estimating the sub-pixel image motion $(\Delta u, \Delta v)$ at center pixel *(i,j)*, we propose a new block matching method as follows;

$$E_{D2}(\mathbf{mp}) = \sum_{\mathbf{x}\in D2} \mathbf{W}^2(\mathbf{x})\,\mathbf{W}_P{}^2(\mathbf{x}) \begin{pmatrix} I_x(\mathbf{x})\Delta u + I_y(\mathbf{x})\Delta v + co_1(\mathbf{x})t_z + co_2(\mathbf{x})w_x \\ + co_3(\mathbf{x})w_y + co_4(\mathbf{x})w_z + I_t(\mathbf{x}) \end{pmatrix}^2 \quad (13)$$

where, $\mathbf{mp} = [\,\Delta u \;\; \Delta v \;\; t_z \;\; w_x \;\; w_y \;\; w_z\,]^T$

At above equation, $\mathbf{W}$ and $\mathbf{W}_p$ are the same ones of equation (5). Here, we can define the equivalent temporal derivative of image intensity as follows;

$$I_{t,E}(\mathbf{x}) \equiv I_t(\mathbf{x}) + co_1(\mathbf{x})t_z + co_2(\mathbf{x})w_x + co_3(\mathbf{x})w_y + co_4(\mathbf{x})w_z \quad (14)$$

where four motion parameters are assumed to be given in the previous robust method initially. Therefore, we propose a block matching method with small motion model for estimating motion correspondence in sub-pixel unit as follows;

$$E_{D2}(\Delta u, \Delta v) = \sum_{\mathbf{x}\in D2} \mathbf{W}^2(\mathbf{x})\,\mathbf{W}_P{}^2(\mathbf{x}) \big(I_x(\mathbf{x})\cdot\Delta u + I_y(\mathbf{x})\cdot\Delta v + I_{t,E}(\mathbf{x})\big)^2 \quad (15)$$

When the motion correspondence is estimated with both pixel unit and sub-pixel unit, the refined motion correspondence of the center pixel (I, J) is as follow;

$$\begin{aligned} u &= (i - I) + \Delta u \\ v &= (j - J) + \Delta v \end{aligned} \quad (16)$$

Therefore, if the refined motion correspondences and refined depths are given, it is possible to formulate a robust estimation of motion parameters as follows;

$$\min_{\mathbf{t},\mathbf{w}} \sum_i \rho\left(\frac{1}{z}\mathbf{A}\cdot\mathbf{t} + \mathbf{B}\cdot\mathbf{w} - \mathbf{u},\ \sigma\right) \quad (17)$$

where,

$$\mathbf{A} = \begin{bmatrix} -1 & 0 & x \\ 0 & -1 & y \end{bmatrix} \quad and \quad \mathbf{B} = \begin{bmatrix} xy & -(1+x^2) & y \\ (1+y^2) & -xy & -x \end{bmatrix}$$

At this time, the robust estimation of global motion parameters must be repeated like as the proposed algorithm [1], because the refined motion correspondence may be inexact. For doing it, equation (14) is reformulated as an iterative scheme as follows;

$$I_{t,E,\,n+1}(\mathbf{x}) \equiv I_t(\mathbf{x}) + co_1(\mathbf{x})t_{z,n} + co_2(\mathbf{x})w_{x,n} + co_3(\mathbf{x})w_{y,n} + co_4(\mathbf{x})w_{z,n} \quad (18)$$

where n denotes the iteration step.

2.3 Proposed Algorithm

As a new method that refines depths and the global motion parameters of calibrated stereo image sequence, our algorithm can be summarized as follows;

Step 1: Do the algorithm in the proposed method [1], which used multi-resolution image level.
At the lowest pyramid level,

Step 2: Select stereo matching points that satisfy equation (4)

Step 3: By using equation (5) and (7), estimate depths about the points of previous step.

Step 4: Among the points satisfying previous step, select motion correspondence points that satisfy equation (11).

Step 5: By using equation (15) and (16), estimate motion correspondence (u, v) about the points of previous step.

Step 6: Estimate motion parameters by using equation (17)

Step 7: Estimate equivalent temporal derivatives of image brightness by using equation (18)

Step 8: Repeat step 5, 6, and 7 as long as the sum of residual errors of equation (17) decreases.

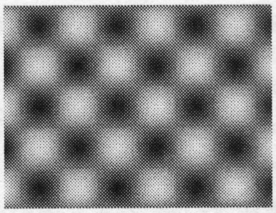
Step 9: After refining motion parameters in *step 8*, refine the depth parameters with equation (17).

3. Experiments

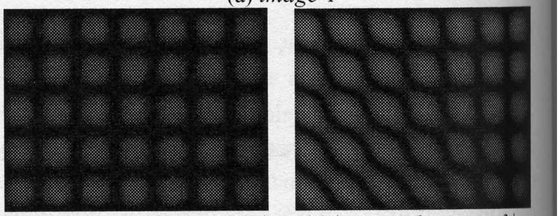
We experimented with synthetic and real consecutive stereo image sequences. For testing the motion correspondence of the proposed method, a synthetic image sequence is used as shown in Fig 3, which has constant depth and its stereo motion is $t_x = t_y = t_z = 50$, $w_x = w_y = w_z = 1°$. In this figure, (d) is the estimation result of pixel unit motion correspondence after implementing stereo matching with sub-pixel accuracy. Experimental results show that the proposed polynomial fitting method improves the estimation accuracy under 3D rotational motion. About the sub-pixel correspondence accuracy, (e) in Fig. 3 shows that the proposed small motion model can prominently improve the correspondence accuracy when compared with previous SAD and u-v constant model. In all experiments of Fig. 3, about 25000 image pixels and 3×3 pixel area is used for corresponding process.

Fig. 4 shows the real test image sequence, where the extrinsic parameters of stereo camera are $t_x = 8mm$, $t_y = t_z = w_x = w_y = w_z = 0$. For constituting the experimental setup causing bas-relief ambiguity, we made-up the stereo camera system having small field of view, of which the focal length of lens is 16mm and the maximum field of view is 23°. In Table 1, the estimated motion parameters by both the robust method [1] and the proposed robust and refinement method (R&R method) are shown. In the table, we can observe that the robust method causes motion ambiguities severely, which is, we think, due to the small depth variation in the scene. Intrinsically, such scene condition is apt to bring about bas-relief ambiguity. And, we can observe that the proposed R&R method can refine the ambiguities. Also, when the iterative refinement process is applied to R&R method like as the robust algorithm [1], the ambiguities of motion parameters are refined by iterations. In Fig 5, it is shown that *Terr* and *Werr*, of which each means translational and rotational error, become smaller as the absolute average of residual errors by equation (17) is smaller, where the absolute average of residual errors has the minimum value at the 7th iteration. Considering the above all experiments, we could observe that the previous robust method also weak for bas-relief ambiguity and the proposed iterative method can be effective for improving the stability in near-ambiguous situations.

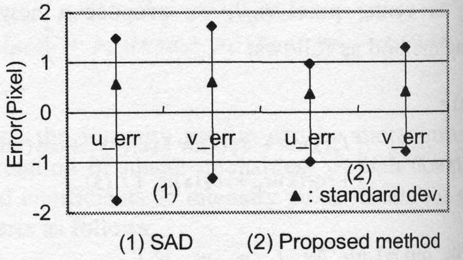
We also obtained the multiple stereo image sequences with the experimental setup of Fig. 4, and its experimental results are shown in Fig. 6. In this figure, it is shown that rotational estimation errors cause much more errors about translation estimates. The 1st, 4th, and 7th frame in Fig. 6 is the case that bas-relief ambiguity is generated. In this figure, we can observe that the proposed R&R method can improve the estimation accuracy under such ambiguity. In Fig. 7, the depth estimation maps, which are the 4th frame in Fig. 6, are shown as an example. The left map is the result in case of that only robust method [1] is applied, where white or black area displays erroneous depth estimates. The right map shows the refined depth estimates by the proposed algorithm. Comparing two depth maps, we can observe that the proposed method also refine the depth estimates.



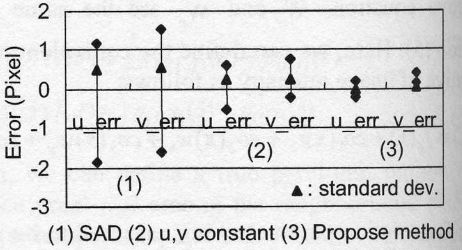
(a) *image-1*



(b) | *image-1* – *image-2* | (c) | *image-1* – *image-3* |



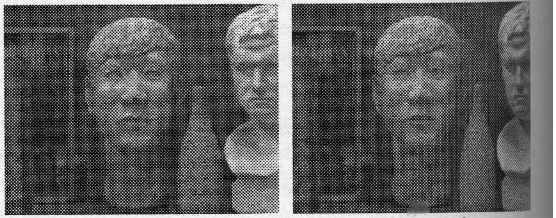
(d) Motion correspondence error (pixel unit)



(e) Motion correspondence error (sub-pixel unit)

Figure 3: *Motion correspondence with synthetic image*

(a) Reference image (*image-1*)



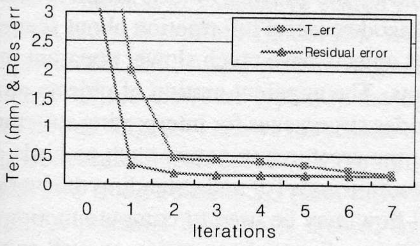
(b) *image-2* (c) *image-3*

Figure 4: *Real experimental stereo image sequence*

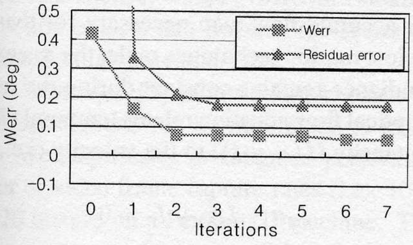
Table 1: *Motion estimation about Figure 4.*

Parameter	*A	B	C	D
t_x(mm)	0.00	6.66	1.96	0.20
t_y(mm)	0.00	-0.10	-0.40	0.09
t_z(mm)	0.00	1.24	0.55	0.27
w_x(Deg)	0.00	-0.00	-0.02	0.00
w_y(Deg)	0.50	0.08	0.34	0.44
w_z(Deg)	0.00	-0.03	-0.02	-0.02

*A : Exact motion parameter
B: Robust method only global optimization
C: Proposed method (1st iteration)
D: Proposed method (7th iteration)

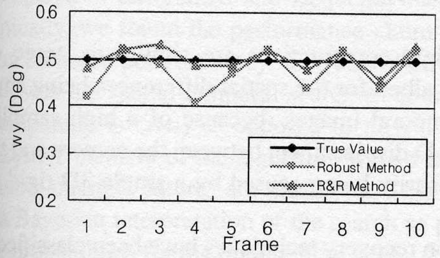


(a) Translational error

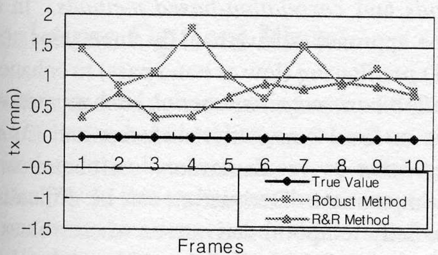


(b) Rotational error

Figure 5: Estimation errors according to iteration

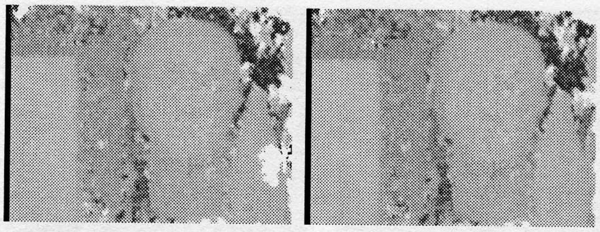


(a) Estimation of w_y



(b) Estimation of t_x

Figure 6: *Motion estimation with multiple frames*



(a) Previous robust method [1] (b) Proposed method

Figure 7: *Depth estimation about 4th frame in Figure 6*

4. Conclusion

In this paper, based on the robust and direct method using stereo image sequence, we proposed a new iterative method for improving bas-relief ambiguity. Its important feature is to refine the stereo and motion correspondence in sub-pixel accuracy, which uses the robust and direct estimation as initial guess. With simulation and real experiments, we showed that the proposed method could improve estimation accuracy. As future research, we will deal correctly with wrong correspondence at occlusion boundaries and convergence properties about our iterative method.

References

[1] S. K. Park and In So Kweon, "Robust and direct estimation of 3-D motion and scene depth from stereo image sequence", Pattern Recognition, pp. 1713-1728, vol. 34(9), 2001.

[2] J. Weng, N. Ahuja and T. S. Huang, "Optimal Motion and Structure Estimation", IEEE Trans. on Pattern Analysis and Machine Intelligence, 15(9):864-884, 1993.

[3] R. Szeliski and S. B. Kang, "Shape Ambiguities in Structure From Motion", IEEE Trans. on Pattern Analysis and Machine Intelligence, 19(5):506-512, 1997.

[4] M. Irani, B. Rousso and S. Peleg, "Recovery of Ego-Motion Using Region Alignment", IEEE Tr. On Pattern Analysis and Machine Intelligence, Vol. 19, No. 3, pp. 268-272, 1997.

[5] A. Jepson and D. Heeger, Linear Space Method for Recovering Rigid Motion, Spatial Vision in Humans and Robots, Cambridge University Press, 1992.

[6] S. Soatto and R. Brockett, "Optimal Structure from Motion: Local Ambiguities and Global Estimates", CVPR, pp. 282-288, 1998.

[7] G. P. Stein and A. Shahsua, "Model-Based Brightness Constraints: On Direct Estimation of Structure and Motion", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 22, No. 9, pp. 992-1015, 2000.

[8] A. M. Tekalp, Digital Video Processing, Prentice Hall PTR, Upper Saddle River, NJ 07458, 1995.

[9] P. J. Rousseeuw and A. M. Leroy, Robust Regression and Outlier Detection, John Wiley & Sons, New York, 1987.