Relative Scale Estimation between Two Camera Motions

Yekeun Jeong and In-So Kweon
RCV Lab, KAIST
373-1 Guseong-dong Yuseong-gu Daejeon, Korea
ykjeong@rcv.kaist.ac.kr, iskweon@kaist.ac.kr

Abstract

In structure from motion, if two metric structures are given, the unknown scale between them can be resolved by constraining the rigidity of the metric space. There exist two well-known approaches. The first one is the pose estimation which aims to find the pose of a camera for known 3D points. The second one is the scale estimation whose goal is to resolve the scale after estimating the motion between cameras. Recently, the former way is preferred because of the vulnerability of the scale estimation, the weakness to the image noise. In this paper, we thus propose a robust method which can overcome the weakness of the scale estimation by considering the uncertainty of reconstructed 3D points. Additionally, the rotation matrix is directly corrected under the structural consistency constraint. To illustrate the performance of the method, we demonstrate some examples of large-scale reconstructions and compare the results.

1. Introduction

Structure From Motion(SFM) is the recovery of structures in a scene using camera motions, and it is currently one of the most active research areas in computer vision. Basically, SFM includes numerous estimations of the camera motion and the structure. Each estimation causes an error because of the uncertainty in measurements and the errors are accumulated over the entire estimation process. In order to minimize these errors, we can use Sparse Bundle Adjustment(SBA) for post-processing. However, the SBA has drawbacks such as the local minima problem and the requirement of heavy computations. Therefore, in the initial result, we need to obtain the finest estimation.

In this paper, SFM implies that the camera setup is weakly calibrated. Then, existing SFM can be classified into two groups: motion(relative pose) estimation based and pose estimation based methods. The motion estimation has a scale ambiguity. Shum et al.[7] performed motion estimation for SFM. They constructed a local structure with an image pair. Subsequently, they merged the local structures by resolving the unknown scales through the minimization of the sum of squared distances between corresponding 3D points. However, their method cannot be applied to a scene with a large depth range unless the accuracy of triangulation for the near and far points is fixed. On the other hand, Nister et al.[5] and Mourag et al.[4] used pose estimation in order to avoid resolving the scale ambiguity. However, pose estimation essentially needs accurate 3D points corresponding to the feature points. Therefore, they carefully reconstructed an initial structure using the information of tracked points between the first and the last frames before beginning the pose estimation. In [5], a re-triangulation scheme and a firewall for propagating errors were additionally used to reduce the error accumulation and Local Bundle Adjustment was used for the same purpose in [4].

In this study, we basically use a relative-motion-based approach, and the scale ambiguity is the same as that in [7]. In a scene that has plenty of features, the determination of the optimal geometric relation between two frames has already been extensively investigated [2, 8, 3]. Because the relative scale between two motions is the only cause of ambiguity, a slight mismatch between the scale factors of the two motions can cause severe distortions. To solve the ambiguity, we propose a novel scale estimation method involving outlier removal and optimization. A method is designed to measure the error in each stage to cancel or alleviate the negative effects of the uncertainties of the 3D points due to their large depth. The M-estimator approach can help deal with 3D points that have a potential uncertainty caused by their large depth. A reprojection process can cancel degradations resulting from triangulations. We also provide an additional correction. Through this procedure, an optimal scale that is sufficient to provide a good ini-

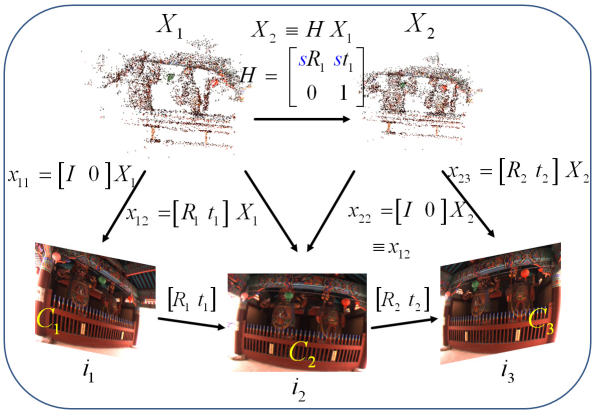

Figure 1. Geometric relations between an image triplet and structures from two pairs.

tial result for certain global optimizations can be efficiently obtained.

This paper is organized as follows. Section 2 explains details of the proposed relative scale estimation method, while section 3 introduces SFM implementation. The experimental results are presented in section 4, and section 5 provides the conclusions.

2. Relative Scale Estimation

Fig.1 shows the geometric relations between three image frames and we shall consider this case in detail. X_1, $X_2(\equiv HX_1)$, $\tilde{X}_1(= HX_1)$ and $\tilde{X}_2(= H^{-1}X_2)$ are 3D points and correspond to relative structures. We need to find the 3D homography H. If we find H, we can transform X_1 to X_2 and vice versa. This means that structures and motions can be expressed in the same coordinate system.

$$H_{met} = \begin{bmatrix} sR & st \\ 0 & 1 \end{bmatrix}, \ H_{aff} = \begin{bmatrix} D_a R & D_a t \\ 0 & 1 \end{bmatrix} \quad (1)$$

In the metric space, we can decompose H_{met} into a rotation matrix R and a translation vector t. In the above discussion, the words "scale" and "scale factor" are used with reference to s. In the affine space, we can find H_{aff} that can be decomposed into D_a, R and t. D_a is an upper triangular matrix and transforms H_{met} to H_{aff}. For an ideal case that noise-free measurements are available, s is constant and all the 3D points have the same scale, s with their corresponding points. Unfortunately, this is not true in practical cases.

In real experiments, the obtained image can suffer from many types of degradations. A noisy image usually causes one or more pixel errors within matched feature points and the errors affect the accuracy of the ray-triangulation. A small pixel error in the image could result in a very large depth uncertainty, especially in the case where the corresponding 3D point of a poorly localized feature is far from the camera. Similar ray directions or a relatively small baseline also causes the same problem. Finally, most of the points have inconsistent scale values. In order to remove the points contaminated by a large localization error or a mismatch, we use RANSAC for determining scale values and the M-estimator technique.

Let us assume that X_R is a real structure, and is the ground truth of the relative structures in Fig.1. All the structures are being viewed from three cameras. Furthermore, let X_i^j and x_{ik}^j be the coordinates of the jth point of X_i and its projection on the image plane of C_k (and therefore, of $\tilde{X}_i^j$ and $\tilde{x}_{ik}^j$), respectively. i_k^j is the observation which corresponds to x_{ik}^j. Then, we can define

$$s^j = \|\tilde{X}_2^j\| / \|X_1^j\| \quad (2)$$

$$e_s^j = w(s^j)[\|s^{ref} - s^j\|] \quad (3)$$

$$w(s^j) = \begin{cases} 1 & , Z(X_1^j) < Z_{med} \\ 1 + \frac{k_w * (Z(X_1^j) - Z_{med})}{(Z_{max} - Z_{min})} & , otherwise \end{cases} \quad (4)$$

where s^j is a sample scale, e_s^j is an error of the sample scale and s^{ref} is the reference scale for each iteration of RANSAC. $Z(X_i^j)$ is the depth of X_i^j, and Z_{med}, Z_{min} and Z_{max} are the median, minimum and maximum depth values of X_1. It is important to note that $w(s^j)$ acts as the $\rho()$ function in the M-estimator. $w(s^j)$ penalizes scale samples that come from a large-depth point, because of its larger uncertainty. Setting k_w to 1 or 2 is appropriate. Using them, we can apply the RANSAC algorithm in which the number of elements in the sample set of each iteration is only one, which is equal to the sample scale. Therefore, if we have n corresponding points, we test all the n sample scales s^j and finally obtain s^{init}, the best supportive scale. Subsequently, we perform an optimization process using another error measure,

$$e_r^j = \left(\frac{\|\tilde{x}_{21}^j - i_1^j\|}{Z(\tilde{X}_2^j)} + \frac{\|\tilde{x}_{13}^j - i_3^j\|}{Z(\tilde{X}_1^j)} \right) \quad (5)$$

This is a reprojection error inversely weighted by the depth of the corresponding 3D point. Actually, the reprojection cancels some amount of uncertainty. Therefore, the denominators in equation (5) could be removed when the scene has a small depth range. With densely distributed scale samples around s^{init}, we seek the optimal scale, s^{opt}, that minimizes the sum of e_r^j. Another way is to use the Levenberg-Marquardt algorithm which

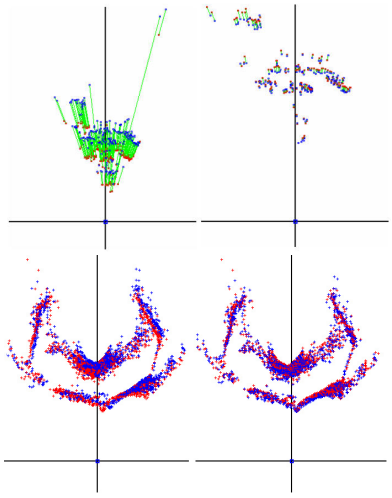

Figure 2. Matched structures. Top left, top right : Before and after the estimated s is used. Bottom left, bottom right : An example of motion correction.

is based on the abovementioned error measure. We have now estimated the optimal scale s_{opt} and thus, we can obtain a scale matching result like that shown at the top of Fig.2. The processing time is less than 0.05s. Thus, the optimal scale estimation does not affect the performance in terms of computation time.

In some cases, an erroneous motion could occur. To correct it, we employ an additional correction step. Using the matched scale (denoted by a "hat"), we should find H that satisfies $X_1 = H * \hat{\tilde{X}}_2$. If the motion is correct, H should equal to I and if the motion is incorrect, H would be a projective transformation. We approximately estimate an affine transformation, H_{aff}, instead of the projective one and can back-propagate the estimated R to R_1 or R_2 (refer to equation (1) and fig.1). In experiments, the motion that causes a large reprojection error is corrected, and the corrected motion is finally used only if it reduces the error. It is effective when the translation error is relatively small and the total amount of error is not very large (cf. Fig.2(bottom)). When the translation error is large, we should take D_a and t_a into account for correction and the process of correction is more complicated.

3. Structure From Motion Implementations

Our entire framework is iterative and can be divided into several parts.

Feature Extraction and Tracking The Harris corner detector[1, 5] is used to extract feature points and the KLT tracker is used for tracking[6].

Table 1. Pagoda results obtained by the proposed and the pose-based methods.

SFM type	Keyframes	Err. Init.	Err. SBA
proposed	48.25	0.6972	0.6218
pose	46.75	0.6537	0.6173

Table 2. Temple results obtained by the proposed and the pose-based methods.

SFM type	Keyframes	Err. Init.	Err. SBA
proposed	79.5	0.5438	0.4577
pose	81.25	0.5400	0.4648

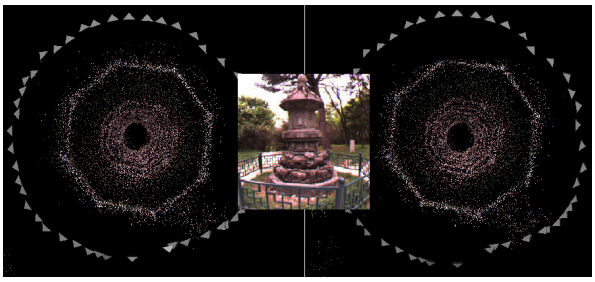

Figure 3. Pagoda result. Middle: A sample image. Left and Right: Reconstructions before and after the SBA.

Keyframe Selection In order to achieve a sufficiently wide baseline, we perform keyframe selection. The selection of the keyframe is made according to the fraction of remaining inliers and the number of tracks observed in the last three keyframes.[4].

Motion Estimation and Model Selection A fundamental matrix and a homography are used for the estimation of the relative motion with the robust methods specified in [2, 3, 8]. We detect the degeneracy of inliers using a single homography test in order to switch between motion models.

Scale Recovery and 3D Point management Our method for scale recovery has been explained in the previous section, and re-triangulation scheme is adopted to obtain the fairly accurate 3D points for each keyframe [5, 4].

4. Experimental Results

In this section, we discuss verification of the performance of the proposed method by presenting the reprojection errors, 3D reconstructions and camera motions

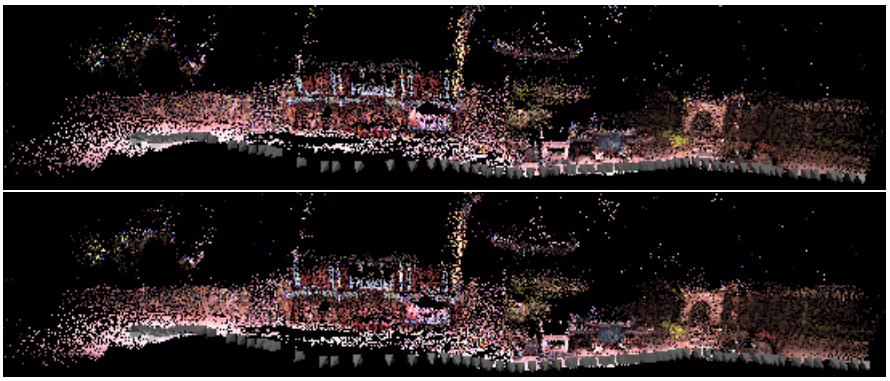

Figure 4. Temple result. Top and Bottom: Reconstructions before and after the SBA.

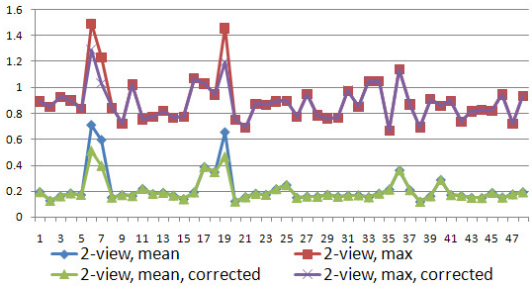

Figure 5. Graph of reprojection errors obtained by the third reconstruction of Pagoda data-set.

obtained by the proposed and the pose-based methods with and without the SBA. For each data-set and each method, SFM is performed four times. The errors and the number of keyframes are averaged. Table 1 compares the results of Pagoda data-set, and Table 4 compares those of Temple data-set. The pose-based method shows better results. However, the gaps between two methods are very narrow.

Fig.3 shows the third reconstruction of Pagoda data-set (49 keyframes among 1184 frames). The total mean reprojection error is reduced (0.665 to 0.612 pixel) by the SBA. The octagonal shapes of the fence and pagoda are nearly preserved even without the SBA. Fig.5 shows the mean and maximum reprojection errors for the keyframes, and also shows that the correction reduces the mean and maximum errors effectively.

Fig.4 shows the first one out of four reconstructions of Temple data-set. Two views are observed: the reconstructed structures with(bottom) and without(top) the SBA. The total mean reprojection error is reduced (0.547 to 0.462 pixel) by the SBA is applied. However, two images look very similar. Thus, the results show that the proposed method provides a near-optimal initial reconstruction.

5. Conclusion

A robust scheme for estimating relative scale is proposed, and a sequential SFM framework is also described. To obtain an accurate relative scale, we consider the similarity of structures and the minimization of the reprojection error to handle noisy observations. The structures are matched in 3D space, and a fine alignment that minimizes reprojection error is achieved. Thus we can observe that the proposed scale estimation is effective. Additionally, when the estimated motions are incorrect, we perform a rotation correction on the basis of the estimated 3D homography between the structures. Finally, experimental results show the improved accuracy and alleviation of the error accumulation problem and also indicate that the proposed method can be applied to the 3D reconstruction of large-scale scenes.

References

[1] C. Harris and M. Stephens. A combined corner and edge detector. *Proc. Fourth Alvey Vision Conference*, 1988.

[2] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003.

[3] Y. Ma, S. Soatto, J. Kosecka, and S. S. Sastry. *An Invitation to 3-D Vision*. Springer, 2004.

[4] E. Mouragnon, M. Lhuillier, M. Dhome, F. Dekeyser, and P. Sayd. Real time localization and 3d reconstruction. *CVPR*, 2006.

[5] D. Nister, O. Naroditsky, and J. Bergen. Visual odometry. *CVPR*, 2004.

[6] J. Shi and C. Tomasi. Good features to track. *CVPR*, 1994.

[7] H. Shum, Q. Ke, and Z. Zhang. Efficient bundle adjustment with virtual key frames: A hierarchical approach to multi-frame structure from motion. *CVPR*, 1999.

[8] Z. Zhang, R. Deriche, O. Faugeras, and Q. T. Luong. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Technical Report, Inria*, 1994.