IMTC 2005 - Instrumentation and Measurement
Technology Conference
Ottawa, Ontario, canada, May 17–19, 2005

Stereo Motion from Feature Matching and Tracking

Sébastien Gilbert[1], Robert Laganière[1], Gerhard Roth[2]

[1]School of Information Technology and Engineering
University of Ottawa
Ottawa, Ontario, CANADA, K1N 6N5
E–mail: laganier@site.uottawa.ca.

[2]Institute for Information Technology
National Research Council
Ottawa, Ontario, K1A 0R6, CANADA
E–mail: Gerhard.Roth@nrc-cnrc.gc.ca.

Abstract – *This paper presents a 3D pose estimation and reconstruction system based on a calibrated stereoscopic vision setup. The proposed approach consist in robustly tracking the movements of the cameras with respect to a rigid scene along a sequence. In addition, a novel correction scheme is proposed, that compensates for the accumulated error in the computed positions, exploiting the detection of loops in the movement. Experiments are presented to assess the accuracy of the resulting 3D measurements.*

Keywords – *Computer vision, stereoscopic vision, non-contact measurement, position measurement, camera calibration, 3D reconstruction.*

I. INTRODUCTION

The knowledge of the position of a camera with respect to a rigid reference frame has important applications for virtual or augmented reality systems, scene reconstruction, object modelling and robotics. In a video sequence in which a camera is moving inside a fixed environment, keeping track of the position of the camera with respect to its surroundings can be challenging. One possible solution consists in installing calibration targets, precisely registered with respect to a global reference frame. By making them visible inside the scene, it becomes possible to compute the camera position as the camera moves with respect to the global reference frame. In practice, this is not always possible. It is therefore desirable to develop a method to compute the camera motion in a rigid scene, when no calibration targets are present.

A pair of cameras whose intrinsic and extrinsic calibration parameters are known forms a calibrated stereoscopic vision setup. It allows 3D reconstruction of matched points [1]. If the feature points on a rigid object identified at capture N are tracked in both images at capture $N + 1$, the two clouds of 3D points can be registered [6], leading to the new position of the cameras. This is the idea that is exploited in this paper to robustly track the movements of the cameras with respect to a rigid scene along a sequence. In addition, a novel correction scheme is proposed, that compensates for the accumulated error in the computed positions, exploiting the detection of loops in the movement.

A. Literature Review

In [2], a binocular or trinocular stereoscopic setup is used and its path along a sequence is computed by using tridimensional reconstruction and registration. The robustness to matching and tracking errors is provided by two means. First, trilinear tensors are computed between image triplets. The features that support the trilinear tensors are known to be reliable. Second, a random sample consensus (RANSAC) [5] algorithm is applied to the 3D registration procedure. It is assumed that the disparities of the tracked feature points across the whole sequence is less than one third of the image size, thus constraining the camera movements.

In [3], the goal is to compute the registration of two consecutive scene captures along with the extrinsic calibration parameters of the stereo setup and the 3D location of a minimum of four matched and tracked feature points. The essential matrix of the stereo setup is calculated from the eight correspondences given by the four feature points in both captures, and nonlinear methods are used to enforce its constraints. It is decomposed to retrieve the extrinsic calibration parameters up to a scale factor of the translation vector. At this point, 3D reconstruction can be applied to the feature points, yielding two clouds of a minimum of four 3D points. The registration between the two captures can then be calculated. It differs from the proposed method in the fact that they do not compute the extrinsic calibration parameters of the stereo setup prior to the computation of the registration. As a consequence, the matching process cannot be guided by the epipolar constraint. No experimental results along a sequence were showed to display the accumulation of error.

In [4], stereoscopic vision and shape-from-motion are combined in an attempt to exploit the strengths of both approaches, i.e. accurate 3D reconstruction for stereo and easy feature tracking for visual motion. It computes 3D reconstruction of

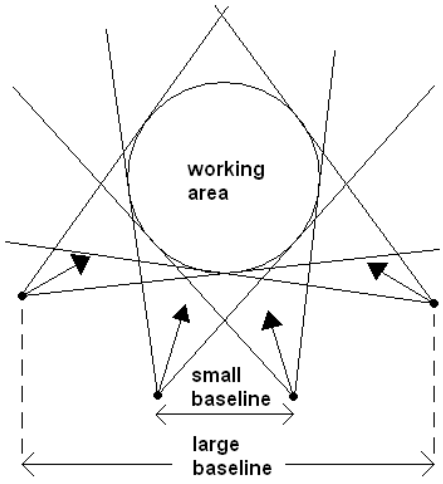


Fig. 1. Varying the baseline of the stereo setup

feature points and the camera motion in two separate steps. They limited their experimentations to short sequences where the viewpoints don't change dramatically from the first to the last capture.

II. PROPOSED APPROACH

A. Calibration

Calibration aims at computing the projection matrices of two cameras [1]. Let us assume that we have a set of n 3D points for which we know the global homogeneous coordinates $\vec{X}_i$. Each point, along with its corresponding image coordinates $\vec{u}_i$, allows to write:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix}_i = \lambda_i \begin{bmatrix} p_{00} & p_{01} & p_{02} & p_{03} \\ p_{10} & p_{11} & p_{12} & p_{13} \\ p_{20} & p_{21} & p_{22} & p_{23} \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}_i \quad (1)$$

Eliminating the λ_i and rearranging the expressions yields a pair of homogeneous linear equations in 12 unknowns, the entries of the projection matrix. Putting together the information of the n 3D points ($n \geq 6$) gives $2n$ homogeneous linear equations in 12 unknowns $p_{00}, p_{01}, ..., p_{23}$. This system can be solved up to a scale factor, through SVD. The quality of the computed projection matrix depends on the linearity of the camera model and the accuracy in the measured 3D location of the points.

Once the projection matrices are computed for both cameras, they can be decomposed to retrieve their intrinsic and extrinsic calibration parameters [1].

In order to determine the optimal angle between the Z-axes of the cameras, we performed an experiment in which we built

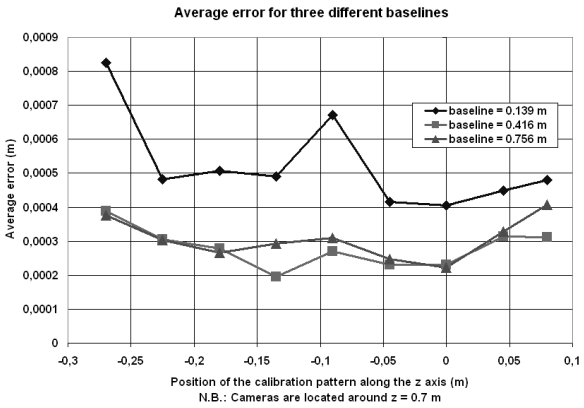


Fig. 2. Average reconstruction error, for three different baselines

three stereo setups with different baselines (0.139 m, 0.416 m and 0.756 m). The angles between the Z-axes of the two cameras were adjusted in such a way that a given working volume was preserved (see Figure 1), resulting in angles of 0.112 rad, 0.463 rad and 1.05 rad respectively. A calibration pattern was used, allowing easy detection of its feature points with subpixel resolution, through Hough transform. The position of the calibration pattern with respect to the table was measured with a ruler. This procedure provides the ground truth value of the feature points position, with an estimated accuracy of 0.3 mm.

Figure 2 shows the reconstruction error ($|\vec{x}_{calculated} - \vec{x}_{measured}|$) averaged over the 20 feature points of a calibration pattern as a function of the Z position of the calibration pattern, for three different baselines. It can be observed that the reconstruction error is higher for the stereo setup with the smallest baseline, as expected. No significant difference can be observed by comparing the results of the stereo setups with baselines of 0.416 m and 0.756 m. Since matching requires the cameras to be as parallel as possible, we can state that there is no need to increase the baseline of our stereo setup above 0.4 m, since it does not provide any improvement in reconstruction accuracy and it would make the matching process more difficult.

B. Matching and Tracking

It is assumed that the two cameras are sufficiently close and parallel to each other to allow matching through correlation. It is also assumed that the movement of the cameras is slow enough to allow feature tracking through correlation.

C. 3D Reconstruction

Let us assume the projection matrices P_1 and P_2 of the cameras are known, and we want to compute the 3D location $\vec{X}$ of a feature point whose image coordinates in the two images, $\vec{u}_1$ and $\vec{u}_2$, are known. The projection equations have the form $\vec{u}_j = \lambda_j P_j \vec{X}$, $(j = 1, 2)$. They can be manipulated to yield 4

linear equations in 3 unknowns, X, Y and Z:

$$\begin{bmatrix} (p_{00}-up_{20})_1 & (p_{01}-up_{21})_1 & (p_{02}-up_{22})_1 \\ (p_{10}-vp_{20})_1 & (p_{11}-vp_{21})_1 & (p_{12}-vp_{22})_1 \\ (p_{00}-up_{20})_2 & (p_{01}-up_{21})_2 & (p_{02}-up_{22})_2 \\ (p_{10}-vp_{20})_2 & (p_{11}-vp_{21})_2 & (p_{12}-vp_{22})_2 \end{bmatrix} \times \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} (up_{23}-p_{03})_1 \\ (vp_{23}-p_{13})_1 \\ (up_{23}-p_{03})_2 \\ (vp_{23}-p_{13})_2 \end{bmatrix} \quad (2)$$

This system can be solved through a least-square method.

D. Robust Registration

After having found matches and tracked the corresponding points in both sequences, two clouds of $3D$ points can be reconstructed. Based on the matches at instant N and their tracked correspondents at instant $N+1$, these two clouds of $3D$ points can be registered to find the rigid motion of the object [6] (or, alternatively, the rigid motion of the stereo setup, when the reference frame is attached to the object). Unfortunately, one cannot use the raw data, since the false matches and the tracking errors will corrupt the result. Instead, it is necessary to incorporate a random sample consensus (RANSAC) algorithm [5] that will filter out the bad pairs of 3D points [10].

One of the main problems associated with applying successive 3D registration procedures is the accumulation of error, due to the fact that every new position is computed from the previous. It is assumed that no special target points that could allow recalibration are available on the object. Instead, one must rely on the knowledge of the approximate camera positions to identify points of view that were previously captured (loop detection). This information will be used to correct for the drift, each time the cameras pass by a location where they have been before. The proposed scheme of automatic identification of loops in the movement and position correction by interpolation, as depicted in the next two subsections, is the main novelty of this paper.

E. Detection of Previously Viewed Locations

This procedure aims at identifying, in a sequence, camera positions that are close to their previous positions in an earlier image capture.

As pointed out in section B, we won't address the situations of wide-baseline matching or tracking. This means that, in order to be able to match images captured at non-consecutive instants, two conditions must be met:

1. The Z-axes of the two views must be nearly parallel;
2. The distance between the center of projection of the views must be sufficiently small.

The angle between the Z-axes of two views can be computed through a scalar product of unit vectors parallel to the Z-axes of the two cameras, as expressed in the world reference frame:

$$\cos(\theta) = \left(Q_{C_M/W} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \right) \cdot \left(Q_{C_N/W} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \right) \quad (3)$$

where $Q_{C_M/W}$ and $Q_{C_N/W}$ are the homogeneous transformation matrices linking a camera at capture M and at capture N with respect to the world reference frame (attached to the object).

The angle between the Z-axes of the left camera at capture M and N need not be the same as the equivalent for the right camera. In a sequence, the minimal angle (or distance) with respect to a given frame may not happen at the same frame for the left and the right camera. When trying to identify the best capture to be matched with an earlier capture, we must find a compromise between the two cameras.

Whenever a view is detected as having been previously captured, the drift of the later view can be compensated for. Of course, it is assumed that the earlier the view, the better the accuracy, since its location has been computed from a smaller number of cascaded transformations [10].

F. Interpolation of the Correction Matrix

After a loop has been detected between the earlier capture M and the later capture N, allowing for correction of the camera positions at view N, the intermediate views $M+1$, $M+2$, ..., $N-1$ can be corrected by interpolation. Under the assumptions of uniform error distribution along the sequence and small rotation amplitude both in the overall error and the individual registration matrices, it can be shown that the homogeneous transformation matrix of an intermediate view P ($M < P < N$) can be corrected by premultiplication of the correction matrix $Q_{correction,P}$:

$$Q_{correction,P} = Q_{correction,N}^{\frac{P-M}{N-M}} \quad (4)$$

III. RESULTS

Figure 3 shows the *Duck* sequence, augmented with its attached reference frame, after detection of a loop in the movement and correction of its projection matrices. The natural movement of the augmented reference frame confirms the validity of the corrected projection matrices.

In a second experiment, a sample of the images were supplied to a commercial bundle adjustment software, and the obtained camera positions were compared with those of the proposed method.

Figure 4 shows the disagreement (in the position and the orientation of one of the cameras) between bundle adjustment and the proposed method, without error correction. As expected, the magnitude of the disagreement increases with the number of registrations, as the proposed method accumulates error.

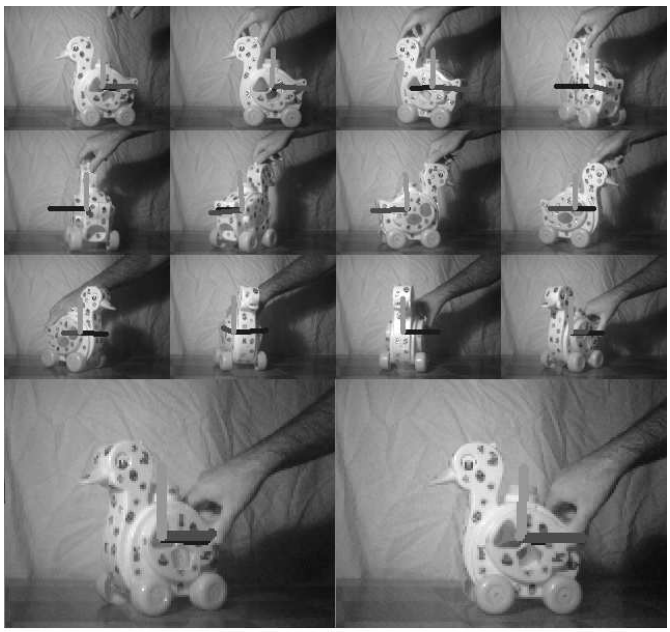
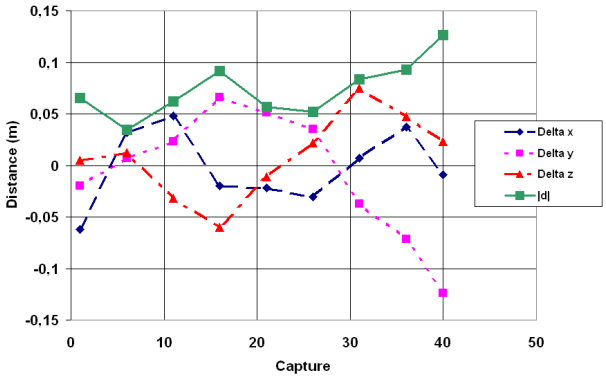
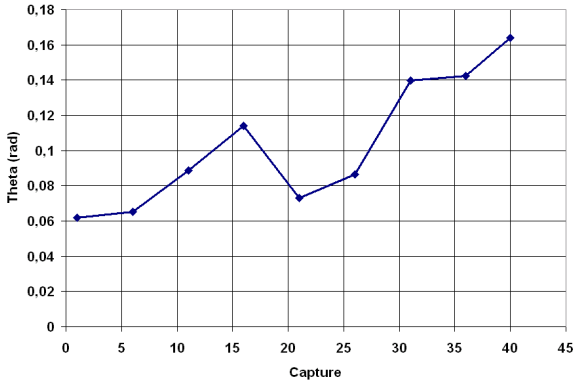


Fig. 3. Samples of the Duck sequence, as seen by the left camera, augmented with its attached reference frame

Figure 5 shows the disagreement between bundle adjustment and the proposed method after error correction through uniform distribution of the correction matrix. It can be seen that the disagreement magnitude does not increase with the number of registrations, indicating that the error correction provided an improvement in the projection matrices.



(a)



(b)

Fig. 4. (a) Position disagreement between bundle adjustment and the proposed method, without error correction (b) Z-axis orientation disagreement between bundle adjustment and the proposed method, without error correction

[1] Trucco E., Verri A. 1998. Introductory Techniques for 3-D Computer Vision, Prentice-Hall (eds)

[2] Roth Gerhard "Computing Camera Positions from a Multi-Camera Head", in *Proc. IEEE 3rd International Conference on 3-D Digital Imaging and Modeling*, 2001, pp. 135-142

[3] Zhang Zhengyou "Motion and Structure of Four Points from One Motion of a Stereo Rig with Unknown Extrinsic Parameters", in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 12, December 1995, pp. 1222-1227

[4] Ho Pui-Kuen, Chung Ronald "Stereo-Motion with Stereo and Motion in Complement", in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 2, February 2000, pp. 215-220

[5] Fischler Martin A., Bolles Robert C. "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography", in *Communications of the ACM*, vol. 24, no. 6, June 1981, pp 381-395

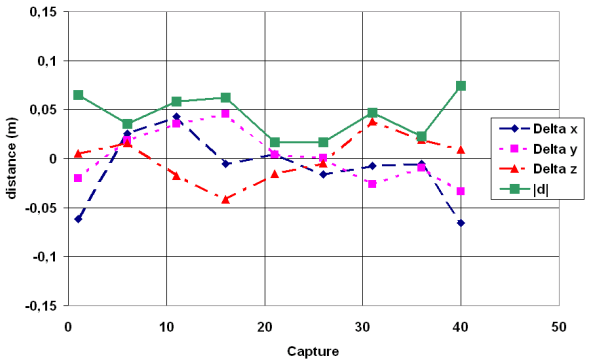
[6] Arun K.S., Huang T.S., Blostein S.D. "Least-Squares Fitting of Two 3-D Point Sets", in *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. PAMI-9, no. 5, Sept. 1987

[7] Laurentini Aldo "The Visual Hull Concept for Silhouette-Based Image Understanding", in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 2, February 1994, pp 150-162

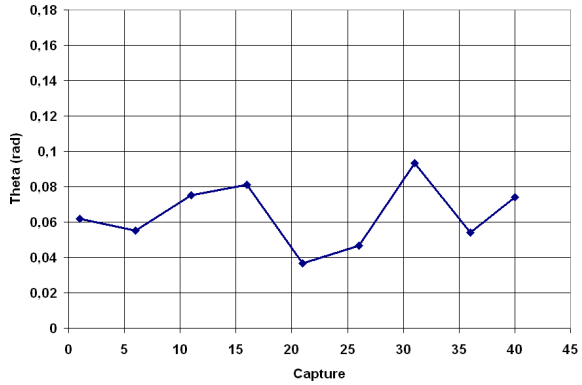
[8] Vincent Étienne, Laganière Robert "Matching Feature Points in Stereo Pairs: A Comparative Study of Some Matching Strategies", in *Machine Graphics and Vision*, vol. 10, no. 3, 2001, pp 237-259

[9] Lucas B., Kanade T. "An Iterative Image Registration Technique with an Application to Stereo Vision", in *Proc. of 7th International Joint Conference on Artificial Intelligence (IJCAI)*, 1981, pp. 674-679

[10] Gilbert Sébastien, Laganière Robert "Registration of a Moving Rigid Object Using a Stereoscopic Vision Setup", in *Proc. of the 3rd IEEE International Workshop on Haptic, Audio and Visual Environments and their Applications - HAVE 2004*, 2004, pp. 171-175



(a)



(b)

Fig. 5. (a) Position disagreement between bundle adjustment and the proposed method, after error correction (b) Z-axis orientation disagreement between bundle adjustment and the proposed method, after error correction