CAMERA MOTION ESTIMATION FOR OBJECT-BASED VIDEO CODING

YUNHAI LIU
Institute of Information and Communication Engineering, Zhejiang University, China
E-mail: liuyh@zju.edu.cn

THORSTEN THORMAEHLEN AND HELLWARD BROSZIO
Information Technology Laboratory, University of Hannover, Germany
E-mail: {thormae, broszio}@tnt.uni-hannover.de

This paper deals with motion estimation of an airborne camera and generation of 3D landscape models from image sequences for object-based video coding. While most camera motion estimators use a random sampling outlier elimination, the presented new approach uses a priori knowledge to simplify the complexity of the outlier elimination and to achieve real time coding. Test results with synthetic sequences show an estimation error of maximal one percent for the motion parameter which leads to a good quality for the generated 3D object models.

1 Introduction

Traditional block-based motion compensated hybrid coding is widely used and has been adopted by international image coding standards due to excellent coding efficency. With the increasing performance of modern computer systems new, more complex, coding techniques become practicable, such as object-based, knowledge-based or semantic coding, which promise higher coding efficiency and can be combined together with block-based coding in a layered coding system[1]. For these new techniques a scene is represented by object models, which are described by their motion, shape and texture parameters. It is still a key research topic to estimate these parameters automatically. In MPEG-4 object-based coding is adopted, but no standard tools are defined to generate 2D or 3D models of the objects.

In practice mainly monocular image sequences are coded. In this case the relative translation (or translation and rotation) between the camera and the objects can be used to estimate a 3D object model. In this paper the image sequence is captured by a camera in an airplane. For simplification a stationary landscape and a camera motion model of pure translation can be assumed because the airplane's motion can be approximated as translation during a small time period.

The next Section describes the estimation of camera translation. Section 3 explains the generation of the landscape models. In the last Section test results are shown and a conclusion is given.

2 Estimation of camera translation

To achieve robust and exact camera motion estimation the presented algorithm includes four steps. First feature points are detected with subpel accuracy in each image. In the second step correspondences of the feature points of two successive images are assigned by measuring normalized cross-correlation between windows surrounding the feature points. The correspondences are established for those feature points which have the highest cross-correlation. Due to misalignments, usually some of the correspondences are incorrect, called outliers. In the third step an outlier elimination by case deletion is applied. In the final step motion parameters are estimated by linear least square optimization using only the selected good correspondences, called inliers. In the following the new outlier elimination algorithm is described. Detailed explanations of the other steps can be found in reference 3.

Outlier elimination is an necessary step to achieve robustness in camera motion estimation. Recommended[2] outlier elimination methods are Random Sample Consensus (RANSAC) or Least Median of Square (LMS). These techniques can cope with large numbers of outliers but the computational effort is high. The presented simplified outlier elimination is useful if the amount of outliers is small.

If focal length is assumed to 1 and with pure translational motion the simplified epipolar constraint is given by:

$$\begin{pmatrix} y_i' - y_i \\ x_i - x_i' \\ x_i' y_i - x_i y_i' \end{pmatrix} \cdot \begin{pmatrix} t_x \\ t_y \\ t_z \end{pmatrix} = \mathbf{V}_i \cdot \mathbf{t} = 0 \quad \forall\ i = 1...M \tag{1}$$

where $\mathbf{p}_i = (x_i, y_i)^\top$ is the image coordinate of a feature point in the first image plane and $\mathbf{p}_i' = (x_i', y_i')^\top$ the corresponding feature point in the second image plane, see Fig. 1. $\mathbf{t} = (t_x, t_y, t_z)^\top$ is the translation vector between the two views and M is the number of correspondences. If all vectors $\mathbf{V}_i$ are normalized to unit vectors $\mathbf{v}_i$, they are distributed on a sphere. As expressed with Eq. 1 the translation vector $\mathbf{t}$ should be orthogonal to $\mathbf{v}_i$ for all inliers. This means all inliers $\mathbf{v}_i$ should be distributed on one inlier plane. Vectors $\mathbf{v}_i$ which lie outside this plane are considered as outliers, see Fig. 2. The quality of a single vector $\mathbf{v}_i$ is measured by the sum of the vector products

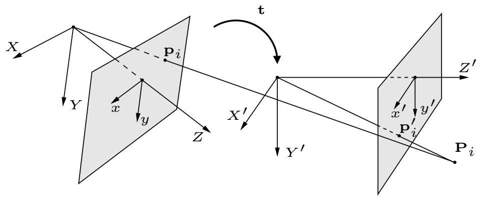

Figure 1. The projection of a 3D point $\mathbf{P}_i$ determines the corresponding feature points $\mathbf{p}_i$ and $\mathbf{p}_i'$ in the image plane of the first and second camera.

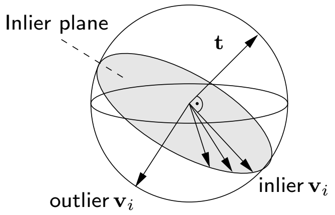

Figure 2. Outlier elimination which analyses the position of normalized vectors $\mathbf{v}_i$ on a sphere. Vectors $\mathbf{v}_i$ in the inlier plane are inlier correspondences.

$$\text{quality}(\mathbf{v}_i) = \sum_{j=1}^{M} \mathbf{v}_i \cdot \mathbf{v}_j \tag{2}$$

because the product of two normalized vector is the cosine of the angle between the vectors. If all inlier vectors $\mathbf{v}_i$ are distributed on the inlier plane equally, this criteria fails. But experiments with sequences from an airborne camera have shown that the distribution is concentrated. A vector $\mathbf{v}_i$ with high quality has high probability to be an inlier but these candidates are only a subset of all inliers. All vectors $\mathbf{v}_i$ can be ordered according to the calculated quality. The M ordered vectors are classified into two sets, the 2 vectors with the lowest quality and the remaining vectors. The translation vector $\mathbf{t}$ for both sets is estimated by

$$\sum_{\text{set}} (\mathbf{v}_i \cdot \mathbf{t})^2 \rightarrow \min \tag{3}$$

If the magnitude of the distance of the estimated two translation vectors exceeds a given threshold, the 2 vectors $\mathbf{v}_i$ with the lowest quality are deleted. The ordering with Eq. 2 and comparison with Eq. 3 are repeated until the difference of the estimated two translation vectors is below the given threshold. The remaining vectors $\mathbf{v}_i$ are considered as inliers. Thus the random selection of subsets in RANSAC is substituted by an ordered selection using quality criterion of Eq. 2.

3 Generation of landscape models

Known camera motion enables the calculation of 3D point coordinates. The triangulation of two lines of sight from two different camera positions gives the 3D coordinate for each correspondence. Due to erroneous detection of feature points, the lines of sight do not intersect in most cases (Fig. 3). Therefore, two 3D points $\tilde{\mathbf{P}}_i$ and $\tilde{\mathbf{P}}'_i$ can be determined for each feature point separately. The 3D points are located where the lines of sight have their smallest distance. The arithmetic mean of $\tilde{\mathbf{P}}_i$ and $\tilde{\mathbf{P}}'_i$ gives the final 3D coordinate $\mathbf{P}_i$. Depending on the 2D triangle meshes in the image planes the 3D points are connected to a 3D triangle mesh (Fig. 4).

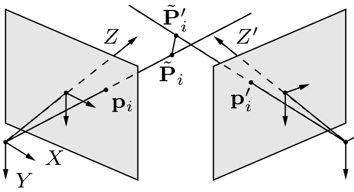

Figure 3. Calculating 3D point $\mathbf{P}_i$ by triangulation of lines of sight.

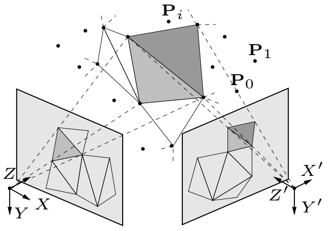

Figure 4. Connecting 3D points $\mathbf{P}_i$ advised by the 2D triangle meshes in the image plane.

4 Test results and conclusion

Two different test sequences have been used for experimental investigations. The first sequence shows a synthetic scene with known camera motion, which is used to measure the motion estimation error. Referencing Fig. 6 the motion estimation error for the synthetic sequence has a maximal value of one percent. Fig. 7 illustrates the generation of 3D object models by displaying the reconstructed depths as a gray scale image. High intensity values indicate a small distance of the objects to the camera. The second sequence (Fig. 8) is a real scene captured by an airborne camera. Fig. 9 shows the motion estimation result.

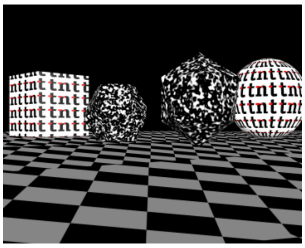

Figure 5. Synthetic scene with virtual objects.

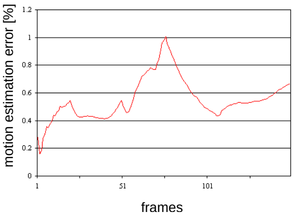

Figure 6. Motion estimation error in percent.

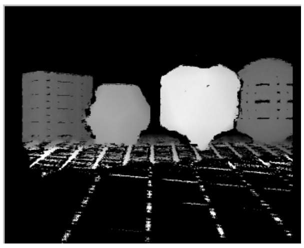

Figure 7. Depth map of generated object models.

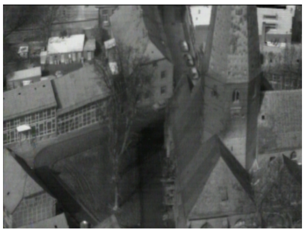

Figure 8. Scene captured by an airborne camera.

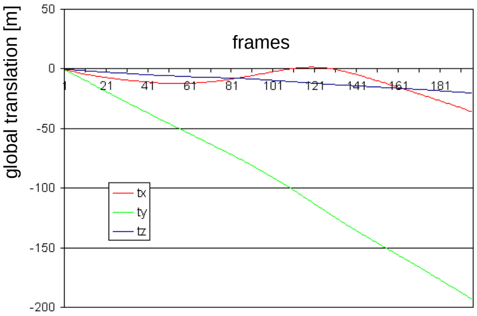

Figure 9. Estimated global (accumulated) camera translation.

This paper presents an algorithm for robust motion estimation of an airborne camera and generation of 3D models from an image sequence. In future work the generated object models have to be checked for use in an object-based video coder.

References

1. J. Ostermann, M. Kampmann, *Source Models for Content-Based Video Coding*, International Conference on Image Processing 2000 (ICIP-2000), Vancouver, Canada, 10-13 September 2000 (2000)
2. R. Hartley, A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press (2000)
3. T. Thormaehlen, H. Broszio, P. N. Meier, *Three-Dimensional Endoscopy*, Falk Symposium No. 124, Medical Imaging in Gastroenterology and Hepatology, Hannover, 28-29 September 2001 (2002)