CAMERA CALIBRATION OF LONG IMAGE SEQUENCES WITH THE PRESENCE OF OCCLUSIONS

Miguel Sainz
Image Based Modeling and Rendering Lab
Dept. of Electrical and Computer Science
University of California, Irvine, USA
msainz@ece.uci.edu

Antonio Susin
Dynamic Simulation Lab
Dept. Matematica Aplicada 1
Universitat Politecnica de Catalunya, SPAIN
toni.susin@upc.es

Nader Bagherzadeh
Image Based Modeling and Rendering Lab.
Dept. of Electrical and Computer Science
University of California, Irvine, USA
nader@uci.edu

ABSTRACT

Camera calibration is a critical problem in applications such as augmented reality and image based model reconstruction. When constructing a 3D model of an object from an uncalibrated video sequence, large amounts of frames and self occlusions of parts of the object are common and difficult problems. In this paper we present a fast and robust algorithm that uses a divide and conquer strategy to split the video sequence into sub-sequences containing only the most relevant frames. Then a robust stratified linear based algorithm is able to calibrate each of the subsequences to a metric structure and finally the subsequences are merged together and a final non-linear optimization refines the solution. Examples of real data reconstructions are presented.

INTRODUCTION

In recent years Image Based Modeling and Rendering (IBMR) techniques have demonstrated the advantages of using real image data to greatly improve the rendering quality in virtual environments. New rendering algorithms have been presented that reach a photorealistic quality at interactive speeds when rendering 3D models by using images of real objects and some additional shape information (i.e. a geometric proxy). While these methods have emphasized the rendering speed and quality, they generally require extensive preprocessing in order to obtain accurately calibrated images and geometric proxies of the target objects. Moreover, most of these algorithms require user interaction for the camera calibration and image registration part or need the use of expensive equipment such as calibrated gantries and 3D scanners.

In this paper we present a method to calibrate and extract a set of key-frames from a video sequence that contain enough three dimensional information to be used as the reference views for an image based reconstruction algorithm. More specifically, the goal is to recover the 3D geometry of a scene from the 2D projections obtained from the digital images of multiple reference views, taking into account the motion of the camera. Moreover, since to obtain an robust reconstruction the image sequence must show the object from different perspectives, self-occlusions are constantly present increasing the difficulty of the problem.

Inspired in [1] and [2], we present a different novel approach based on a divide and conquer strategy to fully calibrate a long sequence of images with a high degree of feature occlusions such as video sequences of objects in rotating platforms. The complete sequence is automatically divided into subsequences and, in each of them, a set of key-frames is selected and calibrated using an improved version of the algorithm presented in [5], recovering both camera parameters and structure of the scene. When the different subsequences have been successfully calibrated a merging process groups them into a single set of cameras and reconstructed features of the scene. A final non-linear optimization is performed in order to reduce the overall reprojection error.

One advantage of the presented approach is that it allows to recover an Euclidean reconstruction of the scene without any initial solution or prior information, which is one of the drawbacks of most of the existing methods. Another important feature is that the entire calibration process is based on solving *linear systems* using the SVD decomposition algorithm. The knowledge of the geometric meaning and rank properties of the different transformations represented by the matrices of the process allows to enforce a valid Euclidean reconstruction. The proposed solution is designed to be versatile in respect to the input data allowing the use of (1) automatically tracked video sequences, (2) manually tracked sequences which usually contain less frames or (3) a set of still images with features and correspondences manually selected. Here on, we assume that the input data is given as a sequence of images and a list of features in each image and the correspondences with the rest of the frames.

SEQUENCE FRAGMENTATION

The fragmentation algorithm starts with the set of features of frame $i=0$ and keeps track of them in the subsequent

frames until one of the following is satisfied:

- A minimum of key frames has been selected.
- The end of the sequence is reached.
- More than a user selected percentage of the original features are lost.

When this occurs, a new subsequence is created as the set of key-frames starting from the first frame to the last keyframe before the ending condition has been triggered.

At the same time, each frame is tested against the last found key-frame for a planar homography fit in order to determine how much three dimensional variation is present. Typically the homography error is small when there is little variation in the camera positions between the two frames. We use a RANSAC based approach to determine the percentage of inliers of the 2D homography between the two frames [3]. If this value is below a user selected threshold, it means that there exists some significant camera motion not modeled by the 2D homography, and the frame is marked as a key-frame. Otherwise, the frame is discarded and we proceed to the next frame.

To guarantee connectivity between the different subsequences so they can be merged, the last key-frame of a subsequence is the first key-frame of the following one.

FRAGMENT CALIBRATION

Once a subsequence has been determined, we proceed to perform a complete metric calibration by extracting the measurements that appear in all the frames of the fragment into a measurement matrix *W*. The presented calibration solution is a stratified reconstruction based on linear factorization algorithms with a non linear optimization process to reduce the overall reprojection error. Moreover, a robust statistical 3D analysis based on RANSAC [3] is used to improve the quality of the reconstructions.

Inlier determination

Due to limitations and errors of the tracking algorithms, not all the features contained in the initial full measurement matrix are suitable to be used for the reconstruction. Therefore an initial filtering based on a RANSAC-type random sampling approach is needed in order to extract the set of inlying measures.

Using this method we randomly select sets of four features, which is the minimum amount required to obtain a projective reconstruction. A solution for the projection matrices *P* is calculated for each set and the rest of the points are obtained as a least square solution for *P* and the 2D measurements. Then all the reconstructed points can be classified into inliers or outliers depending on the reprojection error. The solution that presents the largest number of inliers is kept.

However, the proposed projective reconstruction algorithm is a robust but slow iterative approach and it is not suitable to be used multiple times as required by the RANSAC filtering. To accelerate the selection of the best set of inliers an closed-form affine reconstruction [3],[4] is performed based on the randomly selected sets.

Once an initial set of inliers has been determined, an improved projective reconstruction is computed by iteratively reevaluating the set of inliers and calculating a new projective solution with the new set of inliers until the number of inlying measures remains constant. Usually this robust estimate converges in less that ten iterations, allowing the use of the most costly projective reconstruction algorithm.

Projective Reconstruction

The projective factorization method is a generalization of the factorization method which was first developed in [4], for the orthographic and the paraperspective projection models respectively. It provides a more general framework to recover shape and motion from multiple view images.

Let $X_j = (X_j, Y_j, Z_j, 1)^T$, $j = 1, \ldots, n$, be the unknown homogeneous 3D point vectors, P_i, $i = 1, \ldots, m$ the unknown 3 *x* 4 image projections, and $x_{ij} = (u_{ij}, v_{ij}, 1)$ the measured homogeneous image point vectors. We call **projective depths** the non-zero scale factors λ_{ij} relating the world points and its projections

$$W = \begin{bmatrix} \lambda_{11}x_{11} & \ldots & \lambda_{1n}x_{1n} \\ \ldots & \ldots & \ldots \\ \lambda_{m1}x_{m1} & \ldots & \lambda_{mn}x_{mn} \end{bmatrix} = \begin{bmatrix} P_1 \\ \ldots \\ P_m \end{bmatrix} \begin{bmatrix} X_1 & \ldots & X_n \end{bmatrix} \qquad \text{Eq (1)}$$

The matrix *W* has to be a rank-4 matrix if it is the matrix associated to a projection of a set of real points. Consequently, for points in general positions, a rank-4 factorization of the scaled matrix produces a projective reconstruction of the points.

An iterative algorithm presented in [5] is used to perform a rank 4 decomposition while determining the best set of projective depths for that factorization.

Metric Upgrade

We want to update our projective scene reconstruction to a metric one. Essentially the autocalibration process can be summarized into the problem of computing a *projective distortion matrix* (PDM), in other words, an homography *H* such that, $P_i^M = P_i H$ and $X_j^M = H^{-1}X_j$ are the metric reconstruction of the scene. The metric camera matrices can be decomposed in terms of the internal and external parameters as

$$P_i^M = K^i[R^i|t^i], \; i = 1, \ldots m \qquad \text{Eq (2)}$$

where the *3 x 3* symmetric matrix K^i is the internal camera parameters matrix, R^i is the euclidean rotation matrix and t^i is the translation vector.

To derive the auto-calibration equations [3], [7], let us choose the world coordinate frame aligned with the first camera. If we write the plane at infinity as an usual plane, $\pi_\infty = (p^T,1)^T$, then

$$H = \begin{bmatrix} K^1 & 0 \\ -p^T K^1 & 1 \end{bmatrix} \quad \text{Eq (3)}$$

To derive the auto-calibration equations, first we express the rest of the projective cameras distinguishing between the first three and the last columns, $P^i = [A^i | a^i]$, and from equations (1) and (2) we obtain (see [3])

$$K^i K^{iT} = (A^i - a^i p^T) K^1 K^{1T} (A^i - a^i p^T), i = 2 \ldots m \quad \text{Eq (4)}$$

The symmetric matrix $K^i K^{iT}$ is known to be the dual image of the absolute conic, ω^{*i}, which is related with the *absolute dual quadric*, Q_∞^*, by

$$\omega^{*i} = P^i Q_\infty^* P^{iT} \quad \text{Eq (5)}$$

As we are using the first camera frame as the origin, we can write

$$Q_\infty^* = \begin{bmatrix} K^1 K^{1T} & -K^1 K^{1T} p \\ -p^T K^1 K^{1T} & p^T K^1 K^{1T} p \end{bmatrix} = \begin{bmatrix} \omega^{*1} & -\omega^{*1} p \\ -p^T \omega^{*1} & p^T \omega^{*1} p \end{bmatrix} \quad \text{Eq (6)}$$

A linear system can be obtained from (6) and (5) assuming some knowledge on the camera internal parameters. For instance, if the principal point is at the origin then $\omega_{13}^{*i} = \omega_{23}^{*i} = 0$. From (5) two linear equations can be derived for the entries of Q_∞^*. When, in addition, zero skew is assumed another linear relation is added $\omega_{12}^{*i} = 0$. Finally, if an equal aspect ratio is assumed a last linear equation $\omega_{11}^{*i} = \omega_{22}^{*i}$ can be added. This means that 4 linear equations can be derived from every frame when the previous assumptions are considered. When looking at Q_∞^* this assumptions gives (Q_∞^* is symmetrical)

$$Q_\infty^* = \begin{bmatrix} q_{11} & 0 & 0 & q_{14} \\ 0 & q_{11} & 0 & q_{24} \\ 0 & 0 & 1 & q_{34} \\ q_{14} & q_{24} & q_{34} & q_{44} \end{bmatrix} = HH^T \quad \text{Eq (7)}$$

The autocalibration equations (5) become an overdetermined linear system of *4 x m* equations and only five unknowns. A supplementary non linear condition $det(Q_\infty^*) = 0$ must be added to the least square solution.

If we express the solution of the overdetermined system as a linear combination of the solution vector and a constant λ by the eigenvector with the lowest eigenvalue, a third degree polynomial in terms of λ can be obtained to enforce the supplementary condition. Then, using SVD to obtain H and back-substituting in the equations, a final metric reconstruction is computed under the assumption of known principal points and skew values.

Optimization

The solution presented above is a closed form least squared constrained approximation of the structure from motion problem. A final non-linear optimization process is required in order to reduce the reprojection error accounting for all the nonlinearities not recovered in the metric solution. Moreover, if a more sophisticated camera internal parameter description is required, it can be incorporated into the optimization process as well.

Additionally, since the cameras have been recovered, several of the features not included in the reconstruction, because they are not present in all the frames, can be recovered by least squared approximation, increasing the number of reconstructed points.

A bundle adjustment is then computed using the standard sparse Levenberg-Marquardt algorithm described in the literature (e.g. in [3] or [8]), using the reconstruction as the initial solution.

SEQUENCE MERGING

The merging of two subsequences is performed in metric space by determining the set of common points, between the last frame of one sub-sequence and the first of the next one, that by construction correspond to the same camera and measurements.

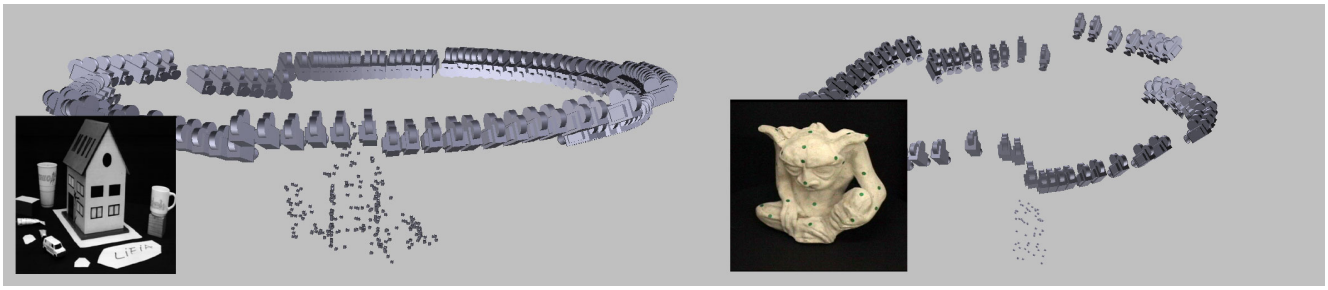
Since both subsequences have been calibrated, the set of reconstructed common points must be similar up to a scaling and a translation assuming both frames aligned with the world reference. An overdetermined linear equation system can be built and solved relating the reconstructed features. Then, the transformations are applied and the sub-sequences are merged.

Due to noise related errors during calibration or because the set of features used in each of the subsequences is different, small disparities in the focal length between the two overlapped frames can appear. This does not affect the overall reprojection error since it is a reconstruction fully compatible with the input data. In the case that a correction was necessary, a recalibration of one of the subsequences should be performed using the desired focal length as a constraint. However for the purpose of the presented work this is not required since the 3D reconstruction of the features is accurate and the reprojection error is small.

Once all the subsequences have been merged, a final Levenberg-Marquardt adjustment is computed over the complete sequence to reduce the overall error.

RESULTS

Several validation and verification tests have been performed with real imagery. The two datasets presented in this paper consist of video sequences of a circular camera motion around a set of objects. All tests were performed on a PC system with a 2.2Ghz P4, 1Gb of RAM running MS



	# of Frames	# of used frames	# of features	Total time(sec)	Selection time (sec.)	Calibration time (sec.)	Bundle time (sec.)	# sub sequences	Mean error (pixels)	Error Var.
Monster	423	68	52	130.1	5.1	2.8	122.1	7	0.64	0.58
House	114	114	936	371.5	23.9	127.7	219.7	12	0.81	0.96

Table 1: Statistics of the datasets

Windows XP.

The first dataset, *monster*, consists of 423 frames and was captured using an on-the-shelf digital video camcorder at 720x480 pixels of resolution. The object is a 20cm tall concrete statue on top of a turning table. The camera was set to autofocus and the object was manually rotated. Several markers where located on the surface of the object and a semi-automatic tracking tool was used to generate the measurement data.

The second dataset, *house*, is the well know sequence from the project MOVI and consists of 114 frames. The measurement data was obtained using the KLT tracker [6] set to track 500 features per frame replacing the ones that are lost due to noise or occlusions.

Some numerical results are presented in Table 1 showing the reconstruction speed and how much time each of the parts of the method use. The most expensive is the final bundle adjustment because it uses all the selected frames and all available features. The final reprojection mean errors are very low (below 1 pixel) showing the accuracy of the calibration method.

On top of the page, two figures show the 3D reconstruction of the features and the cameras as well as one of the frames of each sequence.

CONCLUSIONS

In this paper we have presented a novel divide-and-conquer approach to the problem of calibrating a moving camera during long sequences with feature occlusions, with the purpose of identifying and recover 3D information to be used in an image based modeling tool.

An automatic key-frame selection and sub-sequence construction strategy is used to select and group those frames from the sequence that most likely contain significant three dimensional information.

A fast and robust calibration tool is used to recover the structure and motion from each of the subsequences. This algorithm is mainly linear in the unknowns and does not require any prior knowledge of the scene or camera parameters.

The extensive use of RANSAC-based techniques, both in 2D and 3D space, provide extra robustness to the reconstruction algorithms. Moreover the proposed method is computationally fast in standard PC computers, making it a very attractive solution.

Acknowledgments. This research was supported by the National Science Foundation under contract CCR-0083080 and by the Comissio Interdepartamental de Recerca i Innovacio Tecnologica, Gaspar de Portola grant C02-03.

REFERENCES

[1] Fitzgibbon A. and Zisserman A. *Automatic camera recovery for closed or open image sequences*. In Proc. European Conference on Computer Vision, pages 311ñ326. Springer-Verlag, June 1998.

[2] Gibson S., Cook J., Howard T.L.J., Hubbold R.J., and Oram D. *Accurate Camera Calibration for Off-line, Video-Based Augmented Reality.* IEEE and ACM International Symposium on Mixed and Augmented Reality, Darmstadt, Germany, September 2002.

[3] Hartley R., Zisserman A., Multiple View Geometry in Computer Vision. Cambridge Univ. Press, 2000.

[4] Poelman C.J. and Kanade T., A paraperspective factorization method for shape and motion recovery *Technical Report* CMU-CS 93-219, Carnegie Mellon University, December 1993.

[5] Sainz M., Bagherzadeh N. and Susin A., *Recovering 3D Metric Structure and Motion from Multiple Uncalibrated Cameras*. In IEEE Proc. International Conference on Information Technology: Coding and Computing, pp 268-273, 2002.

[6] Tomasi C. and Kanade T. *Detection and tracking of point features*. Technical Report CMU-CS-91-132, Carnegie Mellon University, April 1991.

[7] Triggs B., Autocalibration and the Absolute Quadric *IEEE CVPR96*, 845-851, 1996.

[8] Triggs B., McLauchlan P.F., Hartley R.I., and Fitzgibbon A.W. Bundle adjustment - a modern synthesis. In B. Triggs, A. Zisserman, and R. Szeliski (eds.), Vision Algorithms: Theory and Practice, pp. 298--472. Springer-Verlag, 2000.