PRIORITIZED SEQUENTIAL 3D RECONSTRUCTION IN VIDEO SEQUENCES WITH MULTIPLE MOTIONS*

Evren İmre, Sebastian Knorr†, A. Aydın Alatan, and Thomas Sikora†

Dept. of Electrical & Electronics Eng.,
M.E.T.U.
Balgat, 06531 Ankara, Turkey
E-mail:{eimre@, alatan@eee.}metu.edu.tr

† Communication Systems Group,
Technische Universität Berlin
Einsteinufer 17, Berlin, Germany
E-mail: {knorr, sikora}@nue.tu-berlin.de

ABSTRACT

In this study, an algorithm is proposed to solve the multi-frame structure from motion (MFSfM) problem for monocular video sequences in dynamic scenes. The algorithm uses the epipolar criterion to segment the features belonging to independently moving objects. Once the features are segmented, corresponding objects are reconstructed individually by using a sequential algorithm, which is also capable of prioritizing the frame pairs with respect to their reliability and information content, thus achieving a fast and accurate reconstruction through efficient processing of the available data. A tracker is utilized to increase the baseline distance between views and to improve the F-matrix estimation, which is beneficial to both the segmentation and the 3D structure estimation processes. The experimental results demonstrate that our approach has the potential to effectively deal with the multi-body MFSfM problem in a generic video sequence.

Keywords: Machine vision, image analysis

1. INTRODUCTION

When the aim is extraction of the 3D information from a typical mono video sequence, the multi-frame structure from motion (MFSfM) problem should be solved in order to exploit the information redundancy in the video sequence to obtain a robust estimate. Moreover, in the presence of multiple moving objects, the problem definition should be extended to include the estimation of the structure and the motion of the independently moving objects (IMOs), as well as the background structure and the motion of the camera.

The literature on the multi-body MFSfM problem is shaped by the observation that, when the feature set is segmented into partitions corresponding to the background and the individual objects, the problem can be decomposed into several static MFSfM problems, one for the background and each of the IMOs. Hence, the segmentation and the reconstruction are the subproblems in this aspect.

The solution approaches for the segmentation part of the problem can be classified into three categories. Optical flow based methods assume that the scene is composed of planes at various depths, and use a simple clustering to achieve the desired segmentation [1]. Another set of solutions utilize the eigen decomposition of the *affinity matrix*, a structure which contains the similarity information among the features [11]. Finally, geometric methods exploit the constraints imposed by the epipolar geometry and the rigid body motion assumption. The most common constraint is the fundamental matrix [1].

For the reconstruction part, the basic approaches are the batch and the sequential methods. The best known example of the former is the popular *factorization method* [10], with variants covering many different camera models [3]. In the sequential methods, the problem is either cast into the framework of state estimation in dynamic systems [8], or the framework of inverse-MSE filtering to estimate an unknown constant vector (structure) [6].

While uncommon, there also exist techniques to solve the MFSfM problem simultaneously for all bodies involved, employing the multi-body extension of the factorization method [1], or the particle filter [7].

In this paper, both the segmentation and the reconstruction aspects of the multi-body MFSfM problem in video sequences are studied. The organization of the paper is as follows: In the next section, the proposed solution is outlined. In Section 3 and 4, the segmentation and the reconstruction stages are described, respectively. The experimental results are presented in Section 5. Finally, in Section 6, the paper is concluded by a discussion of the results and the future work.

2. PRIORITIZED RECONSTRUCTION FOR MULTI-BODY MFSFM

The proposed algorithm attempts to deal with two problems: Decomposition of the dynamic scene structure and efficient processing of the huge amount of data a video sequence presents to achieve a good reconstruction.

In order to solve the first problem, the geometric segmentation approach is employed, as it offers a robust method that easily incorporates more information with fewer assumptions,

* This work is funded by EC IST 6th Framework 3DTV NoE

than its competitors in the literature. The segmentation is performed, not at the feature, but at the trajectory level, as video sequences allow the utilization of the Kanade-Lucas tracker to build trajectories accurately. Once the segmentation is complete, each partition of points can be constructed by the proposed reconstruction algorithm.

As for the reconstruction part of the problem, the batch methods can easily handle such vast amount of data. However, they lack a significant advantage of the sequential methods: Intermediate results obtained from the already processed frames can be incorporated into the processing of the remaining ones, to improve the final result. Obviously, this approach renders the result dependent on the processing order of the frames, which brings up the issue of how to determine an adventageous order. This paper proposes a novel solution for such a prioritization.

Another motivation to study this question is the fact that consecutive frames in a video sequence have very narrow baseline. Hence, it is not possible to process them in their default (temporal) order, since a wide baseline is often critical for the success of the structure estimation. A common practice is to employ frame skipping, but, obviously, a reliable frame pair is not guaranteed unless some properties of the motion is known beforehand.

For proper prioritization of the frame pairs, a priority metric should be designed according to the following two criteria:

- **Fast convergence to a reliable estimate:** In a sequential approach, since the quality of the subsequent reconstructions depend on the current (intermediate) structure estimate, failure to compute an accurate estimate in the first few pairs may cause the entire estimation procedure to collapse.
- **Fast recovery of the scene structure:** The number of reconstructed 3D points should be maximized, while processing a minimum number of frame pairs.

A priority metric that takes the baseline distance and the number of feature matches into account should cover both of these criteria. Hence, the pairs that are to be used in the reconstruction are selected based on a weighted sum of the baseline distance and the number of matching features. Notice that another important reliability indicator, trajectory length, is not considered, as it sacrifices many sufficiently good, yet relatively short-lived features while trying to attain reliability, hence conflicts the second criterion.

3. FEATURE TRACKING AND SEGMENTATION

Prior to dealing with any 3D reconstruction of a dynamic scene, the scene should be decomposed into its individual elements (i.e. the background and the IMOs). Since a wide baseline is needed both for the segmentation and the reconstruction processes, a slightly modified version of the well-known pyramidal Kanade-Lucas tracker is used to track features, selected with the Harris corner detector, along a sequence of consecutive frames.

The first modification is the padding of the lost tracks, i.e., if some features are lost during the tracking process, some additional features are selected again by the Harris corner detector in the current frame. The second modification is the key-frame selection to handle the baseline problem for the segmentation and the reconstruction part. After initializing the first frame of the sequence as the key-frame and proceeding frame by frame, we use the median distance between the features in the current frame transferred through an average planar-homography and the corresponding points in the previous key-frame, as introduced in [6].

Once the trajectories are constructed and the key-frames are selected, trajectory segmentation is handled via geometric means. For each independent motion in the sequence, there exists a corresponding F-matrix, $\mathbf{F}_i$, which fulfills the epipolar constraint

$$\mathbf{x}_1^T \mathbf{F}_i \mathbf{x}_2 = 0 \,, \tag{1}$$

where $\mathbf{x}_1$ and $\mathbf{x}_2$ are corresponding points in two views. A RANSAC-based F-matrix estimation algorithm identifies the feature pairs belonging to the dominant motion and labels the rest as outliers. If the same procedure is repeated with the remaining outliers, some of these should satisfy the epipolar constraint according to another F-matrix, which corresponds to the motion of an IMO. This procedure is repeated until no more reliable F-matrices can be determined. Hence, upon successive iteration of the procedure for all key-frames, the feature trajectories can be classified, either as background or belonging to one of the IMOs. Trajectories, which are labeled as outliers after RANSAC and re-RANSAC, are removed and not used in further computations. The segmentation algorithm can be summarized as follows:

Trajectory segmentation algorithm

1. Compute the F-matrix for the 1st and the 2nd key-frame by using a RANSAC-based procedure and label the inliers as background trajectories.
2. Compute the F-matrix on the outliers of Step 1 by using again RANSAC and label the inliers as trajectories belonging to the first IMO.
3. Repeat Step 2 as long as F-matrix estimation is still reliable and most of the remaining features are spatially close.
4. Proceed to the next key-frame. Estimate the F-matrix between the last and the current key-frame for each motion using the labeled trajectories and classify new trajectories by using Step 1 to 3.
5. Repeat Step 4 for all key-frames.

4. PRIORITIZED SEQUENTIAL RECONSTRUCTION

For the sake of clarity of the following discussion, two definitions are necessary.

Definition 1: A *sub-estimate* is a structure estimate obtained by the triangulation of the matching features in a single frame pair[1].

Definition 2: A *sub-reconstruction* is an intermediate structure estimate obtained from a collection of sub-estimates belonging to a subset of frames of the video sequence. Two distinct sub-reconstructions cannot have any common frames. Global motion and structure estimate is computed by merging the sub-

[1] It should be stressed that, while in this study a 2-view reconstruction approach is preferred due to the availability of relatively simple, mature and reliable techniques, a sub-estimate can be constructed by any of the methods existing in the literature.

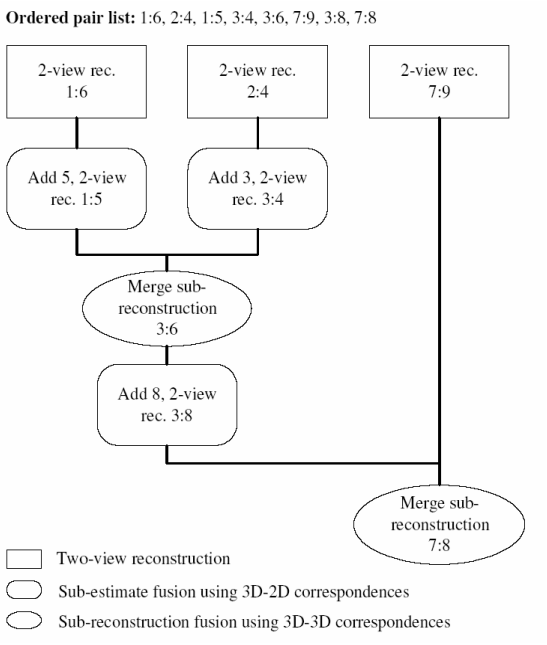

Figure 4.1: Sequential pair processing scheme

reconstructions.

The core of the reconstruction algorithm is based on [6], and its implementation is detailed in [9]. The basic idea is, starting with an initial reconstruction by triangulation, and adding new frames by first estimating their pose by 3D-2D matches, and then computing the sub-estimate corresponding to the last and the current frame, again via triangulation. This sub-estimate is used to add new points and refine the reconstruction for the existing ones.

This algorithm is designed to process the frame pairs in a certain order (e.g, F_1-F_2, F_2-F_3, F_3-F_4 , …), and changing the order of frame pairs requires some modifications.

Consider the pairs F_m-F_n and F_p-F_q., which are assumed in priority order with respect to the proposed metric. If the pairs have one common frame, then they can be processed by using the original algorithm [6], (i.e. n=q, then F_m-F_n, F_n-F_p) to obtain a single sub-reconstruction. If they have no common frames, two separate sub-reconstructions (each including a single sub-estimate) can be computed for each frame pair. Assume the latter occurs and let the sub-reconstructions be T_1 and T_2. Next, consider a third pair F_r-F_s. The cases that it has no common frames with neither of the sub-reconstructions, or has one with either of them are already handled. A new possible case is, one member belongs to T_1 and the other to T_2 (i.e. r=m and s=q).

The fusion of T_1 and T_2 requires a 4x4 projective transformation mapping the points in one of the sub-reconstructions to the other, as the coordinate system of each sub-reconstruction is determined by its first sub-estimate and different from the others. These coordinate systems are related by a rotation, a translation and a scale factor, caused by the normalization of the translation in the two-view reconstruction. The estimation procedure is derived from the robust 2D projective transformation estimation algorithm described in [4]. The basic idea is first to determine 3D-3D matches, then to use RANSAC to find a projective transformation that maps as many matches as possible, then to refine the estimate by using all available pairs and finally to further refine the estimate by a nonlinear minimization.

One possible final case is when both frames in the pair are already included in a single sub-reconstruction. In this case, one may skip the pair, or process it to obtain additional points. A typical reconstruction procedure is depicted in Figure 4.1. The complete reconstruction algorithm is summarized below:

Prioritized sequential 3D reconstruction

Given the internal calibration parameters and the correspondence information for all frames as trajectories:

1. Compute the initial reconstruction

2. Estimate the pose of each frame with respect to the first frame in the initial reconstruction by using the 3D-2D correspondences

3. Compute the priority metric and order the pairs

4. While the priority metric is above the threshold or all pairs are not processed

a. If no member of the pair belongs to any of the existing sub-reconstructions, initialize a new sub-reconstruction

b. If one member of the pair belongs to an existing sub-reconstruction, add the other frame to this sub-reconstruction (Algorithm in [6])

c. If two members of the pair belong to the same sub-reconstruction, process using the algorithm in [6].

d. If two members of the pair belong to different sub-reconstructions, merge the sub-reconstructions

5. If the number of remaining sub-reconstructions is greater than one, merge them all into a global estimate.

One last remaining issue is the choice of the pair for the initial reconstruction in Step 1. This pair should both be reliable and have as many common features as possible with the rest of the sequence, since the quality of the pose estimates depends on the number of matches. The key-frame pairs determined in the segmentation part are the obvious candidates for the initial frame pair.

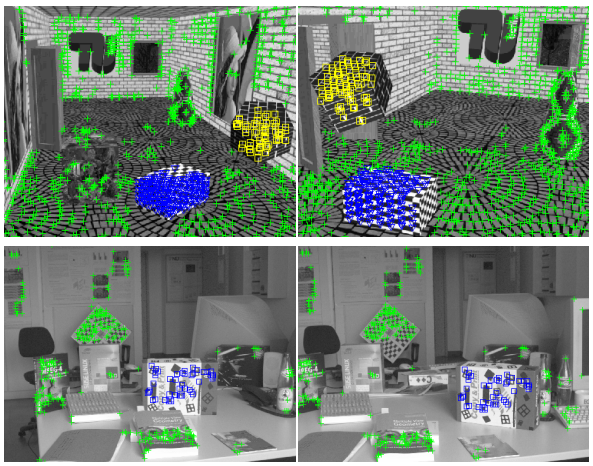

Figure 5.1: Segmentation results. TUB-Room (top) and Desk-sequence (bottom)

5. EXPERMENTAL RESULTS

The segmentation algorithm is tested on both synthetic and real data. In Figure 5.1, the segmentation results are presented. The pictures on the top show frame 1 and 170 of the 240-frame synthetic sequence "*TUB-Room*" with two independently moving

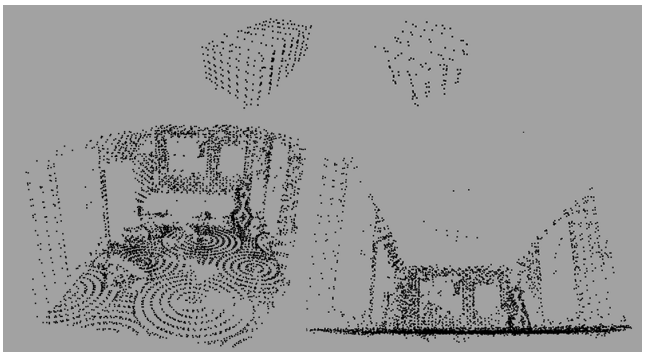

Figure 5.2: 3D reconstreuction results of the IMOs (top row) and the backround for "TUB-Room" (2 different front views)

objects in the scene. The features on the two IMOs are indicated with blue triangles and yellow squares, respectively. The background features are labeled with green crosses. The lower picture illustrates frames 1 and 13 of the "*Desk*"-sequence which was captured in an office. The features on the IMO are labeled with blue squares and the features of the background again with green crosses. The results confirm the good performance of the segmentation algorithm

The reconstruction of the synthetic scene is presented in Figure 5.2. The top row indicates the two IMOs. 331 features (left) and 113 features (right) were used for the reconstruction, respectively. The reconstructed background is shown from two different viewpoints (middle and bottom row). Here, the number of reconstructed features is 5014, out of 6095.

In Figure 5.3, the reconstruction results for *"Palace"*, a 208-frame sequence acquired from TV is depicted. Out of a total of 3546 features, 2771 of them are successfully reconstructed.

6. CONCLUSION

In this paper, an algorithm for the reconstruction of dynamic scenes in video sequences is proposed. The algorithm utilizes the epipolar constraint to partition the feature set into independent motions. Each partition is reconstructed separately by a sequential algorithm designed to efficiently process the large amount of information available in a video sequence. The key to achieve this objective is observed to be processing the pairs in an order that allows the extraction of the structure reliably from a small number of pairs. The experiments indicate that the algorithm performs well, as long as enough features are present. However, in practice, the lack of features is likely to cause significant problems, especially for IMOs significantly smaller than the background. A possible remedy for such cases is employing higher level geometric entities, such as lines or planes, to characterize the IMOs. The proposed method is an important step towards robust extraction of 3D information from an arbitrary 2D video content.

7. REFERENCES

[1] J. P. Costeira, T. Kanade, "A Multibody Factorization Method for Independently Moving-Objects", IJCV(29), No. 3, p. 159-179, September 1998

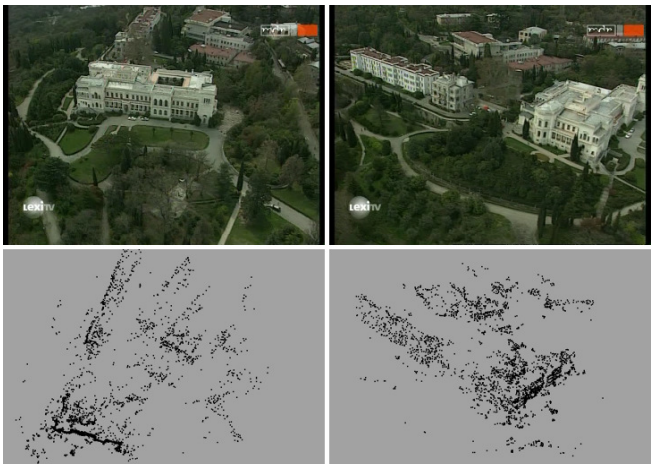

Figure 5.3: *Top row*: First and last frames of "Palace". *Bottom row:* Top and top-left views

[2] W. Fitzgibbon, A. Zisserman, "Multibody Structure and Motion: 3D Reconstruction of Independently Moving Objects", ECCV 2000, 2000

[3] M. Han, T. Kanade, "Perspective Factorization Methods for Euclidean Reconstruction", CMU-RI-TR-99-22, 1999

[4] R. Hartley, A. Zisserman, Multiple view geometry, Cambridge University Press, UK, 2003

[5] M. Irani, P. Anandan, "A Unified Approach to Moving Object Detection in 2D and 3D Scenes", IEEE Trans. On PAMI Vol. 20, Issue 6, pp. 577-589, June 1998

[6] M. Pollefeys, "Tutorial on 3D Modeling from Images", ECCV 2000, 2000.

[7] G. Qian, R. Chellappa, Q. Zheng, "Bayesian Algorithms for Simultaneous Structure from Motion Estimation of Multiple Independently Moving Objects", IEEE Trans. on Image Processing, Vol. 14, No.1, January 2005

[8] S. Soatto, P. Perona, "Reducing 'Structure from Motion': a General Framework for Dynamic Vision Part 1: Modeling", Pattern Analysis and Machine Intelligence, 20(9), September 1998

[9] E. Tola, "Multiview 3D Reconstruction of a scene containing independently moving objects", MS Thesis, METU Library, 2005

[10] C. Tomasi, T. Kanade, "Shape and Motion from Image Streams: A Factorization Method", Journal of Computer Vision 9(2), p.137-154, 1992

[11] Y. Weiss, "Segmentation Using Eigenvectors: A Unifying View", Proceedings of ICCV99, pp. 975-982, 1999