A GEOMETRIC SEGMENTATION APPROACH FOR THE 3D RECONSTRUCTION OF DYNAMIC SCENES IN 2D VIDEO SEQUENCES[*]

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

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

[†] EEE Department
Middle East Technical University
Balgat, 06531 Ankara, Turkey
E-mail:{eimre@, alatan@eee.}metu.edu.tr

ABSTRACT

In this paper, an algorithm is proposed to solve the multi-frame structure from motion (MFSfM) problem for monocular video sequences with multiple rigid moving objects. The algorithm uses the epipolar criterion to segment feature trajectories belonging to the background scene and each of the independently moving objects. As a large baseline length is essential for the reliability of the epipolar geometry, the geometric robust information criterion is employed for key-frame selection within the sequences. Once the features are segmented, corresponding objects are reconstructed individually using a sequential algorithm that is capable of prioritizing the frame pairs with respect to their reliability and information content. The experimental results on synthetic and real data demonstrate that our approach has the potential to effectively deal with the multi-body MFSfM problem.

1. INTRODUCTION

Structure from motion in static scenes is an extensively studied problem with some well established solutions [1]. However, these solutions are not capable of dealing with dynamic scenes with multiple moving objects, which are often encountered in practice. Hence, the intention of this study is to achieve both an accurate segmentation and reconstruction of the whole 3D scene including the dynamic elements.

In the literature, analysis of video sequences of dynamic scenes falls under the category of the multi-body MFSfM problem, which has the following definition for this special case:

Given a set of N features belonging to a background scene and K independently moving objects (IMOs), L views and the correspondence information, estimate the locations of the feature points in 3D world coordinates and the external calibration parameters.

Once 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. In this study, we will focus on the segmentation of multiple independently moving objects and on their reconstruction using a prioritized MFSfM approach.

The segmentation techniques handling the multi-body MFSfM problem can be divided into three categories. Optical flow based methods [2][3] assume a scene composed of planes of varying depths. In this case, a simple clustering of the optical flow values is sufficient to achieve the desired segmentation. Statistical techniques belong to the second category. In [4], the *sequential importance sampling* is employed to estimate simultaneously the structure from motion for multiple independently moving objects. The empirical posterior distribution of object motion and feature separation parameters is approximated by weighted samples.

Finally, it is possible to exploit the constraints derived from the epipolar geometry and the rigid body motion assumptions. The most common approach is to estimate the individual F-matrices for each motion, and to use the epipolar constraint for the classification [5][6][7]. However, in [8], different geometric constraints are available, and both the partitions and the models for each partition are determined after utilizing the geometric robust information criterion. Yet another technique is presented in [9], which exploits the rank constraint on the shape interaction matrix. The basic approach is an extension of the factorization method for SfM.

In this study, we use a geometric segmentation approach which is based on the epipolar constraint. The final 3D reconstruction step employs a two-frame triangulation. Both techniques achieve a better reliability for the large baseline case. However, a small baseline facilitates the solution of the correspondence problem. It is observed that the use of a tracker reduces the need for a compromise, providing a satisfactory solution to the correspondence problem, while providing a larger baseline.

The organization of this paper is as follows: In the next section, the tracking and segmentation algorithm is outlined. Section 3 describes the prioritized MFSfM approach. The experimental results are presented in Section 4. Finally, in Section 5, the paper is concluded by a discussion of the results and the future work.

2. FEATURE TRACKING AND SEGMENTATION

2.1. Feature Tracking

As a large baseline is needed for both the segmentation and the reconstruction processes, a slightly modified version of the well known pyramidal Lucas-Kanade tracker is used to track features

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

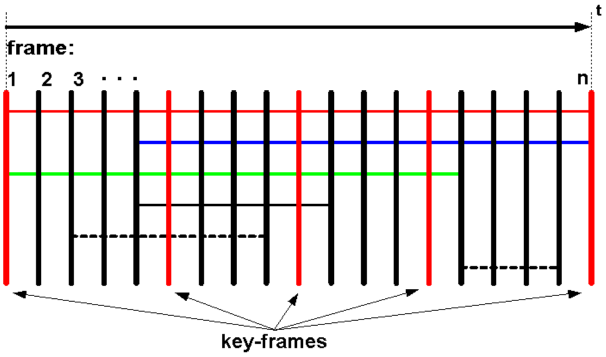

Figure 1: Trajectory classification and key-frame selection

(corners) along a sequence of consecutive frames. The first modification is the padding of lost tracks, i.e., if features get lost during the tracking process, additional features are selected again with the Harris corner detector in the current frame. The second one is the key-frame selection to handle the baseline problem for the segmentation and reconstruction part. Since the baseline between consecutive frames is small, a 2D motion model **H** (homography) can be used to transfer features from one frame to their corresponding positions in a second frame. If the baseline increases during the tracking process and if the features belong to a 3D scene structure, the projection error increases as well, i.e. the 2D motion model must be upgraded to a 3D motion model **F** (epipolar geometry).

The Geometric Robust Information Criterion (GRIC) [10] is a robust model selection criterion to extract key-frames and is defined as:

$$GRIC = \sum \rho\left(e_i^2\right) + \lambda_1 dn + \lambda_2 k \,, \tag{1}$$

where $\rho(e_i^2) = \min\left(e_i^2/\sigma^2,\ \lambda_3(r-d)\right)$. The parameters are defined as follows: d is the dimension of the selected motion model (**H** has the dimension two and **F** dimension three), r is the dimension of the data (i.e. four for two views), k is the number of the estimated model parameters (seven for **F** and eight for **H**), n is the number of tracked features, σ is the standard deviation of the error on each coordinate and e_i is the distance between a feature point transferred through **H** and the corresponding point in the target image or the Euclidian distance between the epipolar line of a feature point and its corresponding point in the target image (dependent on the selected model **M**):

$$e_i = D\left(x_i', \mathbf{M}x_i\right) \tag{2}$$

The parameters λ_1, λ_2, and λ_3 are tuning parameters with λ_1=2, λ_2=log(4n) and λ_3=2 [10].

Initializing the first frame of the sequence as key-frame and proceeding frame by frame, the next key-frame is selected if the GRIC value of the motion model **F** is below the GRIC value of **H**, i.e. a 2D motion model is no longer an accurate representation of the camera motion with respect to the 3D structure.

Figure 1 illustrates schematically n frames of a video sequence with some key-frames, indicated as vertical red lines, and 5 different kinds of feature trajectories. We use only the four upper kinds of trajectories for the segmentation and reconstruction, i.e., the ones that are visible in at least two consecutive key-frames.

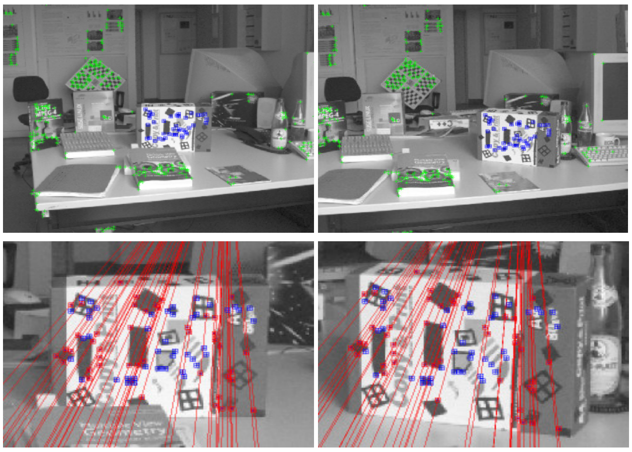

Figure 2: Trajectory segmentation (top) and guided matching (bottom) of the “Desk”-sequence

2.2. Segmentation

Once the features are tracked throughout the sequence and the key-frames are selected, trajectory segmentation is handled by geometric means. For each independent motion in the sequence, there exists a corresponding F-matrix, **F**$_i$, which fulfills the epipolar constraint

$$\mathbf{x}_1^T \mathbf{F}_i \mathbf{x}_2 = 0 \,, \tag{3}$$

where $\mathbf{x}_1$ and $\mathbf{x}_2$ are corresponding points in two views. A RANSAC (RANdom Sample Consensus)-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 outliers, some of the outliers should satisfy the epipolar constraint according to a new F-matrix, which corresponds to the motion of an IMO. This procedure is repeated as long as reliable F-matrices can be found for each IMO in the scene. Hence, upon successive iteration of the procedure for all key-frames, the feature trajectories can be classified, either as background or IMOs.

In case, where one of the IMOs has a similar motion as the camera, some features may be classified as IMO features even when they belong to the background or another IMO. We handle this problem by considering the distance of each of the segmented features to their centroid. A feature is rejected if its distance is higher than a predefined threshold dependent on the standard deviation of the distances:

$$dist_i > D\left(x_i, c\right) + \nu \cdot \sigma \,, \tag{4}$$

where c is the centroid of the data set, ν is a weighting factor and σ is the standard deviation.

Trajectories, which are labeled as outliers after RANSAC, re-RANSAC and distance check are removed and not used in further computations. Finally, we employ guided matching [1] along the epipolar lines in the key-frames to increase the number of IMO features: First, a bounding box is placed around the features on the IMO in the previous key-frame and additional features are selected with the Harris corner detector. These features are searched along their epipolar lines in the current key-frame, i.e. the search range is restricted to one dimension. Since the optical flow of the already segmented features is known, the search range on these lines can further be limited. A feature is considered as a match, if the normalized cross correlation (NCC) value is the highest among its neighbors and is above 0,8. The segmentation algorithm can be summarized as follows:

Algorithm 1: Trajectory segmentation

1. Compute the F-matrix corresponding to the first and the second 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 IMO trajectories.
3. Compute the centroid of the inliers of step 2 and check their distances. If the distance is higher than a threshold, reject the feature.
4. Increase the number of features on the IMO in consecutive key-frames with guided-matching.
5. Repeat step 2 to 4 as long as the F-matrix estimation is still reliable and most of the remaining features are spatially close.
6. 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 using step 1 to 5.
7. Repeat step 6 for all key-frames.

Figure 2 gives an example of the background and IMO trajectory segmentation (frame 1 and 13 of the "Desk"-sequence captured in an office). The green crosses indicate the background trajectories and the blue squares belong to the IMO. The lower two images show a close up of the IMO with the guided matching results, i.e. red squares on the corresponding epipolar lines.

3. PRIORITIZED RECONSTRUCTION

3.1 Prioritization

There are two motivations to study the prioritization problem. Firstly, one of the most significant advantages of the sequential methods is their ability to utilize the intermediate results from already processed frames to process the remaining ones. However, this provides the final reconstruction quality dependent on the processing order. Secondly, consecutive frames in a video sequence have a very narrow baseline, not allowing a reliable reconstruction. So, it is impossible to use the default (temporal) order.

Once it is established that there is an ordering problem, the next step is to determine a favorable order with respect to a metric. To design a proper prioritization metric, the following criteria should be considered:

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

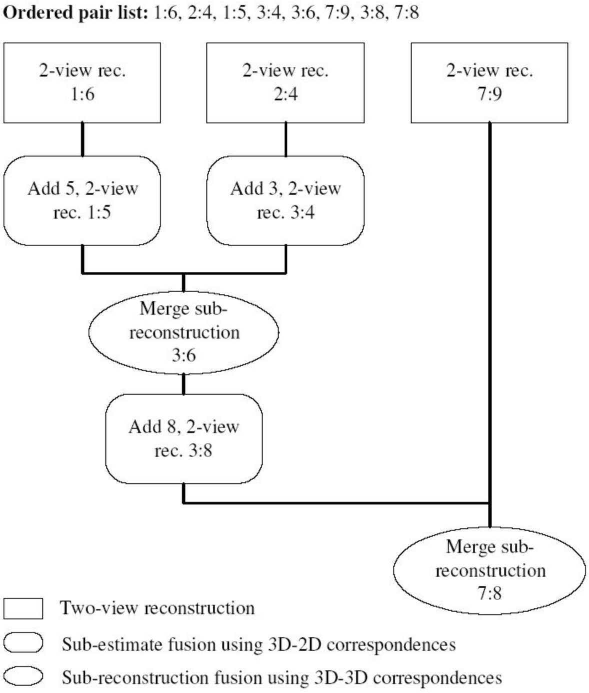
Ordered pair list: 1:6, 2:4, 1:5, 3:4, 3:6, 7:9, 3:8, 7:8
2-view rec. 1:6
2-view rec. 2:4
2-view rec. 7:9
Add 5, 2-view rec. 1:5
Add 3, 2-view rec. 3:4
Merge sub-reconstruction 3:6
Add 8, 2-view rec. 3:8
Merge sub-reconstruction 7:8
Two-view reconstruction
Sub-estimate fusion using 3D-2D correspondences
Sub-reconstruction fusion using 3D-3D correspondences

Figure 3: Sequential pair processing scheme

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.

3.2 Reconstruction

Prior to the description of the reconstruction algorithm, two definitions are necessary:

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

Definition 2: A *sub-reconstruction* is an intermediate structure estimate obtained form 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-reconstructions.

The core of the reconstruction algorithm is based on [11], and its implementation is detailed in [12]. The basic idea is, starting with an initial reconstruction by triangulation, and adding new frames

[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 existing methods.

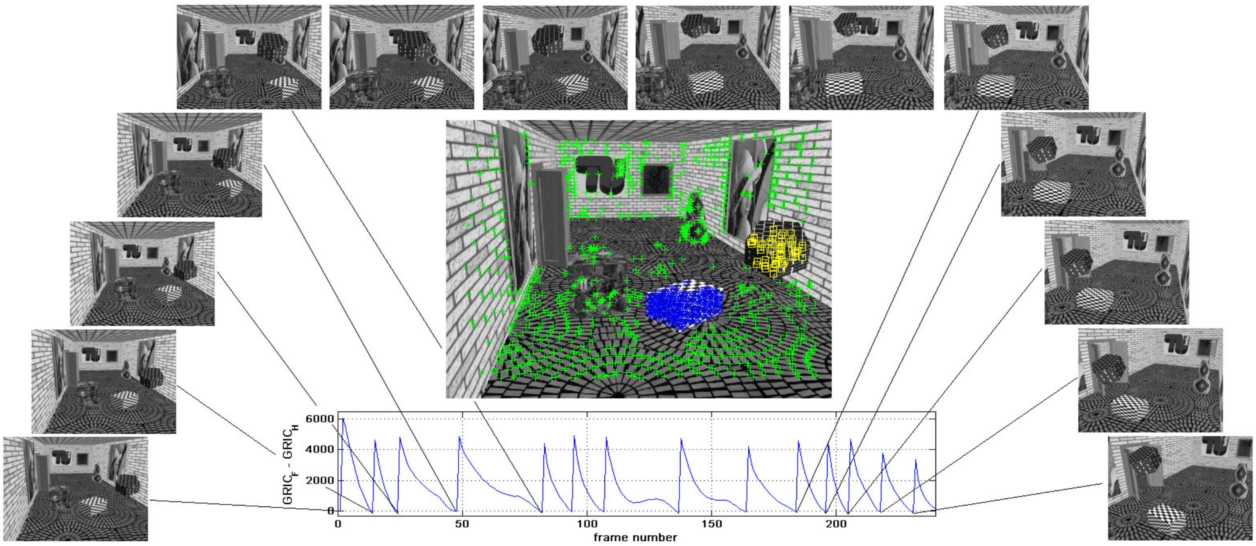

Figure 4: Segmentation results, GRIC score and selected key-frames for the 240 frame sequence "TUB-Room"

by first estimating their pose by 3D-2D matches, then computing the sub-estimate corresponding to the last and the current frame, again via triangulation and finally incorporating this sub-estimate into the reconstruction

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 two pairs, F_m-F_n and F_p-F_q. There are two possible cases: They may have one common frame (e.g. n=q, then F_m-F_n, F_n-F_p), in which case the sub-reconstruction can be computed using the original algorithm [11], or no common frames, which leads to two distinct sub-reconstructions, with a single sub-estimate each. Assume the latter occurs and let the sub-reconstructions be T_1 and T_2. Next, consider a third pair F_r-F_s with r=m and s=q, i.e. each frame belongs to separate sub-reconstructions.

The fusion of T_1 and T_2 requires the estimation of a similarity transformation defining a mapping between the points of the sub-reconstructions. The fundamentals of the estimation procedure are described in [1]. 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 3. The complete reconstruction algorithm is summarized below:

Algorithm 2: 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 [11]).

c. If two members of the pair belong to the same sub-reconstruction, process using again the algorithm in [11].

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 be both 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.

4. EXPERMENTAL RESULTS

The segmentation algorithm is tested on both synthetic and real data. In Figure 2, the segmentation results of the "Desk"-sequence are presented as mentioned in Section 2. Figure 4 shows the 14 key-frames, which are selected when the GRIC score falls below the zero-crossing, and the segmentation results of the synthetic sequence "TUB-Room" (240 frames) with two independently moving 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 reconstruction of the synthetic scene is presented in Figure 5. 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 6, the reconstruction results for *"Palace"*, a 208-frame sequence without IMOs acquired from TV is depicted. This sequence was chosen to show the good performance of the

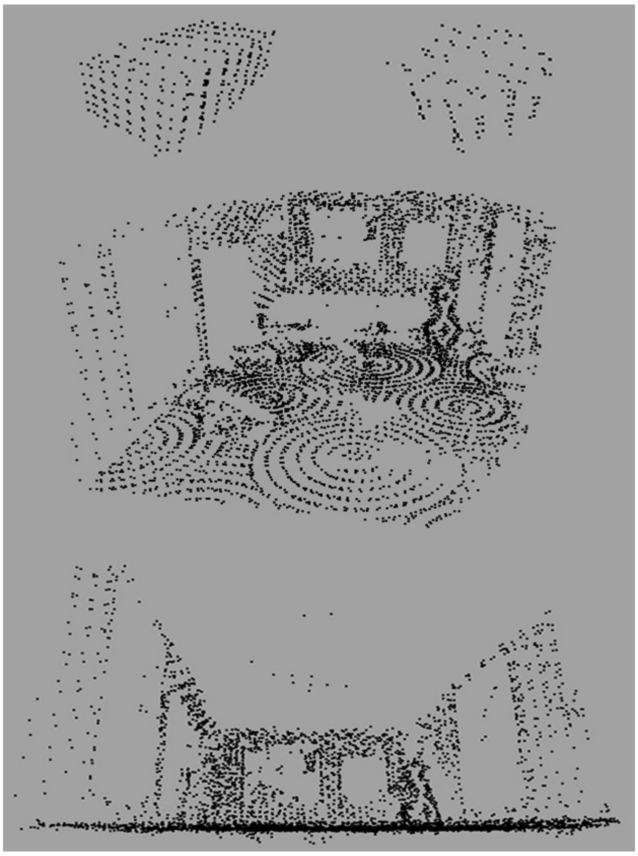

Figure 5: 3D reconstreuction results of the IMOs (top row) and the backround for “TUB-Room” (2 different front views)

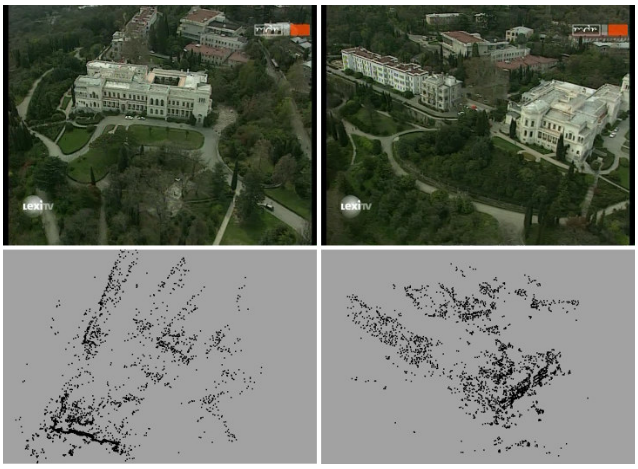

Figure 6**:** *Top row*: First and last frames of “Palace”. *Bottom row:* Top and top-left views

prioritized reconstruction approach for real data. Out of a total of 3546 features, 2771 of them are successfully reconstructed.

Due to sparse features on the walls, the reconstructed background of the "desk"-sequence is not shown. Moreover, the reconstruction of the IMO would fail because all features are located on a planar surface.

5. CONCLUSION

In this paper, a segmentation and reconstruction approach for dynamic scenes using video sequences is proposed. The algorithm utilizes the epipolar constraint to partition the feature set into independent motions. Since a large baseline is needed for a reliable F-matrix estimation, the geometric robust information criterion was employed for the key-frame selection. Once the features are segmented according to their motion, each partition is reconstructed separately by a sequential algorithm designed to efficiently process the large amount of information in the video sequence. The key to achieve this objective is processing the pairs in an order that allows the extraction of the structure reliably from a small number of pairs. The experiments indicate that for both the segmentation and the reconstruction the algorithm performs well, if enough features are present. However, in many real world sequences, the lack of features is likely to cause significant problems, especially for IMOs significantly smaller than the background.

The proposed method is an important step towards robust extraction of 3D information from an arbitrary TV broadcast video. Future works will focus on dense matching techniques to get more completed 3D models of the scene.

6. REFERENCES

[1] R. Hartley, A. Zisserman, Multiple View Geometry, Cambridge University Press, UK, 2003

[2] 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

[3] M. Lourakis, A. Argyros, S. Orphanoudakis, “Independent 3D Motion Detection Using Residual Parallax Normal Flow Fields”, Proceedings of ICCV98, Bombay, 1998

[4] G. Qian, R. Chellappa, Q. Zheng, “Bayesian Algorithms for Simultaneous Structure from Motion Estimation of Multiple Independently Moving Objects”, IEEE Trans. on IP, Vol. 14, Issue 1, Jan 2005

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

[6] R. Vidal, S. Soatto, Y. Ma, S. Sastry, “Two-view Multibody Structure from Motion”, IJCV 2002

[7] L. Wolf, A. Shashua, “Two-body Segmentation from Two Perspective Views”, Proc. of CVPR 2001, Vol. 1, 2001

[8] P.H.S. Torr, “Geometric Motion Segmentation and Model Selection”, Phil. Trans. Royal Society of London A 356, 1740,1321–1340, 1998

[9] J.P. Costeira, T. Kanade, “A Multibody Factorization Method for Independently Moving Objects”, IJCV’ 98

[10] P.H.S. Torr, A.W. Fitzgibbon and A. Zisserman, “The Problem of Degeneracy in Structure and Motion Recovery from Uncalibrated Image Sequences”, International Journal of Computer Vision, 32(1):27-44, August 1999

[11] M. Pollefeys, “Tutorial on 3D Modeling from Images”, ECCV 2000, 2000.

[12] E. Tola ., “Multiview 3D Reconstruction of a Scene Containing Independently Moving Objects”, MS Thesis, METU Library, 2005