ROBUST KEY FRAME EXTRACTION FOR 3D RECONSTRUCTION FROM VIDEO STREAMS

Mirza Tahir Ahmed, Matthew N. Dailey
School of Engineering and Technology, Asian Institute of Technology, Pathumthani, Thailand
mirza.tahir.ahmed@ait.ac.th, mdailey@ait.ac.th

Jose Luis Landabaso, Nicolas Herrero
Telefonica Research, Barcelona, Spain
jlldiaz@tid.es, nhm@tid.es

Keywords: 3D reconstruction, Key frame extraction, 3D video player, Geometrical robust information criterion (GRIC), 3D reconstruction degeneracy

Abstract: Automatic reconstruction of 3D models from video sequences requires selection of appropriate video frames for performing the reconstruction. We introduce a complete method for key frame selection that automatically avoids degeneracies and is robust to inaccurate correspondences caused by motion blur. Our method combines selection criteria based on the number of frame-to-frame point correspondences, Torr's geometrical robust information criterion (GRIC) scores for the frame-to-frame homography and fundamental matrix, and the point-to-epipolar line cost for the frame-to-frame point correspondence set. In a series of experiments with real and synthetic data sets, we show that our method achieves robust 3D reconstruction in the presence of noise and degenerate motion.

1 INTRODUCTION

Reconstructing a 3D scene from video requires choosing a number of representative (key) frames from the video stream. Automatic 3D reconstruction from snapshots and manually extracted video frames has been a focus of the structure-from-motion research community for a long time, but only a few researchers have carefully considered automatic selection of key frames prior from a video prior to the reconstruction process.

Estimation of 3D camera poses and recovery of 3D scene geometry are two very expensive processes in 3D reconstruction if performed with all frames in a video sequence. If the frames are decimated then these processes become less expensive. Additionally, consecutive frames may have baselines that are too short for accurate triangulation. Another important factor is that while the fundamental matrix provides extremely useful information about the relationship between two images of a general 3D structure related by general camera motion, in degenerate cases, when these generality assumptions for do not hold, fundamental matrix estimation fails.

We introduce a method for automatic key frame selection that takes all of these factors into account. It is based on the number of frame-to-frame point correspondences obtained, Torr's geometrical robust information criterion (GRIC, Torr, 1998), and the point-to-epipolar line cost for the frame-to-frame correspondence set to identify key frames. In a series of experiments with real and synthetic data sets, we show that our method achieves robust 3D reconstruction in the presence of noise and degenerate motion.

2 REQUIREMENTS

There three main reasons for extracting key frames from video sequences: computational performance, triangulation accuracy, and avoidance of degeneracy.

2.1 Computational performance

The same level of 3D reconstruction can be achieved from a few frames instead of processing all the frames in a video sequence. This will not only improve the performance but also the estimation of the 3D camera pose and recovery of 3D scene geometry will be computed more efficiently.

2.2 Triangulation accuracy

The baseline is the line between two camera centers. The baseline length is typically very small in consecutive frames. Long baselines are required for accurate triangulation. The size of a 3D point's region of uncertainty increases as the distance between two frames decreases. Therefore, the frame selection process should seek to maximize the baseline between the camera positions for key frames, subject to the constraint that a sufficient number of correspondences are retained.

2.3 Degeneracy avoidance

There are two conditions for non-general camera motion and non-general position of structure known as *degenerate cases* when the epipolar geometry is not defined and methods based on estimation of the fundamental matrix will fail (although note that the frame pair may still be useful for *resectioning*, in which we estimate only the camera position from known 3D-2D correspondences):

Motion Degeneracy: If the camera rotates about its center with no translation, the epipolar geometry is not defined.

Structure Degeneracy: When all of the 3D points in view are coplanar, the fundamental matrix cannot be uniquely determined from image correspondences alone.

3 PREVIOUS WORK

Here we provide an overview of the most relevant recent work in key frame selection. We mention the most relevant. Seo et al. (2003) consider three factors: (a) the ratio of the number of point correspondences found to the total number of point features found, (b) the homography error, and (c) the spatial distribution of corresponding points over the frames. Hartley and Zisserman (2004) state that the homography error is small when there is little camera motion between frames. Homography error is a good proxy for the baseline distance between two views. Seo et al. also encourage the use of evenly distributed correspondences over the entire image to obtain the fundamental matrix. They derive a score function from the above mentioned factors which is used to select key frames. The pair with the lowest score is selected as a key frame. The authors do not discuss any measure for degenerate cases.

Pollefeys and van Gool (2002) select key frames for structure and motion recovery based on a motion model selection mechanism (Torr et al., 1998) to select next key frame only once the epipolar geometry model explains the relationship between the pair of images better than the simpler homography model. The distinction between the homography and the fundamental is based on the geometric robust information criterion (GRIC, Torr, 1998). They discard all frames based on degenerate cases.

Seo et al. (2008) use the the ratio of the number of correspondences to the total number of features found. If the ratio is close to one this means the images overlap too much and the baseline length will be small. Under these assumptions, a frame should not be selected as a key frame. The second measure is the reprojection error. The pair of frames with minimum reprojection error are categorized as key frames. But as in their earlier work, no measures are taken for degenerate cases.

4 METHOD

We treat key frame selection as constrained optimization. Given the first frame of a video sequence, we seek to find the successor frame that 1) has a sufficiently long baseline (via a correspondence ratio constraint), 2) does not lead to degenerate motion or structure, and 3) has the best estimated epipolar geometry. We introduce our methods to achieve these criteria in this section.

4.1 Correspondence ratio constraint

We use Seo et al.'s (2008) *correspondence ratio* R_c as a proxy for baseline length:

$$R_c = \frac{T_c}{T_f}, \tag{1}$$

where T_c is the number of frame-to-frame point features in correspondence for the frame pair under consideration, and and T_f is the total number of point features considered for correspondence. R_c is inversely correlated with camera motion: as the camera moves, features in view tend to leave the scene, and the appearance of objects in view tends to change with perspective distortion, occlusion, and so on.

Although a long baseline is desirable for triangulation accuracy, if the number of corresponding features is too low, camera pose estimation accuracy will suffer. We therefore constrain candidate key frames to those having a correspondence ratio R_c between upper and lower thresholds T_1 and T_2. Currently, we set these thresholds through experimentation.

4.2 Degeneracy constraint

The relationship between a pair of images with general camera motion and general structure is appropriately defined by a fundamental matrix, whereas degenerate camera motion is more fittingly defined by a homography. We can thus use the relative quality of fit to distinguish general motion from degenerate motion. To assess the relative quality of fit, we use Torr's geometric robust information criterion (GRIC, Torr, 1998). GRIC is based not only on goodness of fit but also on the relative parsimony of the two types of models. The score, summed over the point correspondences, is

$$GRIC = \sum_i \rho(e_i^2)_i + \lambda_1 dn + \lambda_2 k, \qquad (2)$$

where $\rho(e_i^2)$ is a robust function

$$\rho(e_i^2) = \min(\frac{e_i^2}{\sigma^2}, \lambda_3(r-d))$$

of the residual e_i, d is the number of dimensions modeled ($d = 3$ for a fundamental matrix or 2 for a homography), n is the total number of features matched across the two frames, k is the number of degrees of freedom in the model ($k = 7$ for a fundamental matrix or 8 for a homography), r is the dimension of the data ($r = 4$ for 2D correspondences between two frames), σ^2 is the assumed variance of the error, $\lambda_1 = \log(r)$, $\lambda_2 = \log(rn)$, and λ_3 limits the residual error.

Given a candidate key frame, we calculate the GRIC score for the homography and the fundamental matrix models. If the GRIC score for the homography model is lower than the GRIC score for the fundamental matrix, we eliminate the frame as a candidate key frame.

4.3 Key frame selection criteria

We assume that the ith key frame has already been identified as the frame with index k_i (k_0 is just the first frame of the video sequence). Here we describe our method to select the next key frame k_{i+1}. Let $\phi(k_i)$ be the set of frame indices succeeding k_i for which the upper and lower bounds on the correspondence ratio R_c are satisfied and for which the GRIC score for the fundamental matrix model is better than the GRIC score for the homography model. We let

$$k_{i+1} = \text{argmax}_{j \in \phi(k_i)}(f(k_i, j)) \qquad (3)$$

where $f(i, j)$ is an objective function expressing one or more key frame goodness criteria for current key frame i and candidate next key frame j. We consider two criteria, GRIC difference and the point-to-epipolar line cost (PELC).

4.3.1 GRIC difference criterion

If the GRIC score of the fundamental matrix model is much better than that of the homography model, the relationship between the frames is much better represented by the fundamental matrix model, indicating a good candidate key frame. We use the normalized GRIC difference as one possible criterion for selecting the next key frame:

$$f_G(i, j) = \frac{GRIC_{\mathrm{H}}(i, j) - GRIC_{\mathrm{F}}(i, j)}{GRIC_{\mathrm{H}}(i, j)}, \qquad (4)$$

where $GRIC_{\mathrm{H}}(i, j)$ is the GRIC score from Equation (2) for the homography between frames i and j, and $GRIC_{\mathrm{F}}(i, j)$ is the GRIC score for the fundamental matrix for frames i and j. As we shall see, this measure is good for selecting key frames because it provides very low variation in reprojection error as compared to uniformly sampled frames.

4.3.2 PELC criterion

The GRIC difference method tends to stabilize variation in reprojection error, but as we shall see in the experimental results, it has little effect on the mean error compared to uniformly sampled frames. We analyzed the GRIC difference scores and the point-to-epipolar line cost over many frames in real image sequences and observed some frames in which the variation in the GRIC difference was very small but the variation in the point-to-epipolar line cost (PELC) was very high, as shown for example in Figure 1. We found that high PELC values tended to occur due to inaccurate correspondences with blurry images in our video sequences.

We thus consider PELC as an additional criterion for key frame selection. As we shall see in the experimental results, including both the GRIC difference and the PELC in the key frame selection criteria helps us find key frames that are both well explained by the epipolar geometry and have highly accurate correspondences. We therefore propose the alternative key frame score

$$f_{GP}(i, j) = w_G f_G(i, j) + w_P(\sigma - PELC(i, j)), \qquad (5)$$

where σ is the assumed standard deviation of the error and $PELC$ is the standard geometric reconstruction error measure for the fundamental matrix (Hartley and Zisserman, 2003). The weights w_G and w_P could be set automatically, but we currently set them experimentally.

4.4 Algorithm summary

The complete method for key frame selection is summarized in Algorithm 1.

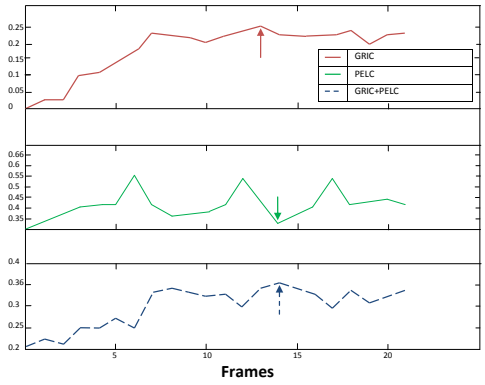

Figure 1: Variation of GRIC difference and PELC. Frame 0 is an assumed previous key frame. The GRIC difference is maximal for frame 13, but PELC has a local minimum at frame 14. Since there is only a small change in GRIC difference between frame 13 and 14 but a much improved PELC, the GRIC+PELC method (Equation 5) selects frame 14 as the next key frame.

Algorithm 1 KEYFRAMEEXTRACTION

Input: A video stream with n frames.
Output: Key frame index sequence $k_0, k_1, \ldots$
$i \leftarrow 0$; $j^* \leftarrow 0$
while $j^* \neq \perp$ **do**
  $k_i \leftarrow j^*$; $i \leftarrow i+1$; $j^* \leftarrow \perp$
  for candidate frame $j \in k_i + 1..n$ **do**
    Match keypoints between frames k_i and j
    Compute H and F using RANSAC
    Discard outlier matches
    Calculate correspondence ratio R_c
    if $R_c < T_{min}$ or $R_c > T_{max}$ **then**
      continue
    end if
    if $GRIC_{\mathrm{H}}(k_i, j) \leq GRIC_{\mathrm{F}}(k_i, j)$ **then**
      continue
    end if
    if $f_{GP}(k_i, j)$ is best so far **then**
      $j^* \leftarrow j$
    end if
  end for
end while

5 EXPERIMENTS AND RESULTS

We performed experiments with both synthetic and real data. The synthetic data is useful because we can precisely identify degenerate motion and structure; the real data is useful for validate the method's robustness to real-world noise.

5.1 Video sequences

Here we provide details about each experimental video sequence.

Synthetic data: Church We created a 930-frame synthetic sequence with ground truth data using Blender (Blender Community, 2009) and a sample 3D model of a church (Blender Artists, 2000). The scene is outdoors, with sky in the background. We inserted degenerate cases of both types by rotating the camera view point about its center or zooming in on planar surfaces. We extracted 3D points, projected 2D points, camera projection matrices, and the ground truth depth for every 2D point.

Indoor data These sequences were captured indoors at Telefonica Research, Barcelona, with a Sony HDR camera. We performed manual calibration of the camera intrinsic parameters using a chessboard pattern.

Library A 1500-frame sequence in the Telefonica library.

Lunch Room A 1500-frame sequence in the Telefonica lunch room.

Imagenio A 1500-frame sequence in the Imagenio room at Telefonica.

Nico A 1000-frame sequence of a person sitting still in a chair.

Photocopy Machine A 1200-frame sequence of a photocopy machine.

5.2 Experiment 1: Degeneracy

In Experiment 1, we tested degenerate case identification in the Church sequence. We processed every 10th frame as a candidate key frame. We manually identified 26 frames consisting of degenerate motion or structure.

Figure 2 shows the $GRIC_{\mathrm{H}}(i, i+10)$ and $GRIC_{\mathrm{F}}(i, i+10)$ scores for each frame i considered. Frames 201–271 and 441–571 consist of pure camera rotations, and frames 761–881 only contain coplanar points.

Table 1 shows the detection rate and error rate for the degenerate cases in the data set. The method is able to identify the actual degenerate cases perfectly, with only 3 false positives among the 93 frames tested.

5.3 Experiment 2: 3D reconstruction

In Experiment 2, we compared uniformly sampled key frames, key frames selected based on the GRIC

Table 1: Detections and errors for degenerate motion and structure detection in Experiment 1. FPs = false positives; FNs = false negatives.

Sequence	Positives	FPs	FNs
Church	26	3	0

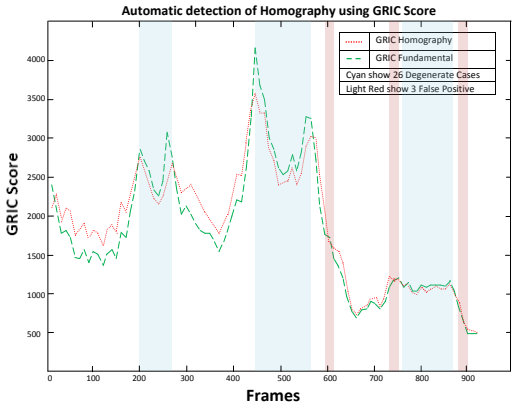

Figure 2: GRIC scores for the synthetic Church sequence. The red dash-dot line shows $GRIC_{\mathrm{H}}(i, i+10)$; the green dash-dot line shows $GRIC_{\mathrm{F}}(i, i+10)$. The cyan areas show when the when the homography model is preferred. The light red areas indicate false positives.

difference score, and key frames selected using PELC as well as the GRIC difference score. For each sequence, we performed key frame selection then applied Telefonica's structure from motion pipeline (metric reconstruction from the essential matrix for the first pair followed by resectioning and bundle adjustment for subsequent key frames) to obtain a 3D point cloud from the key frames. We computed the root mean reprojection error for each frame then computed the min, max, mean, and standard deviation statistics over the entire sequence. A numerical comparison of the three methods is shown in Table 2, and the per-frame reprojection errors are shown for three sequences in Figure 3.

The GRIC difference score method yields much lower reprojection error lower variance than uniform sampling in almost every case, but the mean reprojection error is not much better than that for uniform sampling, due to a few outlier frames. A manual inspection revealed that the outlier frames tended to be those with significant blur, leading to inaccurate correspondences, even for the inlier correspondences. Including PELC in the objective function eliminates these outlier frames and leads to lower mean reprojection error and lower variance for all of the real sequences. PELC does not help much on the noise-free synthetic sequence, however.

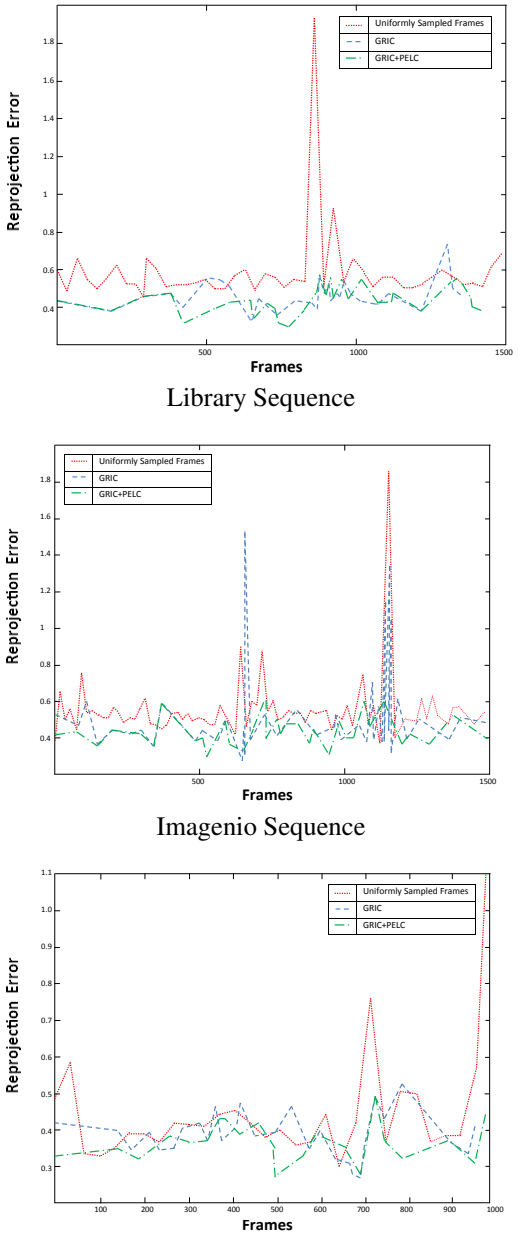

Figure 3: Reprojection error comparison. Red dotted lines: Uniform sampling. Blue dash lines: GRIC. Green dot-dash lines: GRIC+PELC.

6 DISCUSSION AND CONCLUSION

We have demonstrated the feasibility of automatic key frame selection using a combination of constraints based on the correspondence ratio and the GRIC score for the homography and fundamental matrix,

Table 2: Results of Experiment 2. The GRIC+PELC method obtains the lowest mean reprojection error and lowest error variance on all real video sequences.

Sequence	Method	Key frames	Reprojection Error			
			Min	Max	Mean	σ
Church (synthetic)	Uniform	30	0.2929	0.9608	0.4451	0.0221
	GRIC	37	0.2627	0.5498	0.3878	0.0029
	GRIC+PELC	29	0.2752	**0.5164**	0.3748	**0.0026**
Library	Fixed	43	0.4681	1.9595	0.6108	0.1434
	GRIC	34	0.3362	0.7222	0.4738	0.0046
	GRIC+PELC	37	**0.2519**	**0.4852**	**0.3971**	**0.0022**
Imagenio	Uniform	84	0.2016	1.8660	0.5415	0.0294
	GRIC	79	0.2786	1.5150	0.4772	0.0363
	GRIC+PELC	65	**0.2493**	**0.6042**	**0.4275**	**0.0049**
Nico	Uniform	30	0.3052	1.2125	0.4668	0.0290
	GRIC	29	0.2760	0.5533	0.4032	0.0043
	GRIC+PELC	29	0.2760	**0.5155**	**0.3780**	**0.0031**
Photocopy machine	Uniform	51	0.3210	1.7264	0.4649	0.0368
	GRIC	51	0.3210	0.7955	0.4707	0.0116
	GRIC+PELC	58	0.3274	**0.5682**	**0.4324**	**0.0036**

followed by optimization of a criterion including the GRIC difference and the point-to-epipolar line cost.

We find that the relative quality of the fundamental matrix and homography models, represented by the GRIC difference, is more important than the point to epipolar line cost, but both are useful in key frame selection when some frames are corrupted by blur.

One possible limitation of our method is the need to specify the thresholds and weights. We currently set these free parameters experimentally. However, since all of the parameters are relative to the number of correspondences obtained or the overall residual error, in principle, it should be possible to find values that work well for most sequences and allow the user to adjust them when necessary.

By limiting further 3D reconstruction processing to the most informative frames, our method helps to minimize the overall compute time of the video processing pipeline. Telefonica is deploying the method in an upcoming product for video surfing, which is in the last phase of development.

Future work will focus on enhancing the system for robustness with arbitrary videos. The key frame selection method may have to interact with other processes such as moving object segmentation and autocalibration to achieve this goal.

ACKNOWLEDGEMENTS

MTA was supported by a graduate fellowship from the Higher Education Commission of Pakistan. We thank Telefonica Research, Barcelona for providing the environment for this research and we thank Guillermo Gallego and Jose Carlos for valuable suggestions.

REFERENCES

Blender Artists (2000). 3D church model. Available at `http://blenderartists.org/cms/`.

Blender Community (2009). Blender [open source software]. Available at `http://www.blender.org/`.

Hartley, R. and Zisserman, A. (2003). *Multiple View Geometry in Computer Vision*. Cambridge University Press, New York, NY, USA.

Pollefeys, M. and Van Gool, L. (2002). Visual modeling with a hand held camera. *Journal of Visualization and Computer Animation (JVCA)*, 13:199–209.

Seo, Kim, Doo, and Choi (2008). Optimal keyframe selection algorithm for three-dimensional reconstruction in uncalibrated multiple images. *Society of Photo-Optical Instrumentation Engineers (SPOIE)*, Vol. 47.

Seo, Kim, Jho, and Hong (2003). 3D estimation and keyframe selection for match move. *International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)*.

Torr, P., Fitzgibbon, A., and Zisserman, A. (1998). Maintaining multiple motion model hypotheses over many views to recover matching and structure. pages 485–491.

Torr, P. H. S. (1998). Geometric motion segmentation and model selection. *Phil. Trans. Royal Society of London A*, 356:1321–1340.