Keyframe Selection for Camera Motion and Structure Estimation from Multiple Views

Thorsten Thormählen, Hellward Broszio, and Axel Weissenfeld

Information Technology Laboratory, University of Hannover,
Schneiderberg 32, 30167 Hannover, Germany
{thormae, broszio, aweissen}@tnt.uni-hannover.de
http://www.tnt.uni-hannover.de/~thormae

Abstract. Estimation of camera motion and structure of rigid objects in the 3D world from multiple camera images by bundle adjustment is often performed by iterative minimization methods due to their low computational effort. These methods need a robust initialization in order to converge to the global minimum. In this paper a new criterion for keyframe selection is presented. While state of the art criteria just avoid degenerated camera motion configurations, the proposed criterion selects the keyframe pairing with the lowest expected estimation error of initial camera motion and object structure. The presented results show, that the convergence probability of bundle adjustment is significantly improved with the new criterion compared to the state of the art approaches.

1 Introduction

The estimation of camera motion and structure of rigid objects in the 3D world using camera images from multiple views has a long and sophisticated research history within the computer vision community.

Usually a mathematical parameter model of a pinhole camera with perspective projection is used to describe the mapping between the 3D world and the 2D camera image. To estimate the parameters of the camera model most approaches establish corresponding feature points in each view. By the introduction of a statistical error model, that describes the errors in the position of the detected feature points, a Maximum Likelihood estimator can be formulated that simultaneously estimates the camera parameters and the 3D positions of feature points. This joint optimization is called bundle adjustment [1].

If the errors in the positions of the detected feature points obey a Gaussian distribution, the Maximum Likelihood estimator has to minimize a nonlinear least squares cost function. In this case, fast minimization is carried out with iterative parameter minimization methods, like the sparse Levenberg-Marquardt method [1][2, Appendix 4.6].

The main difficulty of the iterative minimization is the robust initialization of the camera parameters and the 3D positions of feature points in order to converge

T. Pajdla and J. Matas (Eds.): ECCV 2004, LNCS 3021, pp. 523–535, 2004.

to the global minimum. One possible solution is to obtain an initial guess from two [3,4] or three [5,6] selected views out of the sequence or sub-sequence. These views are called keyframes.

Keyframes should be selected with care, for instance a sufficient baseline between the views is necessary to estimate initial 3D feature points by triangulation. Additionally, a large number of initial 3D feature points is desirable.

By comparison, keyframe selection has been neglected by the computer vision community. In the case of initialization from two views, Pollefeys et al. [3] use the Geometric Robust Information Criterion (GRIC) proposed by Torr [7]. This criterion allows to evaluate which model, homography (H-matrix) or epipolar geometry (F-matrix), fits better to a set of corresponding feature points in two view geometry. If the H-matrix model fits better than the F-matrix model, H-GRIC is smaller than F-GRIC and vice versa. For very small baselines between the views GRIC always prefers the H-matrix model. Thus, the baseline must exceed a certain value before F-GRIC becomes smaller than H-GRIC.

Pollefeys' approach searches for one keyframe pairing by considering all possible pairings of the first view with consecutive views in the sequence. Thus, the first keyframe of the keyframe pairing is always the first view of the sequence. The second keyframe is the last view for which the number of tracked feature points is above 90% of the number of feature points tracked at the view for which F-GRIC becomes smaller than H-GRIC. This approach guarants a certain baseline and a large number of initial 3D feature points.

Gibson et al. [4] propose a quite similar approach. Instead of GRIC they evaluate a score consisting of three weighted addends. The first addend becomes small if the number of reconstructed initial 3D feature points reduces in the actual keyframe pair compared to the previous keyframe pair. The second addend is the reciprocal value of the median reprojection error when a H-matrix is fitted to the feature points and the third addend is the median reprojection error when the F-matrix model is applied. Gibson's approach marks the pairing with the lowest score as keyframes.

The disadvantage of both approaches is, that they do not select the best possible solution. For instance, a keyframe pairing with a very large baseline is not valued better than a pairing with a baseline that just ensures that the F-matrix model fits better than the H-matrix model. Thus, only the degenerated configuration of a pure camera rotation between the keyframe pairing is avoided. Especially, if the errors in the positions of the detected feature points are high, these approaches may estimate a F-matrix, that does not represent the correct camera motion and therefore provides wrong initial parameters for the bundle adjustment.

The approach for keyframe selection presented in this paper formulates a new criterion using techniques from stochastic. By evaluating the lower bound for the resulting estimation error of initial camera parameters and initial 3D feature points, the keyframe pairing with the best initial values for bundle adjustment is selected. It will be shown that this new approach increases significantly the convergence probability of the bundle adjustment.

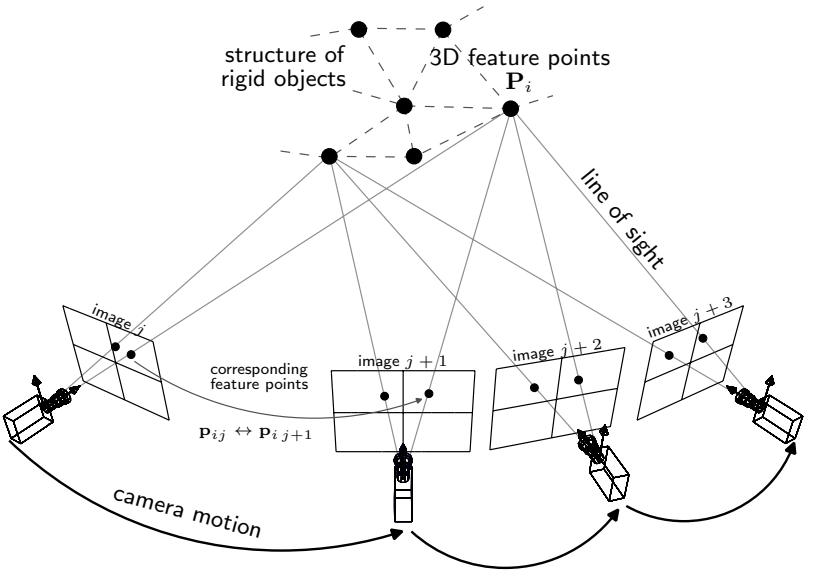

Fig. 1. Projection of 3D feature points on rigid objects in multiple camera views.

The following chapter defines a reference framework for the keyframe selection approaches by describing the processing steps that are used for the estimation of camera motion and structure of the observed objects. In Chapter 3 the new approach for keyframe selection is presented. Chapter 4 compares results of the different approaches in the defined framework and conclusions are drawn in Chapter 5.

2 Reference Framework

For estimation of camera motion parameters from corresponding feature points, the real camera must be represented by a mathematical camera model. The camera model describes the projection of a 3D feature point $\mathbf{P}$ to the image coordinate $\mathbf{p}$ through a perspective camera. Using homogeneous representation of coordinates, a 3D feature point is represented as $\mathbf{P} = (X, Y, Z, 1)^\top$ and a 2D image feature point as $\mathbf{p} = (x, y, 1)^\top$. Where $\mathbf{p}_{ij}$ is the projection of a 3D feature point $\mathbf{P}_i$ in the j-th camera (see Fig. 1), with

$$\mathbf{p}_{ij} \sim \mathbf{K}_j \; [\mathbf{R}_j | \mathbf{t}_j] \; \mathbf{P}_i = \mathbf{A}_j \, \mathbf{P}_i \quad \forall \, j \in \{1, \ldots, J\}, \, i \in \{1, \ldots, I\} \tag{1}$$

where $\mathbf{K}_j$ is the calibration matrix, $\mathbf{R}_j$ is the rotation matrix, $\mathbf{t}_j$ is the translation vector, and $\mathbf{A}_j$ is the camera matrix of the j-th camera. The software system used for estimation of $\mathbf{A}_j$ and $\mathbf{P}_i$ consists of five processing steps, as shown in Fig. 2. Each processing step is described briefly in the following subsections. Detailed related reading may be found in [2].

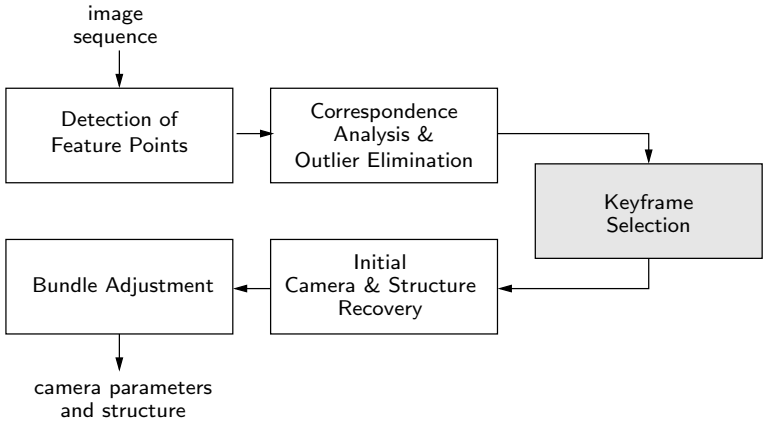

Fig. 2. Processing steps for the estimation of camera parameters and structure from image sequences

2.1 Detection of Feature Points

2D image feature points $\tilde{\mathbf{p}}$ are detected with sub-pixel accuracy using Harris' feature point detector [8]. For each image j of the sequence a list of feature point coordinates $L_j = \{\tilde{\mathbf{p}}_{1j}, \ldots, \tilde{\mathbf{p}}_{ij}, \ldots, \tilde{\mathbf{p}}_{Ij}\}$ is extracted. Due to noise in the intensity values of the images, the positions of the detected feature points $\tilde{\mathbf{p}} = (\tilde{x}, \tilde{y}, 1)^\top$ differ from the true positions $\mathbf{p} = (x, y, 1)^\top$, with

$$\tilde{x} = x + \Delta x \quad \text{and} \quad \tilde{y} = y + \Delta y \tag{2}$$

The error model in this paper assumes that Δx and Δy of all points $\tilde{\mathbf{p}}_{ij}$ are uncorrelated and obey a zero-mean Gaussian distribution with covariance matrix

$$\Sigma_{\tilde{\mathbf{p}}_{ij}} = \begin{pmatrix} \sigma^2_{x_{ij}} & 0 \\ 0 & \sigma^2_{y_{ij}} \end{pmatrix} \tag{3}$$

2.2 Correspondence Analysis and Outlier Elimination

The feature points in list L_j and L_{j+1} of two successive views are assigned by measuring normalized cross-correlation between 15×15 pixel windows surrounding the feature points. The correspondences are established for those feature points, which have the highest cross-correlation. This results in a list of correspondences $L_c = \{q_1, \ldots, q_i, \ldots, q_I\}$, where $q_i = (\tilde{\mathbf{p}}_{ij}, \tilde{\mathbf{p}}_{i\,j+1})$ is a correspondence.

Due to erroneous assignment of feature points arising from moving objects, illumination changes or similarities in the scene, usually some of the correspondences are incorrect. Most of these outliers can be detected because they must fulfill the epipolar constraint between two views:

$$\mathbf{p}^\top_{i\,j+1}\,\mathbf{F}\,\mathbf{p}_{ij} = 0 \quad \forall\, i \quad \text{and} \quad \det(\mathbf{F}) = 0 \tag{4}$$

where $\mathbf{F} = \mathbf{K}_{j+1}^{-\top}[\mathbf{t}_j]_x \mathbf{R} \mathbf{K}_j^{-1}$ is the F-matrix. In the case of motion degeneracy, if the camera does not translate between the views, or structure degeneracy, if the viewed scene structure is planar, a homography is the stricter constraint between two views:

$$\mathbf{p}_{i\,j+1} = \mathbf{H}\,\mathbf{p}_{ij} \quad \forall\, i \tag{5}$$

where $\mathbf{H} = \mathbf{K}_{j+1} \mathbf{R} \mathbf{K}_j^{-1}$ is the H-matrix. $\mathbf{H}$ or $\mathbf{F}$ should be estimated by minimizing the residual error $\bar{e}$ of the Maximum Likelihood cost function for the used error model, consequently here:

$$\bar{e}^2 = \frac{1}{4\,I} \sum_{i=1}^{I} d(\tilde{\mathbf{p}}_{ij}, \hat{\mathbf{p}}_{ij})_\Sigma^2 + d(\tilde{\mathbf{p}}_{i\,j+1}, \hat{\mathbf{p}}_{i\,j+1})_\Sigma^2 = \frac{1}{4\,I} \sum_{i=1}^{I} e_i^2 \longrightarrow \min \tag{6}$$

subject to $\hat{\mathbf{p}}_{ij}$ and $\hat{\mathbf{p}}_{i\,j+1}$ fulfill exactly Eq. 4 for F-matrix estimation and Eq. 5 for the estimation of H-matrix, where $d(\dots)_\Sigma$ denotes the Mahalanobis distance for the given covariance matrices, here $\Sigma_{\tilde{\mathbf{p}}_{ij}}$ and $\Sigma_{\tilde{\mathbf{p}}_{i\,j+1}}$. To achieve a robust estimation the random sampling algorithm MSAC (see [9,10] for details) is employed.

After estimation of $\mathbf{H}$ and $\mathbf{F}$, Torr's GRIC is used to decide which of the both models should be used for outlier elimination and guided matching [7].

$$\text{GRIC} = \left(\sum_{i=1}^{I} \rho(e_i^2) \right) + \lambda_2\, m\, I + \lambda_2\, k \tag{7}$$

$$\text{with} \qquad \rho(e^2) = \left\{ \begin{array}{ll} \frac{e^2}{\sigma^2} & \text{for} \quad \frac{e^2}{\sigma^2} < \lambda_3(r-m) \\ \lambda_3(r-m) & \text{for} \quad \frac{e^2}{\sigma^2} \geq \lambda_3(r-m) \end{array} \right\} \tag{8}$$

where k is number of essential parameters of the model, m is dimension of the fitted manifold, and r is dimension of the measurements, with $k = 7$, $m = 3$, $r = 4$ for F-GRIC and $k = 8$, $m = 2$, $r = 4$ for H-GRIC. The model with the lower GRIC is indicated as more likely.

2.3 State of the Art in Keyframe Selection

In the keyframe selection step keyframe pairings are determined to start the initial camera and structure recovery in the following step.

In general, many possible keyframe pairings exist. To reduce complexity Pollefeys' and Gibson's approaches always set the first keyframe of a keyframe pairing at the first view. Then consecutive views of the sequence are considered. For comparability this procedure is also adopted in our reference framework.

Pollefeys' approach chooses as second keyframe the last view for which the number of tracked feature points is above 90% of the number of feature points tracked a the view where F-GRIC becomes smaller than H-GRIC.

In Gibson's approach the following score S_g is evaluated for each pairing of views:

$$S_g = w_1 \left(1.0 - \frac{I_1}{I_2} \right) + w_2 \frac{1}{\bar{e}_H^2} + w_3\, \bar{e}_F^2 \tag{9}$$

where I_2 is the number of 3D feature points that were reconstructed in the previous pair and I_1 is the number of those features that can also be reconstructed in the currently evaluated pair, $\bar{e}_H$ is the residual error defined in Eq. 6 with the H-matrix model fitted to the data, and $\bar{e}_F$ is the residual error for the F-matrix model. The pairing with the lowest S_g is marked as new keyframe. Gibson suggests to choose the weights $w_1 = 3$, $w_2 = 10$, $w_3 = 1$.

Pollefeys and Gibson apply different optimization strategies in their bundle adjustment step. While Pollefeys' approach uses Incremental Bundle Adjustment, Gibson's approach uses Hierarchical Merging of Subsequences. In the incremental approach one keyframe pairing per sequence must be selected. In contrast, the hierarchical approach divides the sequence into subsequences according to the chosen keyframes. Thus, in this case one keyframe pairing per subsequence is available.

In order to compare the two state of the art approaches and the new approach, a common framework must be defined. In our reference framework the incremental approach is used in the bundle adjustment step. Hence, only one keyframe pairing per sequence is selected.

2.4 Initial Camera and Structure Recovery

After a keyframe pairing is selected a F-matrix between the keyframes is estimated by MSAC using Eq. 6 with Eq. 4 as cost function. The estimated F-matrix is decomposed to retrieve initial camera matrices $\hat{\mathbf{A}}_{k1}$ and $\hat{\mathbf{A}}_{k2}$ of both keyframes. Initial 3D feature points $\hat{\mathbf{P}}'_i$ are computed using triangulation (see [2, Chapter 11]). Now bundle adjustment between two views is performed by sparse Levenberg-Marquardt iteration using Eq. 6 subject to $\hat{\mathbf{p}}_{i\,k1} = \hat{\mathbf{A}}_{k1}\hat{\mathbf{P}}'_i$ and $\hat{\mathbf{p}}_{i\,k2} = \hat{\mathbf{A}}_{k2}\hat{\mathbf{P}}'_i$ as cost function. Initial camera matrices $\hat{\mathbf{A}}_j$, with $k_1 < j < k_2$, of the intermediate frames between the keyframes are estimated by camera resectioning. Therefore, the estimated 3D feature points $\hat{\mathbf{P}}'_i$ become measurements $\tilde{\mathbf{P}}'_i$ in this step. Assuming the errors mainly in $\tilde{\mathbf{P}}'_i$ and not in $\tilde{\mathbf{p}}_{ij}$ the following cost function must be minimized:

$$\bar{\mu}^2_{\text{res}} = \frac{1}{3\,I}\sum_{i=1}^{I} d(\tilde{\mathbf{P}}'_i, \hat{\mathbf{P}}_i)^2_{\Sigma} \longrightarrow \min \tag{10}$$

subject to $\tilde{\mathbf{p}}_{ij} = \hat{\mathbf{A}}_j\hat{\mathbf{P}}_i$ for all i, where $\bar{\mu}_{\text{res}}$ is the residual error of camera resectioning.

2.5 Bundle Adjustment

The final bundle adjustment step optimizes all cameras $\hat{\mathbf{A}}_j$ and all 3D feature points $\hat{\mathbf{P}}_i$ of the sequence by sparse Levenberg-Marquardt iteration, with

$$\bar{\nu}^2_{\text{res}} = \frac{1}{2\,J\,I}\sum_{j=1}^{J}\sum_{i=1}^{I} d(\tilde{\mathbf{p}}_{ij}\,, \hat{\mathbf{A}}_j\hat{\mathbf{P}}_i)^2_{\Sigma} \longrightarrow \min \tag{11}$$

where $\bar{\nu}_{\text{res}}$ is the residual error of bundle adjustment. The applied optimization strategy is Incremental Bundle Adjustment: First Eq. 11 is optimized for the keyframes and all intermediate views with the initial values determined in the previous step. Then the reconstructed 3D feature points are used for camera resectioning of the consecutive views. After each added view the 3D feature points are refined and extended and a new bundle adjustment is carried out until all cameras and all 3D feature points are optimized.

3 Keyframe Selection Algorithm

In this chapter the new approach for keyframe selection is presented. The approach attempts to find the keyframe pairing that minimize the estimation error of the following final bundle adjustment step. Bundle adjustment with iterative minimization is heavily reliant on good initial values for 3D feature points $\hat{\mathbf{P}}_i$ and camera matrices $\hat{\mathbf{A}}_j$. Thus, a keyframe selection criterion that judges the quality of these initial values is needed. Therefore, it should be taken into account, that initial camera matrices $\hat{\mathbf{A}}_j$ are estimated from initial 3D feature points $\hat{\mathbf{P}}'_i$ as described in step 2.4, which rely on the choice of the keyframe paring.

Consequently, the first step of the approach is the estimation of the covariance matrix of initial 3D feature points $\hat{\mathbf{P}}'_i$ for each keyframe pairing where F-GRIC is smaller than H-GRIC. If F-GRIC $\geq$ H-GRIC the keyframe pairing candidate is rejected without evaluation of the covariance matrix. In the second step the estimated covariance matrix is applied to approximate a lower bound for the estimation error of 3D feature points $\hat{\mathbf{P}}_i$ and camera matrices $\hat{\mathbf{A}}_j$ after camera resectioning.

3.1 Covariance Matrix Estimation

For the estimation of covariance matrix of initial 3D feature points $\hat{\mathbf{P}}'_i$ a bundle adjustment between the two analyzed keyframes with camera matrices $\mathbf{A}_{k1}$ and $\mathbf{A}_{k2}$ is performed. As derived in [2, Chapter 4.2], the covariance matrix $\Sigma_{\hat{\mathbf{A}}_k\hat{\mathbf{P}}'_i}$ of both cameras and the 3D feature points is the first order equal to:

$$\Sigma_{\hat{\mathbf{A}}_k\hat{\mathbf{P}}'_i} = (\mathbf{J}^\top \Sigma_{\tilde{\mathbf{p}}}^{-1} \mathbf{J})^+ \tag{12}$$

where $\mathbf{J}$ is the Jacobian matrix calculated at the optimum for $\hat{\mathbf{A}}_{k1}$, $\hat{\mathbf{A}}_{k2}$, and $\hat{\mathbf{P}}'_i$, and where $\Sigma_{\tilde{\mathbf{p}}} = \text{diag}(\dots, \Sigma_{\tilde{\mathbf{p}}_{ij}}, \dots)$ and $(\dots)^+$ denotes the pseudo-inverse. It should be stressed, that the bundle adjustment between two views and the covariance matrix $\Sigma_{\hat{\mathbf{A}}_k\hat{\mathbf{P}}'_i}$ can be estimated with significant time savings using techniques for sparse matrices, because $\mathbf{J}^\top \Sigma_{\tilde{\mathbf{p}}}^{-1} \mathbf{J}$ has a sparse block structure (see [2, Appendix 4.6] for details). By extracting $\Sigma_{\hat{\mathbf{P}}'_i}$ from $\Sigma_{\hat{\mathbf{A}}_k\hat{\mathbf{P}}'_i}$ the total variance of the 3D feature points $\hat{\mathbf{P}}'_i$ is calculated by the trace of $\Sigma_{\hat{\mathbf{P}}'_i}$.

$$E\left[\sum_{i=1}^{I} d(\mathbf{P}_i, \hat{\mathbf{P}}'_i)^2\right] = \text{trace}(\Sigma_{\hat{\mathbf{P}}'_i}) \tag{13}$$

where $\mathbf{P}_i$ are the true 3D feature points, $E[\dots]$ denotes the expectation of the function, and $d(\dots)$ denotes the Euclidian distance.

3.2 Expectation of Estimation Error

Now, a lower bound for the mean estimation error $\bar{\mu}_{\text{est}}$ of 3D feature points $\hat{\mathbf{P}}_i$ and camera matrices $\hat{\mathbf{A}}_j$ after camera resectioning is derived (compare Eq. 10):

$$E[\bar{\mu}_{\text{est}}^2] = E\left[\frac{1}{3\,I}\sum_{i=1}^{I} d(\mathbf{P}_i, \hat{\mathbf{P}}_i)^2\right] \tag{14}$$

The measurements in the cost function defined by Eq. 10 are the 3D feature points $\tilde{\mathbf{P}}'_i$, that correspond to the estimated $\hat{\mathbf{P}}'_i$ in the previous step. The estimated 3D feature points $\hat{\mathbf{P}}'_i$ obey a Gaussian distribution defined on a space of dimension $3\,I$. In order to simplify calculation and reduce computational effort, let us assume $\tilde{\mathbf{P}}'_i$ obey an isotropic Gaussian distribution with $\Sigma_{\tilde{\mathbf{P}}'_i} = \bar{\sigma}^2\,\mathbf{I}$, where $\mathbf{I}$ is the Identity matrix and $\bar{\sigma}^2 = \text{trace}(\Sigma_{\hat{\mathbf{P}}'_i})/3\,I$, so that $\text{trace}(\Sigma_{\tilde{\mathbf{P}}'_i}) = \text{trace}(\Sigma_{\hat{\mathbf{P}}'_i})$.

The constraint $\tilde{\mathbf{p}}_{ij} = \hat{\mathbf{A}}_j\hat{\mathbf{P}}_i$ of Eq. 10 enforces that $\hat{\mathbf{P}}_i$ can be located only on the line of sight defined by $\tilde{\mathbf{p}}_{ij}$ and $\hat{\mathbf{A}}_j$. Thus, the degrees of freedom for every estimated 3D feature point $\hat{\mathbf{P}}_i$ reduces from three to one. This means, that within the measurement space of dimension $3\,I$ a surface of dimension $I+A$ exists, where A is the number of essential parameters of one camera $\mathbf{A}_j$. On this surface all possible solutions for $\hat{\mathbf{P}}_i$ and $\hat{\mathbf{A}}_j$ are located. In the Levenberg-Marquardt algorithm this solution surface is approximated by a tangent surface, which has the same dimension. Because the Gaussian distribution in the measurement space is assumed isotropic and thus invariant to rotation, the projection on the tangent surface is equal to the projection on the first $(I+A)$ coordinate axes of the measurement space (see [2, Chapter 4.1.3] for details). Thus, on the tangent surface one gets an isotropic Gaussian distribution with total variance $(I+A)\,\bar{\sigma}^2$. This results in a expected estimation error

$$E[\bar{\mu}_{\text{est}}^2] = \frac{1}{3\,I}(I+A)\,\bar{\sigma}^2 = \frac{I+A}{(3\,I)^2}\,\text{trace}(\Sigma_{\hat{\mathbf{P}}'_i}) = S_c \tag{15}$$

If F-GRIC < H-GRIC, the score S_c is the new criterion to evaluate a keyframe pairing candidate, where the pairing with a lower S_c indicates a better choice. This is conceivable as a small $\text{trace}(\Sigma_{\hat{\mathbf{P}}'_i})/(3\,I)$ corresponds to a small variances of the estimated initial 3D feature points and the quotient $(I+A)/(3\,I)$ becomes smaller for a larger number of 3D feature points I.

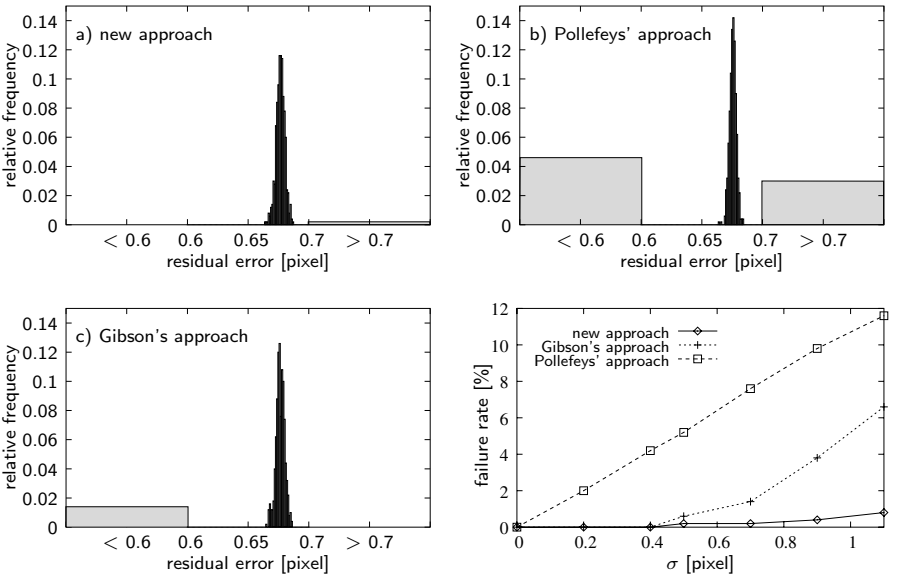

Fig. 3. Relative frequency of residual error $\bar{\nu}_{\text{res}}$ (see Eq. 11) for 500 trials. The errors in the positions of feature points obey a Gaussian distribution with standard deviation $\sigma = 0.7$ pixel. The left most and right most bins capture all residual errors that are smaller than 0.6 pixel and larger than 0.7 pixel. Trials with these residual errors are counted as failures of the bundle adjustment: a) new approach, b) Pollefeys' approach, and c) Gibson's approach. d) Failure rate over standard deviation σ for the different approaches.

4 Results

4.1 Synthetic Data Experiments

This chapter compares the new criterion with the state of the art approaches by Pollefeys and Gibson. Therefore, synthetic data experiments are carried out in the defined reference framework of Chapter 2, whereby only the keyframe selection criterion is changed.

500 synthetic test sequences with random camera motion and random scenes are generated. Each test sequence consists of 40 views. The camera translation between two successive views is uniformly distributed between 0 and 80 mm in all three coordinate directions and the camera rotation around the coordinate axes is uniformly distributed between 0 and 1 degree. 50% of the generated camera motions between two images are purely rotational. The camera has an image size of 720×576 pixel $= 7.68 \times 5.76$ mm and a mean focal length of 10.74 mm. The random scenes consist of 4000 3D feature points, which have a distance from the camera between 800 and 3200 mm. Approximately 35 to 40 of these 3D feature points are used in the final bundle adjustment step. The errors in the positions of

the generated 2D image feature points obey an isotropic Gaussian distribution. 20% of the generated correspondences between 2D feature points are outliers.

Fig. 3a-c opposes the relative frequency of residual error $\bar{\nu}_{\text{res}}$ after bundle adjustment for the three approaches. 500 trials are performed and the errors in the feature point positions have a standard deviation $\sigma = 0.7$ pixel. The expectation value of the residual error

$$E[\bar{\nu}_{\text{res}}] = \sigma \sqrt{1 - \frac{A\,J + 3\,I}{2\,J\,I}} \tag{16}$$

is approximately 0.65 pixel. Therefore, if a residual error is smaller than 0.6 pixel or larger than 0.7 pixel, the bundle adjustment has not converged to the correct minimum and these trials are counted as failures. It is obvious, that the new approach improves the convergence probability of the bundle adjustment significantly because failures occur less frequently. In Fig. 3d the failure rates over standard deviation σ for the different approaches are plotted. Especially, if the standard deviation is large, the new approach shows its improved robustness.

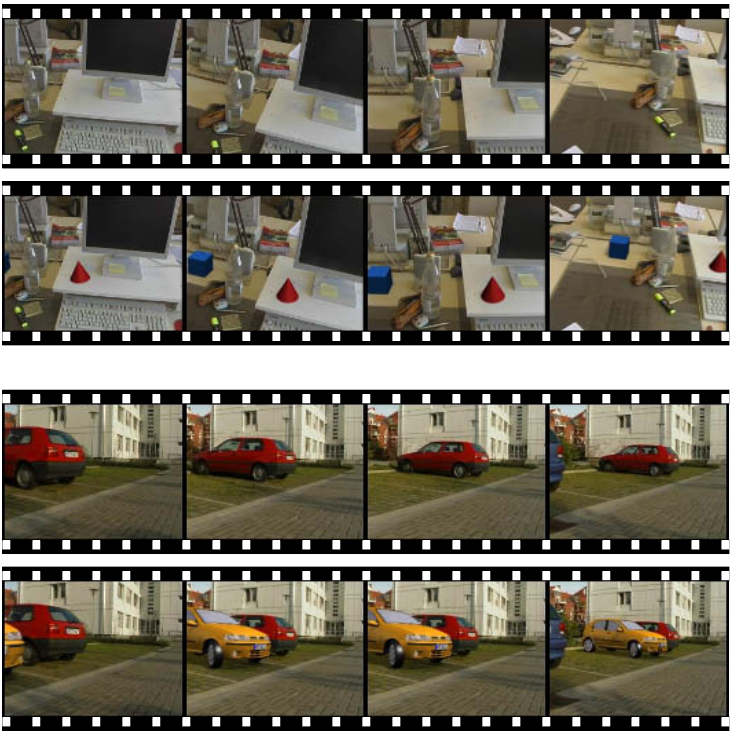

Fig. 4. Examples of augmented image sequences.

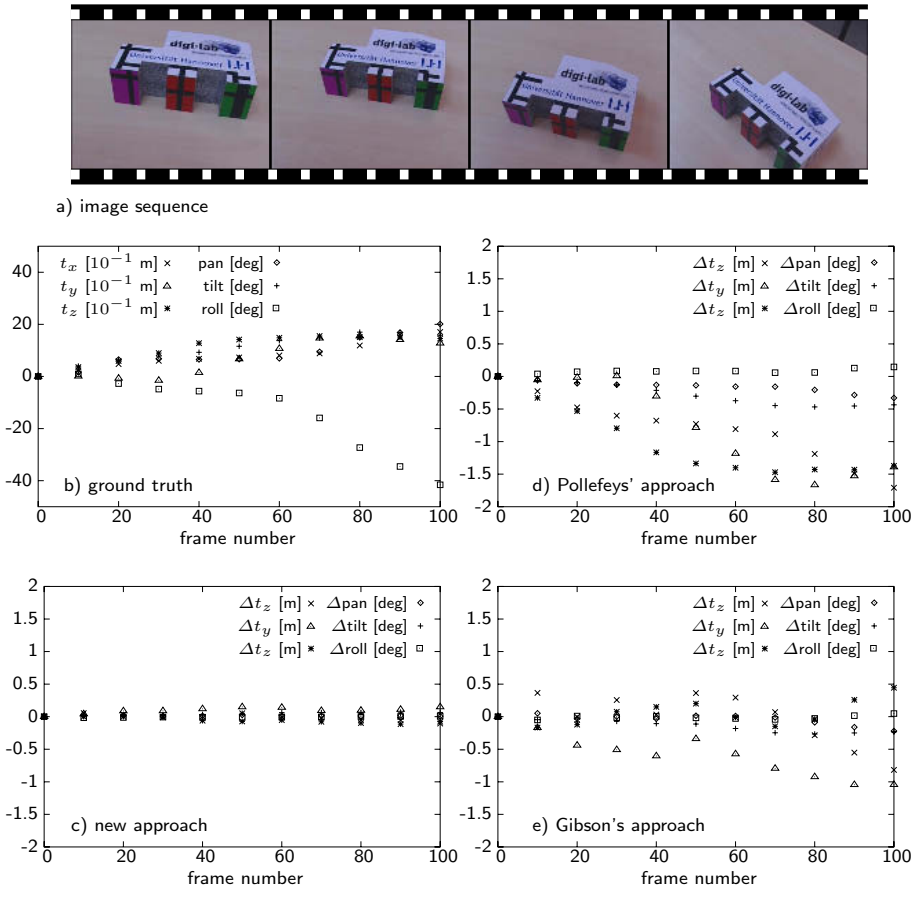

Fig. 5. a) Image sequence showing a test object with an exactly known structure. b) Ground truth extrinsic camera parameter generated with Tsai's camera calibration method. Difference between ground truth and estimation c) new approach, d) Pollefeys' approach e) Gibson's approach.

4.2 Natural Image Sequences

The new keyframe selection criterion has also demonstrated to work well on natural image sequences taken by a moving camera. Results of augmented image sequences that have been calibrated using the technique described in this paper are illustrated in Fig. 4. Videos of these augmented image sequences can be found on our website[1].

An empirical comparison of the new criterion with the state of the art approaches for natural image sequences is difficult because a large database containing natural image sequences with ground truth camera parameters would

[1] http://www.digilab.uni-hannover.de/results.html

be necessary. However, in order to illustrate the practical relevance of the new approach, a real-world example is given in Fig. 5. In Fig. 5a the evaluated image sequence is shown, which contains a test object with exactly known structure. Camera parameters are generated for every 10th view of this sequence with Tsai's camera calibration method [11], whereby the necessary 3D↦2D correspondences are manually edited. Generated camera parameters are exhibited in Fig. 5b and serve as ground truth. In Fig. 5c-e the differences between the ground truth and the estimated camera parameters after bundle adjustment for the different keyframe selection approaches are plotted. In this example the bundle adjustment does not converge to the right solution, if Pollefeys' or Gibson's approach is used. In contrast, the new keyframe selection approach gives satisfying results. It should be stressed, that this single selected example gives no information about the general performance of the three keyframe selection criteria. However, this example reveals, that failures due to wrong keyframe selection can be observed not only in synthetic data experiments but also occur in practice.

5 Conclusion

A new criterion for keyframe selection is proposed. It is derived from the estimation of the covariance matrix of initial 3D feature points and a lower bound for the estimation error of camera resectioning. While the state of the art approaches just avoid degenerated camera motion configurations, the new approach searches for the best possible keyframe pairing. This results in more accurate initial values for 3D feature points and camera parameters. Thus, iterative parameter minimization methods that are applied in the bundle adjustment, like the sparse Levenberg-Marquardt method, converge more frequently into the global minimum.

Furthermore, we see no reason against an adaptation of the new criterion into the three view framework of [5,6], where the trifocal tensor is used and initial 3D feature points are estimated from three views. Though a verification of this is left for future work.

References

1. Triggs, B., McLauchlan, P., Hartley, R.I., Fitzgibbon, A.: Bundle adjustment – A modern synthesis. In: Workshop on Vision Algorithms. Volume 1883 of Lecture Notes in Computer Science. (2000)
2. Hartley, R.I., Zisserman, A.: Multiple View Geometry. Cambridge University Press (2000)
3. Pollefeys, M., Gool, L.V., Vergauwen, M., Cornelis, K., Verbiest, F., Tops, J.: Video-to-3d. In: Proceedings of Photogrammetric Computer Vision 2002 (ISPRS Commission III Symposium), International Archive of Photogrammetry and Remote Sensing. Volume 34. (2002) 252–258
4. Gibson, S., Cook, J., Howard, T., Hubbold, R., Oram, D.: Accurate camera calibration for off-line, video-based augmented reality. In: IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR 2002), Darmstadt, Germany (2002)

5. Fitzgibbon, A., Zisserman, A.: Automatic camera recovery for closed or open image sequences. In: European Conference on Computer Vision. Volume 1406 of Lecture Notes in Computer Science. (1998) 311–326
6. Georgescu, B., Meer, P.: Balanced recovery of 3d structure and camera motion from uncalibrated image sequences. In: European Conference on Computer Vision. Volume 2351 of Lecture Notes in Computer Science. (2002) 294–308
7. Torr, P., Fitzgibbon, A., Zisserman, A.: The problem of degeneracy in structure and motion recovery from uncalibrated images. International Journal of Computer Vision **32** (1999) 27–44
8. Harris, C., Stephens, M.: A combined corner and edge detector. In: 4th Alvey Vision Conference. (1988) 147–151
9. Fischler, R.M.A., Bolles, C.: Random sample consensus: A paradigm for model fitting with application to image analysis and automated cartography. Communications of the ACM **24** (1981) 381–395
10. Torr, P.H.S., Zisserman, A.: MLESAC: A new robust estimator with application to estimating image geometry. Computer Vision and Image Understanding **78** (2000) 138–156
11. Tsai, R.Y.: A versatile camera calibration technique for high-accuracy 3-d machine vision metrology using off-the-shelf cameras and lenses. IEEE Transaction on Robotics and Automation **3** (1987) 323–344