Registration of Sub-Sequence and Multi-Camera Reconstructions for Camera Motion Estimation

Thorsten Thormählen, Nils Hasler, Michael Wand, Hans-Peter Seidel
Max Planck Institute for Computer Science
Saarbrücken, Germany
email: thormae@mpi-inf.mpg.de
www: http://www.mpi-inf.mpg.de

Abstract

This paper presents different application scenarios for which the registration of sub-sequence reconstructions or multi-camera reconstructions is essential for successful camera motion estimation and 3D reconstruction from video. The registration is achieved by merging unconnected feature point tracks between the reconstructions. One application is drift removal for sequential camera motion estimation of long sequences. The state-of-the-art in drift removal is to apply a RANSAC approach to find unconnected feature point tracks. In this paper an alternative spectral algorithm for pairwise matching of unconnected feature point tracks is used. It is then shown that the algorithms can be combined and applied to novel scenarios where independent camera motion estimations must be registered into a common global coordinate system. In the first scenario multiple moving cameras, which capture the same scene simultaneously, are registered. A second new scenario occurs in situations where the tracking of feature points during sequential camera motion estimation fails completely, e.g., due to large occluding objects in the foreground, and the unconnected tracks of the independent reconstructions must be merged. In the third scenario image sequences of the same scene, which are captured under different illuminations, are registered. Several experiments with challenging real video sequences demonstrate that the presented techniques work in practice.

Keywords: camera motion estimation, drift removal, multi-camera registration, structure-from-motion

First presented at the 5th European Conference on Visual Media Production (CVMP 2008) under the title: "Merging of Feature Tracks for Camera Motion Estimation from Video", extended and revised for JVRB

1 Introduction

Camera motion estimation and 3D reconstruction of rigid objects from video (Structure-from-Motion) is a well-established technique in computer vision, and fully automatic algorithms have been developed over the last decades [GCH+02, PGV+04, Tho06].

Most approaches determine corresponding feature points in consecutive frames. For video, the displacement of feature points between two frames is usually small and therefore, feature tracking methods, like the KLT-Tracker [ST94], produce less outliers and less broken tracks than feature matching methods (e.g. SIFT matching [Low04]), which are designed for wide baseline matching between images. Once corresponding feature points are found between consecutive frames, the parameters of a camera model can be estimated for every frame. Also, for each feature track, a corresponding 3D object point position is determined. If the errors in the position of the 2D feature points follow a Gaussian distribution, the Maximum Likelihood estimator for camera parameters and 3D object points is called bundle adjustment. Bundle adjustment minimizes the reprojection error of the 3D object points

into the camera images. The error is consequently distributed equally over the whole sequence.

During the sequential processing of the input frames, feature tracks are often broken. This happens because of occlusion, moving foreground objects, fast camera movements with large feature displacements or motion blur, image noise, or because the tracked 3D object point leaves the camera's field of view. Often, the same 3D position is found and tracked again later in the sequence but a new 3D object point is assigned to the new track. For the performance of bundle adjustment to be optimal, it is essential that those 3D object points are merged.

An application for which the merging of unconnected feature tracks has already been studied in the literature is drift removal [KBK06]. If 3D object points are not merged, errors accumulate and drift occurs during the sequential processing of the input frames. The presence of drift becomes particularly problematic in long sequences where the camera visits the same part of the scene multiple times. Often the 3D object point clouds representing those parts of the scene differ from each other by a significant 3D transformation. If the sequence is closed (i.e. camera position and orientation of the first and the last frame are the same), drift can be removed by enforcing this special constraint for the first and last camera view during the estimation process [FZ98]. A more general approach by Cornelis et al. [CVG04] removes drift by estimating the 3D transformation between the 3D object point clouds with a RANSAC [FB81] approach and afterwards merges those 3D object points that support the estimated transformation. One problem is that the number of possible 3D-3D matches is usually very large and the percentage of false matches (i.e., outliers) is high. In this case, the computational effort of the RANSAC method is excessive because many random samples have to be evaluated until a valid set of inliers is found. To remove the number of outliers, Cornelis et al. propose a proximity as well as a similarity constraint for possible 3D-3D matches. The Bhattacharyya distances of color histograms between the unconnected feature tracks is proposed to evaluate similarity.

In this paper novel scenarios are presented for which successful camera motion estimation and 3D reconstruction from video cannot be achieved without merging of unconnected feature point tracks. We first consider the case of drift removal and evaluate different similarity measures to find one that produces fewer false matches and thereby speeds up the RANSAC approach. As an alternative to the RANSAC approach, we describe how the spectral method by Leordeanu et al. [LH05] can be applied to the problem of merging unconnected feature tracks. The main contribution of this paper, however, is the extension to scenarios where multiple independent structure-from-motion reconstructions are registered into a common global coordinate system. It is shown that a modified algorithm for merging feature point tracks can be applied in novel scenarios where a scene is captured simultaneously by multiple moving cameras; in situations where the tracking of feature points completely fails (e.g., due to large occluding objects in the foreground); or in scenarios where multiple single camera recordings of the same scene need to be registered (e.g., captured at different points in time under possibly different illuminations).

The rest of the paper is organized as follows. In the next section, we describe our approach for finding unconnected feature track candidates and evaluating different similarity measurement scores. Sections 3 and 4 introduce the RANSAC method and spectral method for merging unconnected feature tracks, respectively. In Section 5, a modified version of the algorithm is presented that allows the registration of multiple independent structure-from-motion reconstructions. In Section 6, we report results of our experiments that show the performance of the suggested algorithms. The paper ends with concluding remarks in Section 7.

2 Finding unconnected feature tracks candidates

Let's assume we are given a video sequence with K images I_k, with $k = 1, \ldots, K$, and we have tracked J 3D object points $\mathbf{P}_j$, with $j = 1, \ldots, J$. That is, we have established J trajectories of 2D feature points $\mathbf{p}_{j,k}$ in a number of consecutive images with a feature point tracker (e.g. KLT-Tracker [ST94]).

After estimation of the 3×4 camera matrix $\mathtt{A}_k$ for each image I_k with a sequential structure-from-motion algorithm (e.g., [Tho06]), the reprojection of a 3D object point $\mathbf{P}_j$ in the image k with the camera matrix $\mathtt{A}_k$ should be located on the measured feature point $\mathbf{p}_{j,k}$. This can be seen in Fig. 1 and follows from the bundle adjustment objective function

$$\underset{\mathtt{A}_k, \mathbf{P}_j}{\arg\min} \quad \sum_{j=1}^{J} \sum_{k=1}^{K} d(\mathbf{p}_{j,k}\,,\, \mathtt{A}_k\, \mathbf{P}_j)^2 \qquad , \quad (1)$$

where $d(\dots)$ denotes the Euclidean distance and $\mathbf{p}_{j,k} = (x, y, 1)$ and $\mathbf{P}_j = (X, Y, Z, 1)^\top$ are homogeneous vectors. Optimizing this bundle adjustment equation is usually the final step in structure-from-motion algorithms.

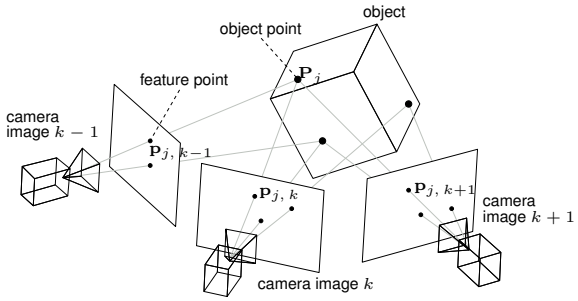


Figure 1: Result after structure-from-motion estimation. The projection of a 3D object point $\mathbf{P}_j$ in the camera image at time k gives the tracked 2D feature point $\mathbf{p}_{j,k}$.

If the covariance of the positional error of the measured 2D feature points is known, then it is possible to calculate covariance matrices for the error of the estimated camera parameters and the 3D object points after bundle adjustment (see [HZ00, Tho06] for details). Therefore, we assume that the 3×3 covariance matrix Σ_j of every estimated 3D object point $\mathbf{P}_j$ is available.

If a long sequence is processed, noise from the 2D feature points accumulates and drift occurs. As illustrated in Fig. 2, 3D object point $\mathbf{P}_j$ and the earlier reconstructed object point $\mathbf{P}_i$ should be at the same physical position. If sequential bundle adjustment is employed, however, due to drift this is not necessarily the case. A real-world example of severe drift is shown in Fig. 3.

In order to remove drift, we need to merge 3D object points $\mathbf{P}_i$ and $\mathbf{P}_j$, which is a pairwise matching problem. In the first step of the matching procedure, merging candidates are discarded if they are not in the vicinity of each other. This proximity constraint is evaluated in the image plane. That is, two object points are merging candidates if

$$d(\mathtt{A}_k^{(j)}\,\mathbf{P}_j, \mathtt{A}_k^{(j)}\,\mathbf{P}_i) < \tau_1 \quad , \tag{2}$$

for all camera images $\mathtt{A}_k^{(j)}$ where $\mathbf{P}_j$ is tracked. Typically, we choose the threshold τ_1 in the range of 20 to 100 pixels, dependent on the amount of expected drift.

The second constraint that two object points need to fulfill is the similarity constraint. The similarity constraint evaluates whether the color intensity in a window around their tracked position is similar in those images where the object points were found. If $S(\dots)$ is a similarity measurement score, then two object points are merging candidates if

$$S(\mathtt{A}_k^{(j)}\,\mathbf{P}_j, \mathtt{A}_k^{(i)}\,\mathbf{P}_i) < \tau_2 \quad , \tag{3}$$

for all camera images $\mathtt{A}_k^{(j)}$ where $\mathbf{P}_j$ is tracked and all camera images $\mathtt{A}_k^{(i)}$ where $\mathbf{P}_i$ is tracked.

In order to find an appropriate similarity measure we evaluate in the following how different approaches in the literature suit our problem. Therefore, we generated a ground truth data set for the drift sequence presented in Fig. 3 by selecting two images out of this sequence and labeled unconnected feature tracks by hand, see Fig. 4. In total we found 30 correct matches between those two images.

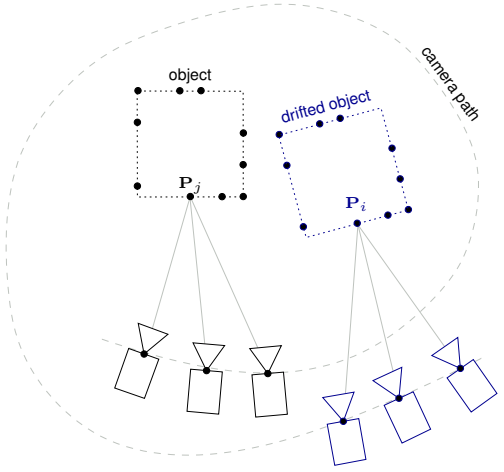


Figure 2: After processing a long sequence 3D object points $\mathbf{P}_j$ and $\mathbf{P}_i$ are not at the same position because of drift.

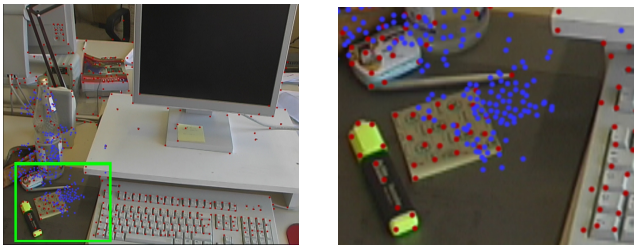


Figure 3: A real-world example of drift. After the camera revisits the same part of the scene the 3D object points of the current image (red) differ strongly from the 3D object points generated earlier in the sequence (blue). The right image shows a detail magnification from the left image. The squared shape of the yellow paper is clearly visible in the shifted blue point cloud.

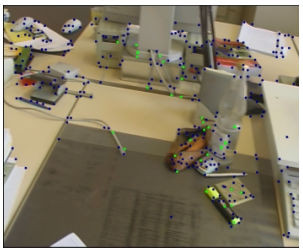



Figure 4: A ground truth data set with 30 hand-labeled unconnected feature tracks, marked in green.

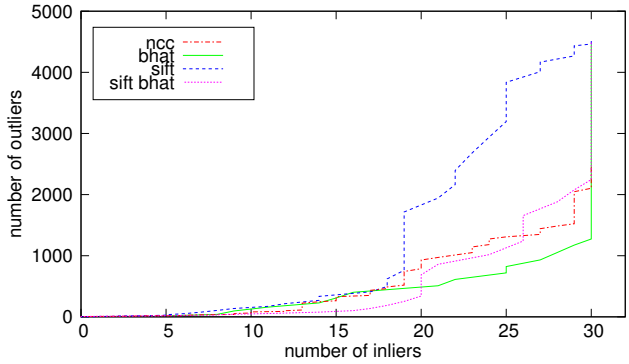


Figure 5: Number of inliers over number of outliers of unconnected feature track candidates for different values of the similarity threshold τ_2. Results for four different similarity measurements are shown: normalized cross correlation (ncc), Bhattacharyya distance of color histograms (bhat), scale invariant feature transform matching (sift), and a combination of SIFT matching and Bhattacharyya distance (sift bhat).

Four different similarity measures were evaluated: normalized cross correlation, Bhattacharyya distance of color histograms, scale invariant feature transform (SIFT) matching, and a combination of SIFT matching and Bhattacharyya distance. To generate the combined SIFT and Bhattacharyya distance measure, the output of both similarity scores is mapped to the range between 0.0 and 1.0 and the mean of both is used as the combined score. We then changed the threshold τ_2 in small steps from 0.0 to 1.0 and plotted the number of inliers over the number of outliers. As shown in Fig. 5 the resulting inlier to outlier ratio is generally very low for this difficult example. If the threshold is chosen quite strict (e.g. so that only 50% of the inliers pass the test), the inlier/outlier ratio is acceptable, especially for the combined SIFT and Bhattacharyya distance score. This combined similarity measure also performed best for the other examples presented in Section 6.

At last, we define a uniqueness constraint. Two object points, $\mathbf{P}_i$ and $\mathbf{P}_j$, are merging candidates if

$$\frac{S(\mathtt{A}_k^{(j)}\,\mathbf{P}_j, \mathtt{A}_k^{(i)}\,\mathbf{P}_i)}{S_{\text{close}}} < \tau_3 \quad , \tag{4}$$

where S_{close} is the best measurement score either $\mathbf{P}_i$ or $\mathbf{P}_j$ achieves with any other 3D object point. This constraint is especially important for scenes that contain repetitive structures. If this constraint is not checked, whole groups of 3D object points may get merged with the wrong repeated structure in the neighbourhood.

3 Merging of unconnected feature tracks with RANSAC

All 3D object point pairs $(\mathbf{P}_i, \mathbf{P}_j)$, that pass all three tests in Eqs. (2), (3), and (4) are candidates for unconnected feature tracks. All of these candidates are added to the set M_{all}. This set usually contains a large number of wrongly assigned pairs (outliers). We now need to separate all candidates within set M_{all} into a set of inliers M_{in} and a set of outliers M_{out}.

As can be seen in Figs. 2 and 3, the drifted object points can be transformed to the object points in the current frame by a common 3D transformation $\mathtt{H}$,

$$w\,\mathbf{P}_j = \mathtt{H}\,\mathbf{P}_i \qquad \forall(\mathbf{P}_i, \mathbf{P}_j) \in M_{\text{in}} \tag{5}$$

where w is an unknown scale factor. The 4×4 matrix $\mathtt{H}$ has 16 elements. Because of the unknown scale factor for homogeneous entities we may fix one element of $\mathtt{H}$ to 1 and, therefore, $\mathtt{H}$ has 15 unknown parameters. These 15 parameters can be estimated from a minimal set of 5 object point pairs, since every pair contributes 3 linear equations [HZ00].

To determine $\mathtt{H}$, we use the RANSAC approach [FB81]. Five object point pairs are randomly sampled out of the set M_{all} and an $\mathtt{H}$ is estimated. Then, the support of $\mathtt{H}$ is measured by evaluating

$$\epsilon = \sum_{M_{\text{all}}} \epsilon_{i,j} \quad \text{with} \tag{6}$$

$$\epsilon_{i,j} = \begin{cases} d(\mathbf{P}_j, \mathtt{H}\,\mathbf{P}_i)^2_\Sigma & \text{if} \quad d(\mathbf{P}_j, \mathtt{H}\,\mathbf{P}_i)^2_\Sigma < \tau_4 \\ \tau_4 & \text{else} \end{cases} ,$$

where $d(\dots)_\Sigma$ denotes the Mahalanobis distance. The Mahalanobis distance can be calculated if the covariance matrices of $\mathbf{P}_i$ and $\mathbf{P}_j$ are available, which is the case here. After randomly sampling 5 pairs from M_{all}

for a sufficient number of times (i.e., until we can assume that we have at least once estimated H with 5 inliers), we choose the H with the smallest ϵ. All object point pairs with $d(\mathbf{P}_j,\, \mathtt{H}\,\mathbf{P}_i)^2_\Sigma < \tau_4$ are added to the set of inliers M_{in} and all others to the set of outliers M_{out}. All 3D object point pairs in M_{in} are considered unconnected feature tracks and are merged. Afterwards, the drift can be removed by a bundle adjustment over the whole sequence.

Please note that for simplicity and speed, we assume in our implementation that the covariance matrix Σ_i of $\mathbf{P}_i$ is not changed by a multiplication with H, which is only an approximation, but works well in practice. The Mahalanobis distance $d(\mathbf{P}_j,\, \mathtt{H}\,\mathbf{P}_i)^2_\Sigma$ obeys a χ^2-distribution of degree 3. In our experiments, we choose the threshold $\tau_4 = 11.34$, which is the 99% quantile for the χ^2-distribution of degree 3. This means that we will reject an inlier in only 1% of cases, if H is estimated correctly.

In practice, it is often the case that the inlier/outlier ratio is small, as was shown in the previous section. It is a known problem of the RANSAC algorithm,that for small inlier/outlier ratios, a large number of random trials have to be performed before it can be assumed that the RANSAC algorithm found the correct solution. In these cases, the computational effort of the RANSAC method can become excessive.

The computational effort can be reduced, if we assume that H is a similarity transformation with only 7 parameters, 3 for rotation, 3 for translation, and 1 for scale. Then, H can be estimated from only 3 inlier object point pairs, which reduces the number of required trials. The assumption that H is a similarity transformation is valid, if the camera matrices $\mathtt{A}_k$ and 3D object points $\mathbf{P}_j$ are reconstructed in the metric and not in the projective space by the structure-from-motion algorithm. In our experiments, we found that this is a valid approximation after auto-calibration [TBM06], even if drift is present in the reconstruction.

4 Spectral method for merging of unconnected tracks

The spectral method by Leordeanu et al. [LH05] is an alternative to the previously described RANSAC algorithm and is adapted to the problem of merging unconnected tracks in case of drift in the following. However, it can only be used in a metric and not in a projective space. Thus, a auto-calibration step must be applied beforehand.

All 3D object point pairs $(\mathbf{P}_i, \mathbf{P}_j)$ that pass the three tests in Eqs. (2), (3), and (4) are candidates for unconnected feature tracks and are added to the set M_{all}. If N is the total number of elements in M_{all}, then we denote the n-th element as $(\mathbf{P}_i^{(n)}, \mathbf{P}_j^{(n)})$, with $n = 1, \ldots, N$.

First, a symmetric non-negative $N \times N$ matrix L is generated. The n-th diagonal element of L is set to the similarity score $S(\mathbf{P}_i^{(n)}, \mathbf{P}_j^{(n)})$ that was determined with Eq. (3) for the n-th element. The off-diagonal elements $\mathtt{L}(n, m)$, with $m = 1, \ldots, N$, are set to the score $D((\mathbf{P}_i^{(n)}, \mathbf{P}_j^{(n)}), (\mathbf{P}_i^{(m)}, \mathbf{P}_j^{(m)}))$, which measures how well the n-th and m-th pair in M_{all} support the same Euclidean transformation:

$$D((\mathbf{P}_i^{(n)},\mathbf{P}_j^{(n)}),(\mathbf{P}_i^{(m)},\mathbf{P}_j^{(m)})) = \begin{cases} e^{-\frac{\Delta d^2}{2\sigma^2}} & \text{if } \Delta d < \tau_5 \\ 0 & \text{else} \end{cases}$$

with

$$\Delta d \quad = d(\mathbf{P}_i^{(n)}, \mathbf{P}_i^{(m)}) - d(\mathbf{P}_j^{(n)}, \mathbf{P}_j^{(m)}) \quad . \qquad (7)$$

If $D(\ldots)$ is close to 1, then the n-th and m-th pair support the same Euclidean transformation.

We then determine the principal eigenvector of the matrix L with the Power method [Saa92], which is usually very fast because it converges after a few iterations.

The value of the eigenvector at position n can be interpreted as the confidence that the n-th pair is a valid match and can be merged (see [LH05] for details). Therefore, first the pair with the highest confidence is moved from M_{all} to the inlier set M_{in}. All other pairs in M_{all} that are in *conflict* with the highest confidence pair are moved to the outlier set M_{out}. Pairs are in *conflict*, if they share either $\mathbf{P}_i$ or $\mathbf{P}_j$ with the highest confidence pair or the corresponding $D(\ldots)$ to the highest confidence pair is zero. Now the remaining second highest confidence pair from M_{all} is processed the same way, and so on, until M_{all} is empty.

For drift removal, all 3D object point pairs in M_{in} are considered unconnected feature tracks and are merged. Finally, a bundle adjustment is executed over the whole sequence.

5 Registration of independent structure-from-motion reconstructions

The algorithms which were described in the previous two sections are not limited to drift detection and removal. With slight modifications, these algorithms can

find the correct transformation between two or more independent structure-from-motion reconstructions.

The first application we want to address here is the registration of multiple moving cameras that capture the same scene simultaneously.

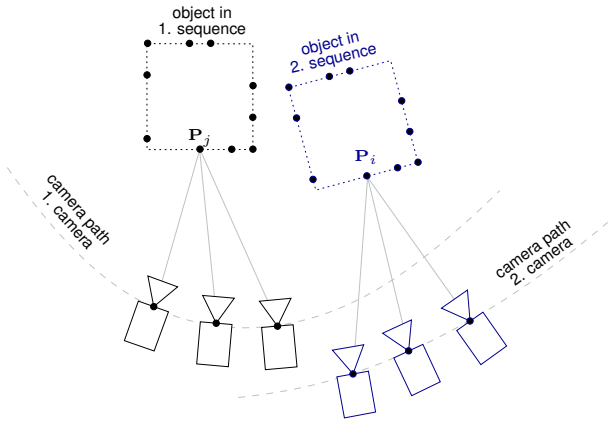


Figure 6: Registration of multiple moving cameras that capture the same scene simultaneously. Drift removal and registration of multiple cameras are similar problems, which becomes obvious, when this figure is compared with Fig. 2.

For each camera, the video sequences are processed independently with a sequential structure-from-motion algorithm. As shown in Fig. 6, the resulting reconstruction of camera motion and 3D object points for each camera is determined only up to a similarity transformation with 7 degrees of freedom, 3 for rotation, 3 for translation, and 1 for scale. In the following, we present an algorithm to register these independent reconstructions into a common global coordinate system.

Registration algorithm:

1. **Run the RANSAC algorithm** of Section 3, where the proximity constraint should not be used. This means all 3D object point pairs $(\mathbf{P}_i, \mathbf{P}_j)$ that pass the similarity and uniqueness constraint in Eq. (3) and Eq. (4) are merging candidates.

2. **Transform** all object points P_i and camera matrices of the second image sequence with the transformation `H`.

3. **Enforce the proximity constraint**. All merging candidates have to pass the three tests in Eqs. (2), (3), and (4). The uniqueness constraint now filters out less candidates because the search for candidates is guided by the proximity constraint.

4. **Run the spectral method** of Section 4 (or, alternatively, the RANSAC algorithm)

5. **Merge** the inlier 3D object point pairs.

6. **Bundle adjust** the merged sequences.

Optionally, after step 4, reduce the proximity threshold τ_1 and go to step 2. (This re-selection of candidates with reduced proximity threshold can also help in the case of drift removal if only a few candidates were found in the first run.)

The same registration algorithm can also be applied if the motion estimation of a single camera cannot be continued and has to restart. This is often the case if no feature points can be tracked because of large occlusions in the foreground or because of extreme motion blur. In these cases, the sequential structure-from-motion algorithm automatically stops and starts again with a new independent reconstruction. Afterwards, the independent reconstruction can be registered with the registration algorithm described above.

Another application example of this algorithm are scenarios in which the same scene is captured at different points of time and a registration of the independent reconstructions is necessary. This is still possible, if the illumination changes because the feature matching is in certain bounds insensitive to linear changes of the image values.

6 Results

In this section, we present three real-world examples of structure-from-motion estimation where merging of unconnected feature tracks is necessary. All examples are recorded with off-the-shelf consumer HDV cameras at a resolution of 1440×1080 pixels and a frame rate of 25 Hz.

The examples are also shown in the video provided with this paper.

6.1 Example 1: Drift removal

In this example, the camera performs two complete loops around an advertising column. A total of 2500 frames were recorded. In Fig. 7, the input sequence and the resulting camera path after sequential structure-from-motion are shown. The white dots are the reconstructed 3D object points. The green dots are the 3D object points of the first frame. The first

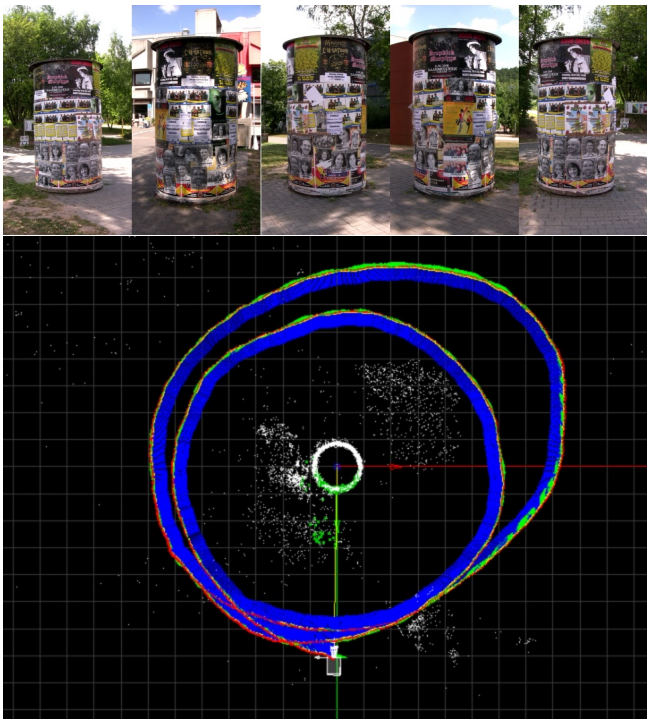
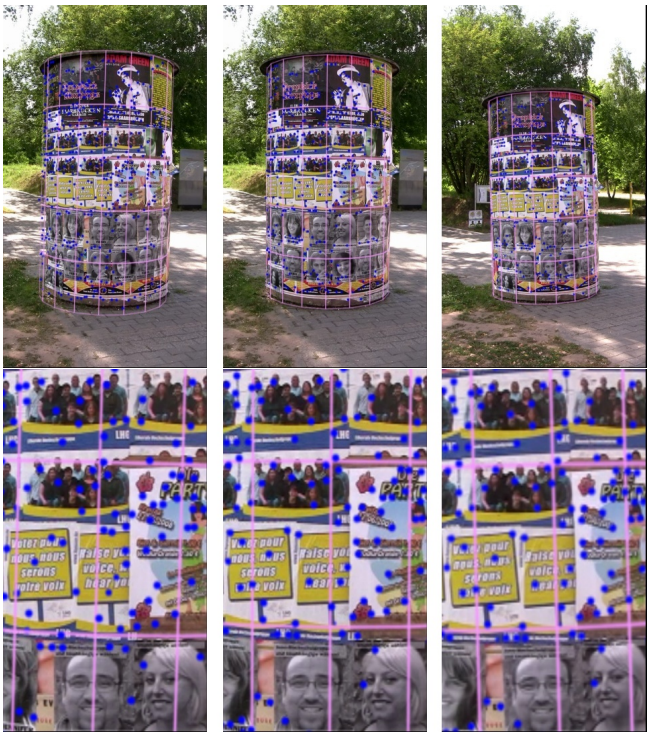


Figure 7: Five images out of the input video of example 1 and a top view on the resulting camera path after sequential structure-from-motion.



a) b) c)

Figure 8: a) result without drift removal after two loops around the column, b) result with drift removal, c) first frame of the sequence as drift-free reference. Only the 3D object points that were generated in the first frame are shown as blue dots. The second row shows detail magnifications of the above images.

frame is also marked with a small camera icon. In total, 34214 object points were reconstructed.

In Fig. 8, the results are compared with and without drift removal. Fig. 8c shows the first frame of the video. Only those 3D object points are shown that were generated out of the 2D feature tracks that started in the first frame. These 3D object points are displayed as blue dots, and they project perfectly onto the detected corners in the first frame. This first frame can therefore be used as a reference. If the estimation contains no drift, these 3D object points of the first frame must project to the same corners even after two complete loops around the advertising column. In Fig. 8a, the result is shown after two complete loops without drift removal. The 3D object points do not project to their original position. To visualize the amount of drift, a 3D model of a column (pink wireframe in Fig. 8) was fitted to the 3D object points of the first frame. The displacement is clearly visible in Fig. 8a. Fig. 8b shows the result after the drift is removed with the spectral method of Section 4. The 3D object points project exactly to the correct positions. The RANSAC method computes very similar results, which are not shown here for this reason. If, however, we compare the computation time of both methods the spectral method needs 459.28 ms on average compared with 613.26 ms for the RANSAC method - an increase in speed of 25 percent. In total, we applied each method 76 times with appropriate frame offsets to find the unconnected feature tracks for this sequence. In total, 8055 of 34214 object points were merged.

6.2 Example 2: Registration of multiple cameras

This example, which is shown in Fig. 9, was recorded simultaneously with four moving cameras. Each of the four video sequences has a length of 188 frames. The camera motion and the 3D object point cloud of every sequence are estimated independently for each sequence. Afterwards, the independent reconstructions are merged with the registration algorithm of Section 5. The sequence is challenging because of the large number of moving objects and because of the repetitive flagging and repeated windows. However, all four independent reconstructions were registered

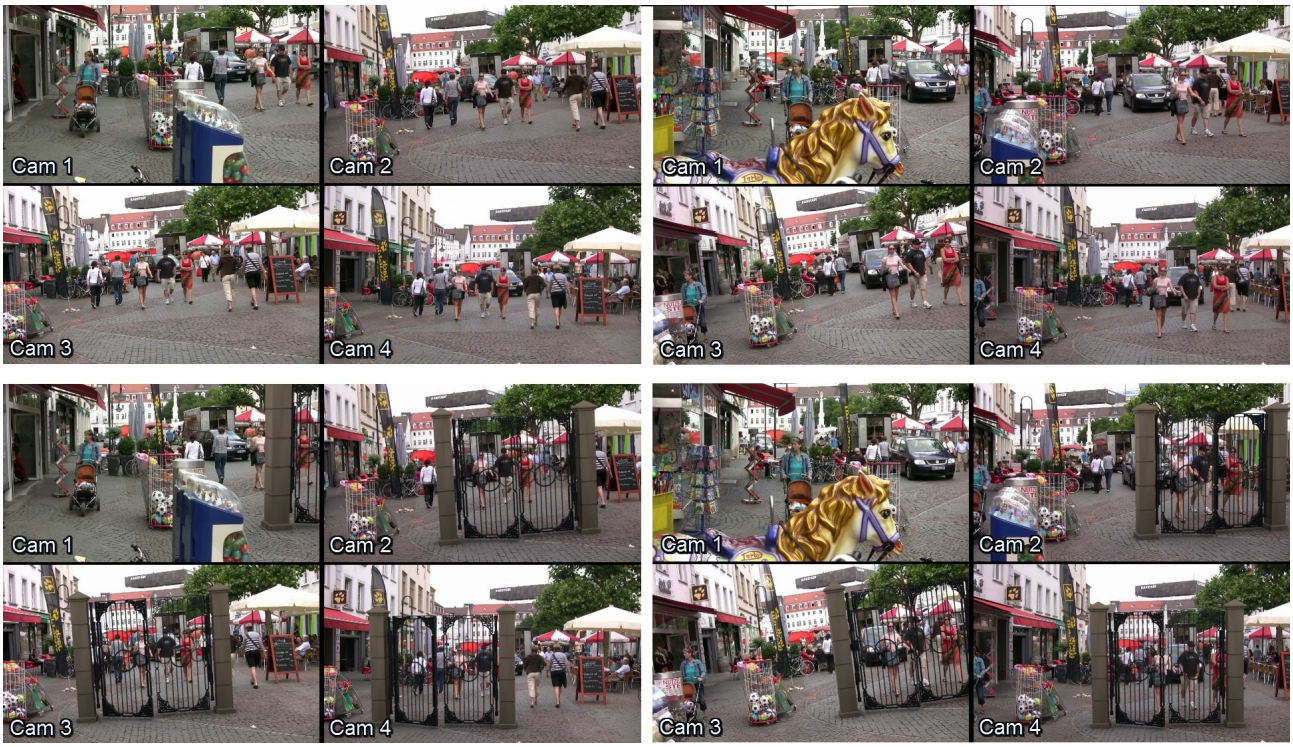


Figure 9: In example 2, four moving cameras simultaneously capture a street scene. Top row: Two sample images out of the sequence for each of the four cameras. Bottom row: The scene is augmented with a 3D model of a gate.

successfully, as can be verified in Fig. 10. In order to test the registration, the scene was augmented with a 3D model of a gate. This virtual gate stays perfectly at its assigned position (see video or Fig. 9).

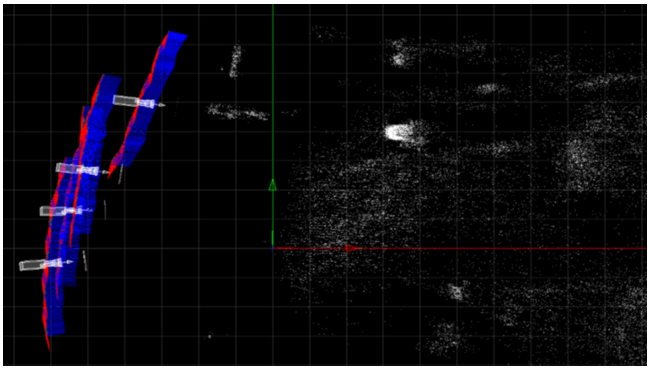


Figure 10: Top view on the resulting camera path of four moving cameras. The four independent reconstructions were registered into a common global coordinate system.

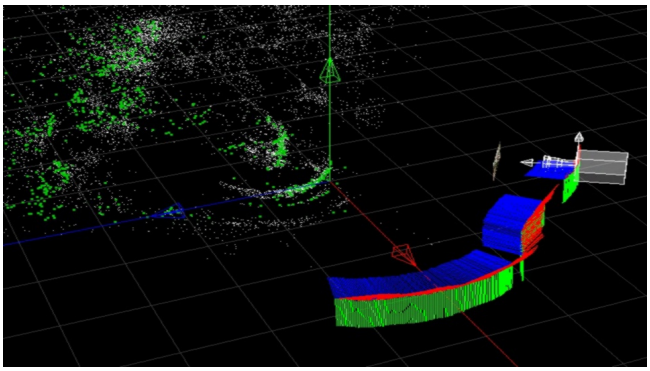


Figure 12: Camera path and 3D object point cloud after registration of three independent structure-from-motion estimations in example 3

6.3 Example 3: Recovery after large occlusions

This challenging example of a market scene has a total length of 319 frames. The sequential structure-from-motion estimation automatically restarts twice because too few feature point tracks were available, due to large occluding objects in the foreground (Fig. 11). After each restart, the structure-from-motion algo-

Figure 11: Top row: Five images out of the input video of example 3. The motion estimation has to restart two times because of large occlusions in the foreground. These large occlusions are visible in the second and fourth image of the top row. Bottom row: Augmented video sequence.

rithm produces an independent reconstruction of the camera motion and 3D object points for that part of the sequence. The registration algorithm of Section 5 was applied, and the three independent reconstructions were successfully registered into a common global coordinate system, as can be seen in Fig. 12. In total, 243 of 13007 object points were merged. To test the results, five virtual 3D objects were rendered into the sequence. As can be verified in Fig. 11 and in the video provided with this paper, these virtual object do not show any visual misalignments or jitter.

6.4 Example 4: Registration of multiple reconstructions of the same differently illuminated scene

In this example three image sequences of the same scene are recorded at different points of time in cloudy, sunny, and dark environments. Each recorded sequence has a length of 438 frames. The sequences are processed independently by the sequential structure-from-motion approach. The resulting reconstructions are merged using the algorithm of section 5. The process is fully automatic for the registration of the sunny and cloudy reconstruction. However, for the extremely challenging dark sequence some feature tracks needed to be edited manually. This has two reasons: Firstly, because of the low light, less high image gradients are available and, consequently, less feature points are detected in large parts of the scene. Secondly, the automatic feature tracking as well as feature matching between the reconstructions are affected by the low signal-to-noise-ratio.

The registration of differently illuminated image sequences allows new applications, like, interactive switching between the different illuminations during playback. A frame selection algorithm is employed that uses the obtained registration to find the best frame, for which the camera view is kept as constant as possible during the switch. For all feature points $\mathbf{p}_{j,k=s}$ of the switching frame s the corresponding 3D object points $\mathbf{P}_{j(s)}$ are considered. These 3D object points are backprojected into each frame of the target sequence. Then for each frame k the following cost is calculated:

$$C = \sum C_j \tag{8}$$

with

$$C_j = \min(\, d(\mathbf{p}_{j,\,s}\,,\, \mathtt{A}_k\, \mathbf{P}_{j(s)})\,,\, d_{\max}) \quad . \tag{9}$$

A threshhold of $d_{\max} = 200$ pixels was used in our experiments. The frame k with the lowest score C is chosen and is played next during the realtime playback. Because of its simplicity this measure can be calculated very quickly, which allows the user to switch between the illuminations at interactive speed. Examples of illumination transitions are shown in Fig. 13 and in the video.

7 Conclusion

In this paper, we presented four examples where merging of unconnected feature tracks was necessary to achieve robust camera motion estimation and 3D reconstruction from video: drift removal, registration of multiple moving cameras, recovery of camera motion after large occlusions, and registration of multiple reconstructions of the same scene under different illumination. For each of these scenarios, we showed results of automatic camera motion estimation for challenging real-world video sequences with repetitive structures and moving objects.

A key ingredient for the successful processing of these videos is our choice of the similarity measure,

Figure 13: After the registration of multiple reconstructions of the same differently illuminated scene, a user can interactively switch (black arrows) between the different illuminations, whereby the camera view is kept as constant as possible during the switch. The automatically selected transitions are shown.

as well as the application of a uniqueness and proximity constraint to find fewer false candidates for unconnected feature tracks. Furthermore, we have adopted the spectral method by Leordeanu et al. [LH05] to the problem of merging unconnected feature tracks.

We believe that this unified technique for merging unconnected feature tracks in different scenarios is an important step towards fully automatic camera motion estimation in difficult situations.

Acknowledgements

This work has been partially funded by the Max Planck Center for Visual Computing and Communication (BMBF-FKZ01IMC01).

References

[CVG04] Kurt Cornelis, Frank Verbiest, and Luc Van Gool, *Drift Detection and Removal for Sequential Structure from Motion Algorithms*, IEEE Trans. Pattern Anal. Mach. Intell. **26** (2004), no. 10, 1249–1259, ISSN 0162-8828.

[FB81] Martin A. Fischler and Robert C. Bolles, *Random Sample Consensus: A Paradigm for Model Fitting with Application to Image Analysis and Automated Cartography*, Communications of the ACM **24** (1981), no. 6, 381–395, ISSN 0001-0782.

[FZ98] Andrew W. Fitzgibbon and Andrew Zisserman, *Automatic Camera Recovery for Closed or Open Image Sequences*, ECCV '98: Proceedings of the 5th European Conference on Computer Vision-Volume I (London, UK), Springer-Verlag, 1998, pp. 311–326, ISBN 3-540-64569-1.

[GCH+02] Simon Gibson, Jon Cook, Toby Howard, Roger Hubbold, and Dan Oram, *Accurate Camera Calibration for Off-line, Video-Based Augmented Reality*, IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR 2002) (Darmstadt, Germany), September 2002, p. 37, ISBN 0-7695-1781-1.

[HZ00] Richard I. Hartley and Andrew Zisserman, *Multiple View Geometry in Computer Vision*, 1 ed., Cambridge University Press, 2000, ISBN 0-521-62304-9.

[KBK06] Kevin Koeser, Bogumil Bartczak, and Reinhard Koch, *Drift-free Pose Estimation with Hemispherical Cameras*, Proceedings of Conference on Visual Media Production (CVMP 2006) (London), 2006, pp. 20–28, ISBN 978-0-86341-729-0.

[LH05] Marius Leordeanu and Martial Hebert, *A Spectral Technique for Correspondence Problems Using Pairwise Constraints*, Proceedings of the Tenth IEEE International Conference on Computer Vision (Washington, DC, USA), IEEE Computer Society, 2005, pp. 1482–1489, ISBN 0-7695-2334-X-02.

[Low04] David G. Lowe, *Distinctive Image Features from Scale-Invariant Keypoints*, Int. J. Comput. Vision **60** (2004), no. 2, 91–110, ISSN 0920-5691.

[PGV+04] Marc Pollefeys, Luc Van Gool, Maarten Vergauwen, Frank Verbiest, Kurt Cornelis, Jan Tops, and Reinhard Koch, *Visual modeling with a hand-held camera*, International Journal of Computer Vision

IJCV **59** (2004), no. 3, 207–232, ISSN 0920-5691.

[Saa92] Yousef Saad, *Numerical Methods for Large Eigenvalue Problems*, Manchester University Press, 1992, ISBN 0719033861.

[ST94] Jianbo Shi and Carlo Tomasi, *Good Features to Track*, CVPR, 1994, pp. 593–600, ISBN 0-8186-5825-8.

[TBM06] Thorsten Thormählen, Hellward Broszio, and Patrick Mikulastik, *Robust Linear Auto-calibration of a Moving Camera from Image Sequences*, Proceedings of the 7th Asian Conference on Computer Vision (ACCV 2006) (Hyderabad, India), Lecture Notes in Computer Science, vol. 385, Springer Verlag, January 2006, ISBN 9783540312444, pp. 71–80.

[Tho06] Thorsten Thormählen, *Zuverlässige Schätzung der Kamerabewegung aus einer Bildfolge [robust estimation of camera motion from an image sequence]*, Ph.D. thesis, University of Hannover, Fortschritt-Berichte, Reihe 10, VDI Verlag, 2006, ISBN 3183765101.

Citation

Thorsten Thormählen, Nils Hasler, Michael Wand, and Hans-Peter Seidel, *Registration of Sub-Sequence and Multi-Camera Reconstructions for Camera Motion Estimation*, Journal of Virtual Reality and Broadcasting, 7(2010), no. 2, July 2010, urn:nbn:de:0009-6-24333, ISSN 1860-2037.