Robust Maximum Likelihood Structure Invariant Merging of Projective Reconstructions

Stuart B. Heinrich[a,*], Wesley E. Snyder[a]

[a]*Department of Electrical and Computer Engineering, NC State University, Box 7911, Raleigh, NC 27695-7911*

Abstract

The ability to merge partial reconstructions into larger reconstructions is an important step used by many structure from motion systems. Merging is typically performed in metric space after autocalibration by solving an absolute orientation problem between structure point correspondences. However, autocalibration is an inherently sensitive procedure that is more reliable if delayed until the reconstruction is larger and more accurate. Additionally, the uncertainty of structure points triangulated in each partial reconstruction may prevent the accurate estimation of the proper orientation from corresponding structure points. In this paper we show how the orientation problem can be solved entirely in image space in a manner that is invariant to the potentially ill-estimated structure points, and is applicable to projective reconstructions because it does not require autocalibration. This method may be integrated into larger structure from motion systems for improved accuracy and reliability.

Keywords: structure from motion, projective reconstruction, merging

1. Introduction

In the structure-from-motion (SfM) problem, the objective is to simultaneously compute a reconstruction of 3D structure points and camera parameters from the motion parallax information encoded in a set of measured

*Corresponding author

Email addresses: sbheinri@ncsu.edu (Stuart B. Heinrich), wes@ncsu.edu (Wesley E. Snyder)

image correspondences. With uncalibrated cameras, projective reconstruction is usually the first step, followed by autocalibration to yield a metric reconstruction [1, p.265].

A projective reconstruction satisfies projection constraints but makes no assumptions about the camera intrinsic parameters, and is ambiguous up to a 4×4 projective ambiguity. This ambiguity can theoretically be resolved by autocalibration, which imposes the prior knowledge that real cameras do not produce skewed or stretched images [2]. However, autocalibration is not a well-posed problem; in general, there will not exist any 4×4 matrix that causes the camera intrinsic constraints to be satisfied exactly. Most autocalibration algorithms are very sensitive to reconstruction quality and will fail ungracefully when the initial projective reconstruction is not sufficiently accurate. Thus, autocalibration should be delayed until the projective reconstruction is as accurate as possible.

A commonly used technique for making a reconstruction that spans an arbitrary number of views is to compute many smaller independent reconstructions and then merge them together in order to obtain a larger reconstruction [3–6]. Most previous merging approaches have derived merging constraints by using correspondences between structure points [3, 5–7]. However, because Euclidean distance is not preserved under the projective ambiguity, this has required autocalibration to be performed on each partial reconstruction prior to merging, which is not only computationally expensive but also increases the risk of system failure due to the instabilities of autocalibration.

In this chapter, we revisit a linear approach to merging that measures distance in image space, thereby avoiding the need for premature autocalibration and also reducing the sensitivity to uncertainty in the structure points (Section 3.1). We show that although this approach usually produces good results, it can be unstable for certain camera configurations, but using the method symmetrically overcomes this problem (Section 4). Next, we propose a maximum likelihood nonlinear improvement of the merging homography that is completely invariant to the uncertainty in structure points (Section 5). We show how to robustly deal with outliers using this approach (Section 6), as well as how to efficiently merge inter-frame correspondences in order to strengthen the projective constraints for larger reconstructions while avoiding the systematic accumulation of errors (Section 7).

2. Merging Homography

The perspective projection of a homogeneous structure point $\mathbf{X} \in \mathbb{P}^3$, as viewed by a camera with 3×4 projection matrix $\mathbf{P}$, is a homogeneous image point $\mathbf{x} \in \mathbb{P}^2$, given by

$$\mathbf{x} \propto \mathbf{P}\mathbf{X}. \tag{1}$$

Let the estimate of the projection matrix for the jth view be denoted by $\widehat{\mathbf{P}}_j$ in the *left* reconstruction when it exists, and by $\widehat{\mathbf{P}}'_j$ in the *right* reconstruction when it exists. Similarly, the estimate of the ith structure point in the left reconstruction will be denoted by $\widehat{\mathbf{X}}_i$, and by $\widehat{\mathbf{X}}'_i$ in the right reconstruction.

Because both the left and right reconstructions are approximately related to some ground truth configuration by a 4×4 homography, there will also exist a 4×4 homography $\mathbf{H}$ that approximately relates the right reconstruction to the left reconstruction,

$$\widehat{\mathbf{P}}_j \propto \widehat{\mathbf{P}}'_j \mathbf{H} \quad \forall j \tag{2}$$

$$\widehat{\mathbf{X}}_i \propto \mathbf{H}^{-1} \widehat{\mathbf{X}}'_i \quad \forall i. \tag{3}$$

The goal of projective merging is to find the best possible estimate of $\mathbf{H}$. Once $\mathbf{H}$ is known, all projection matrices and structure points in the right reconstruction can be placed into the same projective reference frame as the left reconstruction using (2) and (3).

2.1. View Constraints

The homography $\mathbf{H}$ has 16 elements but just 15 degrees of freedom (dof) because it is a homogeneous entity with arbitrary scale; similarly, each 3×4 projection matrix has 11 dof. Any view j for which $\widehat{\mathbf{P}}_j$ exists in the left reconstruction and $\widehat{\mathbf{P}}'_j$ exists in the right reconstruction is an *overlapping view*, and hence by (2), $\mathbf{H}$ is over-determined and can be estimated using linear least squares from two or more overlapping views. However, we avoid this approach for the following reasons:

1. Because there can only be one estimate for each projection matrix, any overlapping views in the right reconstruction will be discarded when merging the right reconstruction into the left reconstruction (see Fig.

1). However, $\widehat{\mathbf{P}}_j$ will not be exactly equal to $\widehat{\mathbf{P}}'_j\mathbf{H}$, and there is no guarantee on how similar they will be. Even a small change in one element of a projection matrix can result in an *arbitrarily large* increase in the reprojection error of a structure point, depending on where that point is in 3D space. Thus, the results could be very unstable.

2. A least squares approach to merging using constraints from overlapping projection matrices would align them by minimizing the Frobenius norm. However, this is not a meaningful quantification of error because it does not consider the location of structure points that were used to estimate the projection matrix. Thus, the transformation that minimizes the Frobenius norm does not even approximately attempt to minimize reprojection error and this would magnify the effect of errors from (1).
3. Because overlapping views are discarded during a merge, it is computationally wasteful to use more overlapping views than necessary. For example, the result of merging two triplets with two overlapping views is only a net increase of one view in the merged reconstruction.

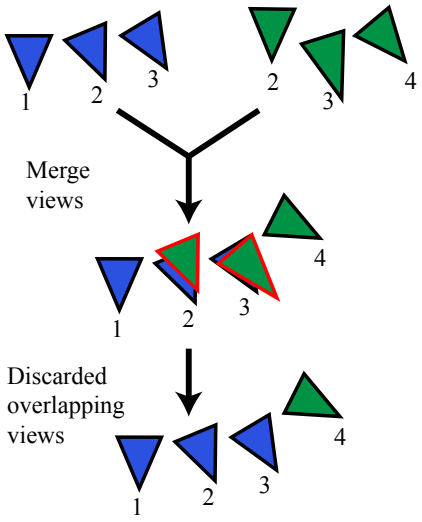

Figure 1. Example of projective merging with two views of overlap. The left reconstruction (blue) uses views $\{1, 2, 3\}$ and the right reconstruction (green) uses views $\{2, 3, 4\}$. In the first step, the right reconstruction is merged into the projective frame of the left reconstruction using only view constraints. Notice that neither of the projection matrices perfectly align. The two overlapping views (identified with red border) are discarded, causing the relative pose between views $\{2, 3, 4\}$ to be altered in a way that, depending on the location of structure points, may result in an unbounded increase of reprojection errors.

If there are no overlapping views then merging is still possible using (3), if corresponding structure points can somehow be identified. However, in order for a correspondence to exist between structure points there must have been a point visible in at least two views of each reconstruction (so that it could be triangulated), and if there are no overlapping views then this means that the point must have been imaged in at least four views. This is undesirable because it becomes exponentially more difficult to find reliable correspondences across more views.

If there is a single view of overlap, then this leaves $15 - 11 = 4$ dof remaining in the estimation of $\mathbf{H}$; thus, it is always possible to align the reconstructions so that *one* overlapping view is exactly equal, and then there will be no additional increase in reprojection error when this identical view is discarded during the merge. Furthermore, a correspondence of structure points requires an image feature to be identified in just three views, as noted in Laveau [8]. Thus, we consider single-view overlap to be the superior choice.

When there is a single view of overlap $\mathbf{H}$ can be parameterized so as to enforce (2) exactly [3]. Multiplying (2) by the pseudo-inverse, one obtains an equation for $\mathbf{H}$,

$$\widehat{\mathbf{P}}_j'^{+}\widehat{\mathbf{P}}_j \propto \mathbf{H}. \tag{4}$$

However, there is actually a 4-parameter family of solutions defined by the choice of $\mathbf{v}$ in

$$\mathbf{H}(\mathbf{v}) = \widehat{\mathbf{P}}_j'^{+}\widehat{\mathbf{P}}_j + \widehat{\mathbf{C}}'\mathbf{v}^{\mathsf{T}}, \tag{5}$$

where $\widehat{\mathbf{C}}'$ is the null space (eg, center of projection) of $\widehat{\mathbf{P}}_j'$. This can be easily verified by left-multiplying again by $\widehat{\mathbf{P}}_j'$ to yield

$$\widehat{\mathbf{P}}_j'\mathbf{H}(\mathbf{v}) = \widehat{\mathbf{P}}_j + \widehat{\mathbf{P}}_j'\widehat{\mathbf{C}}'\mathbf{v}^{\mathsf{T}}, \tag{6}$$

which is identical to (2) because $\widehat{\mathbf{P}}_j'\widehat{\mathbf{C}}' = \mathbf{0}$ by definition.

3. Merging with Single-View Overlap

Given a set of corresponding structure points $\widehat{\mathbf{X}}_i \leftrightarrow \widehat{\mathbf{X}}_i'$, the most obvious way to constrain $\mathbf{H}(\mathbf{v})$ would be to use (3) directly by minimizing

$$\sum_i D_E(\mathbf{H}(\mathbf{v})\widehat{\mathbf{X}}_i, \widehat{\mathbf{X}}_i')^2, \tag{7}$$

where $D_E(\mathbf{a}, \mathbf{b})$ is the *inhomogeneous* Euclidean distance. However, Euclidean distance is not preserved under the projective ambiguity, and attempting to minimize Euclidean distance in a projective space would be truly meaningless [3].

As an example, two points might be measured as having a distance of 1 (with some arbitrary units) in the projective space when in fact the actual distance between those points after metric rectification should be ∞. Thus, in order to make the distance measurement in (7) meaningful, it would be necessary to first autocalibrate the right reconstruction (or the left reconstruction if $\mathbf{H}^{-1}$ is estimated instead).

A second complication is that the structure points in (7) are homogeneous, but the Euclidean distance is measured between inhomogeneous points which means that (7) will require a nonlinear minimization. However, this is not a serious complication because a good initialization can be obtained linearly by minimizing algebraic (rather than Euclidean) distance, as discussed in both Laveau [8] and Fitzgibbon and Zisserman [3].

If both left and right reconstructions have been autocalibrated, then $\mathbf{H}$ should theoretically be a similarity transform with 7 dof. The optimal estimation of a similarity transform between two corresponding point sets, called the absolute orientation problem, can be performed linearly [9–11], and this is perhaps the most commonly used approach to merging (see Farenzena et al. [5], Frahm et al. [6], Repko and Pollefeys [7]).

However, because autocalibration is not a very well posed problem that cannot be solved perfectly (especially for smaller reconstructions), there is always some projective ambiguity remaining which means that the alignment between two autocalibrated reconstructions is not truly a similarity transformation.

Another problem that applies to merging in either projective or metric spaces is that structure points generally have a large degree of uncertainty, and this causes the constraints of (3) to be poorly satisfied even for the best choice of $\mathbf{H}$. The approach of Matei and Meer [11] partially deals with this problem by taking into account the approximate uncertainty of structure points in the absolute orientation problem; however, we prefer to overcome the root of this problem.

3.1. Nister's Linear Method

A more attractive solution is to measure distance in image space, because image space is already a metric space. In other words, if $\tilde{\mathbf{x}}_i^j$ is the measured observation of $\mathbf{X}_i$ in view j, then one could instead minimize

$$\sum_{i,j} D_E(\widehat{\mathbf{P}}'_j \mathbf{H}(\mathbf{v}) \widehat{\mathbf{X}}_i, \tilde{\mathbf{x}}_i^j)^2. \tag{8}$$

Not only does (8) bypass the issue of measuring error in projective spaces, but for most configurations it is fairly robust to structure points that have a large degree of uncertainty in their depth, because re-projecting the structure

point to measure distance on the image plane largely cancels out the uncertainty as long as the views in the right reconstruction are relatively close to the views in the left reconstruction.

Although there is no linear solution to (8), an algorithm was given in Nister [4, p. 65] that minimizes a *very similar* problem linearly. Although it is not explicitly mentioned there, that problem is

$$\sum_{i,j} D_A(\widehat{\mathbf{P}}'_j \mathbf{H}(\mathbf{v}) \widehat{\mathbf{X}}_i, \hat{\mathbf{x}}_i^j)^2, \tag{9}$$

where $D_A(\mathbf{a}, \mathbf{b})$ is the *inhomogeneous* algebraic distance, and $\hat{\mathbf{x}}_i^j$ is the closest point to $\tilde{\mathbf{x}}_i^j$ that can be found by varying $\mathbf{v}$. The details of this method are given below.

Let the index of the overlapping view be denoted by o. Then the epipolar line $\mathbf{l}_i^j$ in the jth view that contains the image of the ith point and the epipole of the oth view is given by

$$\mathbf{l}_i^j = \widehat{\mathbf{P}}'_j {\widehat{\mathbf{P}}'_o}^+ \widehat{\mathbf{P}}_o \widehat{\mathbf{X}}_i \times \widehat{\mathbf{P}}'_j \widehat{\mathbf{C}}'_o. \tag{10}$$

This may be verified as follows. To the right of the cross product, $\widehat{\mathbf{C}}'_o$ is the camera center of the oth view in the right reconstruction, and projecting by $\widehat{\mathbf{P}}'_j$ gives its epipole in the jth view. On the left hand side, we start with $\widehat{\mathbf{X}}_i$, the ith point in the left reconstruction. We multiply by $\widehat{\mathbf{P}}_o$ to get the image of this point in the overlap view, then by ${\widehat{\mathbf{P}}'_o}^+$ to back-project this to a structure point in the right reconstruction, and finally by $\widehat{\mathbf{P}}'_j$ to obtain an image in the jth view. The cross product of two points gives the line joining those points, so $\mathbf{l}_i^j$ is the desired epipolar line.

The closest point on this epipolar line to the measurement $\tilde{\mathbf{x}}_i^j$ is then given by

$$\hat{\mathbf{x}}_i^j = [\mathbf{l}_i^j]_\times [\tilde{\mathbf{x}}_i^j]_\times \Omega_\infty^* \mathbf{l}_i^j, \tag{11}$$

where $\Omega_\infty^* = \mathrm{diag}(1, 1, 0)$ is the absolute dual conic in a metric frame, and $[\mathbf{x}]_\times$ is the 3×3 skew-symmetric cross product matrix of $\mathbf{x}$.

Taking $\widehat{\mathbf{X}}_i$, the ith point in the left reconstruction, multiplying by $\mathbf{H}(\mathbf{v})$ should transform it into the right reconstruction, and then multiplying by $\widehat{\mathbf{P}}'_j$

gives its image in the jth view, which should be $\hat{\mathbf{x}}_i^j$. This is a homogeneous equivalence constraint that implies a zero cross product,

$$\mathbf{0} = \hat{\mathbf{x}}_i^j \times \widehat{\mathbf{P}}'_j \mathbf{H}(\mathbf{v}) \widehat{\mathbf{X}}_i \tag{12}$$

$$= [\hat{\mathbf{x}}_i^j]_\times \widehat{\mathbf{P}}'_j (\widehat{\mathbf{P}}'^{+}_o \widehat{\mathbf{P}}_o + \widehat{\mathbf{C}}'_o \mathbf{v}^\mathsf{T}) \widehat{\mathbf{X}}_i \tag{13}$$

$$= [\hat{\mathbf{x}}_i^j]_\times \widehat{\mathbf{P}}'_j \widehat{\mathbf{P}}'^{+}_o \widehat{\mathbf{P}}_o \widehat{\mathbf{X}}_i + [\hat{\mathbf{x}}_i^j]_\times \widehat{\mathbf{P}}'_j \widehat{\mathbf{C}}'_o \widehat{\mathbf{X}}_i^\mathsf{T} \mathbf{v}. \tag{14}$$

Rearranging and left-multiplying by $([\hat{\mathbf{x}}_i^j]_\times \widehat{\mathbf{P}}'_j \widehat{\mathbf{C}}'_o)^\mathsf{T}$, a single linear constraint on $\mathbf{v}$ is obtained,

$$[\hat{\mathbf{x}}_i^j]_\times \widehat{\mathbf{P}}'_j \widehat{\mathbf{C}}'_o \widehat{\mathbf{X}}_i^\mathsf{T} \mathbf{v} = -[\hat{\mathbf{x}}_i^j]_\times \widehat{\mathbf{P}}'_j \widehat{\mathbf{P}}'^{+}_o \widehat{\mathbf{P}}_o \widehat{\mathbf{X}}_i \tag{15}$$

$$\widehat{\mathbf{X}}_i^\mathsf{T} \mathbf{v} = \frac{-\widehat{\mathbf{C}}'^{\mathsf{T}}_o \widehat{\mathbf{P}}'^{\mathsf{T}}_j [\hat{\mathbf{x}}_i^j]_\times^2 \widehat{\mathbf{P}}'_j \widehat{\mathbf{P}}'^{+}_o \widehat{\mathbf{P}}_o \widehat{\mathbf{X}}_i}{\left|\left| [\hat{\mathbf{x}}_i^j]_\times \widehat{\mathbf{P}}'_j \widehat{\mathbf{C}}'_o \right|\right|^2}. \tag{16}$$

Thus, each correspondence between a structure point in the left reconstruction and an image of that point in the right reconstruction in any view other than the overlapping view provides a linear constraint. A total of at least four such constraints are needed to constrain $\mathbf{H}(\mathbf{v})$.

4. Symmetric Linear Merging

In most cases, Nister's linear algorithm will work well. However, if the baselines between *all* views in the left reconstruction are relatively small, then the structure points that are chosen from the left reconstruction will have a large degree of uncertainty in their depth. If in addition, the baselines between views in the right reconstruction are *not all* small, then this large uncertainty in depth may cause the projection of those points into the right reconstruction to be very inaccurate (see Fig. 2), and this can result in a failure to properly merge the reconstructions.

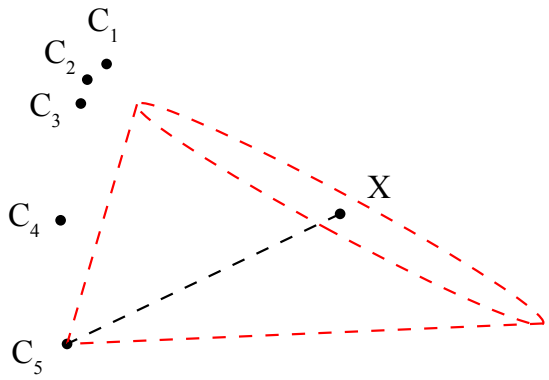

Figure 2. Example of a configuration that may result in unstable merge using Nister's linear algorithm. The true location of the five cameras are indicated by $\widehat{\mathbf{C}}_i$, $i = 1 \ldots 5$. The true location of a structure point $\mathbf{X}$ is also marked. The left reconstruction consists of views $\{1, 2, 3\}$ and the right reconstruction consists of views $\{3, 4, 5\}$. Because all views in the left reconstruction are relatively close together, there is a large degree of uncertainty in the triangulation of any structure point $\mathbf{X}$, indicated by the dotted red ellipse. As a result, the projection of this triangulated point into the 5th view may be far from the measured image point, causing the merging constraint to be bad and preventing the algorithm from identifying a good merging homography.

In this case, one notices that structure points in the right reconstruction would have much less uncertainty in their depth (because the baselines are not all small), and hence these points could be merged into the left reconstruction and projected onto the image plane where they would correctly match up with the measured image points.

In practice, a method of keyframe selection (see, for example Repko and Pollefeys [7], Torr [12], Pollefeys et al. [13], Beder and Steffen [14]) should be used to ensure that there is some sufficient baseline between each view. However, the fact remains that between any two reconstructions that one wishes to merge, one of them will have wider baseline than the other, and because Nister's linear algorithm is asymmetric, the structure points that are used for merging constraints should be chosen from the side that has wider baseline in order to achieve the greatest accuracy.

It is possible (albeit messy) to derive constraints on $\mathbf{v}$ from structure points in the right reconstruction corresponding to image points in the left reconstruction, and combining these with the constraints from (16) would allow $\mathbf{v}$ to be estimated linearly from a set of symmetric constraints. However, this type of symmetry would not actually be a good thing because the

constraints from one direction always have higher error, so a single symmetric estimation would only be as good as the constraints of the lesser direction.

Therefore, our solution is to always apply the algorithm in the forward direction (as described in Section 3.1) as well as the reverse direction, and take the solution that results in lower reprojection error. In the reverse direction we use structure points from the right reconstruction corresponding to image points in the left reconstruction, and solve for $\mathbf{H}^{-1}$ parameterized by $\mathbf{v}'$,

$$\mathbf{H}^{-1}(\mathbf{v}') = \widehat{\mathbf{P}}_j^{+}\widehat{\mathbf{P}}'_j + \widehat{\mathbf{C}}\mathbf{v}'^{\mathsf{T}}, \tag{17}$$

where $\widehat{\mathbf{C}}$ is the null space of $\widehat{\mathbf{P}}_j$. Using the overlap view, and substituting (5) into (17), we can then compute $\mathbf{v}$ from $\mathbf{v}'$,

$$\widehat{\mathbf{P}}'^{+}_o\widehat{\mathbf{P}}_o + \widehat{\mathbf{C}}'_o\mathbf{v}^{\mathsf{T}} = \left(\widehat{\mathbf{P}}_o^{+}\widehat{\mathbf{P}}'_o + \widehat{\mathbf{C}}_o\mathbf{v}'^{\mathsf{T}}\right)^{-1} \tag{18}$$

$$\widehat{\mathbf{C}}'_o\mathbf{v}^{\mathsf{T}} = \left(\widehat{\mathbf{P}}_o^{+}\widehat{\mathbf{P}}'_o + \widehat{\mathbf{C}}_o\mathbf{v}'^{\mathsf{T}}\right)^{-1} - \widehat{\mathbf{P}}'^{+}_o\widehat{\mathbf{P}}_o \tag{19}$$

$$\mathbf{v}^{\mathsf{T}} = \frac{\widehat{\mathbf{C}}'^{\mathsf{T}}_o}{||\widehat{\mathbf{C}}'_o||^2}\left(\left(\widehat{\mathbf{P}}_o^{+}\widehat{\mathbf{P}}'_o + \widehat{\mathbf{C}}_o\mathbf{v}'^{\mathsf{T}}\right)^{-1} - \widehat{\mathbf{P}}'^{+}_o\widehat{\mathbf{P}}_o\right). \tag{20}$$

This still allows us to merge the right reconstruction into the left reconstruction even when using constraints from the reverse direction.

5. Structure Invariant Maximum Likelihood Merging

Once the merging homography has been found, structure points can be retriangulated from all views to obtain a point that is more accurate than the corresponding points previously existing in the left or right partial reconstructions. Thus, the ideal merging homography should seek to maximize the accuracy of the cameras without reference to the previously triangulated structure points.

Although the symmetric modification improves the reliability of Nister's linear method to obtain a better initial estimate of $\mathbf{H}$, it is still less than ideal; in particular, it attempts to minimize distance between the projection of a structure point and an artificial point $\hat{\mathbf{x}}_i^j$ rather than the actual image measurement $\tilde{\mathbf{x}}_i^j$, it minimizes an algebraic rather than Euclidean distance,

and most importantly it is not completely invariant to the uncertainty in previously triangulated structure points.

Therefore, we seek the homography that would maximize the likelihood of the overall reconstruction after retriangulating all structure points from the merged camera matrices using a maximum likelihood method. Assuming measurement noise is Gaussian, it is well known that maximizing likelihood is equivalent to minimizing reprojection error [1, p.102], and therefore the solution is given by

$$\widehat{\mathbf{v}}_{ML} = \underset{\mathbf{v}}{\operatorname{argmin}} \sum_{i,j} D_E(\tilde{\mathbf{x}}_i^j, \overline{\mathbf{P}}_j \widehat{\mathbf{X}}_{MLi})^2, \tag{21}$$

where $\overline{\mathbf{P}}_j(\mathbf{v})$ are the merged projection matrices as a function of $\mathbf{v}$,

$$\overline{\mathbf{P}}_j(\mathbf{v}) = \begin{cases} \widehat{\mathbf{P}}_j, & j \in \mathcal{L}, \\ \widehat{\mathbf{P}}'_j \mathbf{H}(\mathbf{v}), & j \notin \mathcal{L}, \end{cases} \tag{22}$$

and $\widehat{\mathbf{X}}_{MLi}$ is the maximum likelihood triangulation of the ith structure point from all available image measurements $\tilde{x}_i^j$ with respect to the merged cameras, $\overline{\mathbf{P}}_j$.

For points visible in just two views, we compute the maximum likelihood triangulation in closed form as in Hartley and Sturm [15]; for more than two views, we compute the maximum likelihood triangulation by nonlinearly minimizing the sum of squared reprojection errors using Levenberg-Marquardt [16] from the homogeneous linear initialization [1, p. 313]. In order to minimize (21) with respect to $\mathbf{v}$, we initialize using our symmetric linear correction to Nister's method and then use Levenberg-Marquardt with numerical differentiation.

It should be noted that even though (21) provides a maximum likelihood estimate of the merging homography, no estimate of the merging homography will produce a maximum likelihood reconstruction. Therefore, we always follow up merging with bundle adjustment [17, 18], the maximum likelihood nonlinear improvement of a projective reconstruction.

Of course, one could skip (21) and proceed directly to bundle adjustment after using the linear initialization. However, for a system of n points and m views projective bundle adjustment has $12m+3n$ parameters, and in a typical problem there may be hundreds of thousands of free parameters, making bundle adjustment not only computationally expensive but very susceptible

to falling into local minima. Projection matrices in bundle adjustment are almost always parameterized using an absolute coordinate system, and as a result a very small error in the merging homography $\mathbf{H}$ could necessitate rather significant changes to half of the views during the subsequent bundle adjustment. In contrast, $\mathbf{H}$ has only 15 dof, so it will be more efficient and reliable to optimize $\mathbf{H}$ as much as possible prior to bundle adjustment.

6. Robustness to Outliers

Because the measurements $\tilde{x}_i^j$ are obtained using a correspondence finding algorithm on images, there are likely to be some mismatches that result in outliers with very large error. These outliers violate the assumed Gaussian noise model, and it is therefore important to detect and ignore these measurements in order to make a robust estimate of $\mathbf{H}$ using (21).

We use the RANSAC [19] paradigm to handle outliers, specifically MSAC [20]. From the set of structure point correspondences $\widehat{\mathbf{X}}_i \leftrightarrow \widehat{\mathbf{X}}'_i$, our objective is to find the largest sample consensus of correspondences that agree upon a homography which can merge the partial reconstructions while keeping the reprojection error of all retriangulated structure points $\overline{\mathbf{X}}_i$ in (21) below some threshold τ.

This is done by picking random subsets from the set of correspondences and then minimizing (21) (initialized with the symmetric linear method) using *only* the selected subset of correspondences. From this estimate of $\mathbf{v}$ we then enlarge the subset to include all inliers and repeat within the RANSAC framework to find the largest sample consensus. Finally, we iteratively re-minimize (21) and re-classify inliers until convergence.

In general, the minimum number of correspondences that must be used in a random sampling is data dependent because the number of constraints that are provided by each correspondence depends on the number of images that a structure point is viewed in. However, we do not aim to use minimal subsets because a greater robustness to noise is achieved by using larger subsets. When using triplet correspondences to span the overlap view, we use a subset size of 10 and this results in 10 constraints on $\mathbf{v}$, which we find provides a good balance between speed of convergence and robustness to noise.

7. Merging Correspondences

After merging two partial reconstructions into a larger reconstruction with more views it may be possible to triangulate structure points using the additional views for greater accuracy, if the image measurements exist. For example, if the left reconstruction contains views $\{1, 2, 3\}$ and the right reconstruction contains views $\{3, 4, 5\}$, and one has good measurements for $\tilde{\mathbf{x}}_i^j$, $j = 1 \ldots 5$, then after merging an estimate of the point $\mathbf{X}_i$ can be made using all five views that will be more accurate than the estimate of that point in the original left or right reconstructions.

It is typical to detect correspondences in a separate module, either using feature tracking [21–23] or inter-frame matching, that produces as output an increasing list containing the coordinates of an observed feature point in a series of views. However, despite attempts to remove outliers [24–26], some of these measurements will still be erroneous.

In this example, suppose that $\tilde{\mathbf{x}}_i^4$ is a bad measurement. Thus, it is likely that the point $\widehat{\mathbf{X}}_i$ will exist in the left reconstruction, but $\widehat{\mathbf{X}}_i'$ will not exist in the right reconstruction. After merging the two reconstructions together and attempting to triangulate a new point using all the images of this point, this will also fail and hence the ith structure point will be lost from the merged reconstruction even though $\widehat{\mathbf{X}}_i$ was formerly a well-triangulated point.

As feature tracks are increased in length, the probability of an outlier match increases, so it is important to have a method of preventing all the good points from eventually being thrown out when merging together partial reconstructions. Similar to Thormahlen et al. [27], our solution to this problem is to associate measurements independently with each partial reconstruction, and then attempt to merge these correspondences when the partial reconstructions are merged.

In our implementation we have not used feature tracking to find the initial correspondences but rather we have used wide-baseline matching of Harris and Stephens [28] corner points. To compute the initial triplet reconstructions we search for correspondences of triplets, and to merge them together we use a separate set of triplet correspondences. After merging the two projective reconstructions we search for structure points between the left and right reconstruction that can be merged.

In order to identify potential structure points for merging we project all of the structure points from the right reconstruction onto the image plane of the overlapping view. These image points are stored in a uniform grid structure

[29] that allows all points in a fixed radius to be found rapidly. For each structure point in the left reconstruction, we search for potential matches around its projected image point in the overlap view, and for each potential match we compute the maximum likelihood triangulation from the set of merged correspondences associated with those points. If the triangulated point reprojects back to all of the original measurements within some small threshold, then the merge is adopted.

8. Results

We compare the various merging approaches on synthetic data with controlled levels of noise so that the statistical differences between algorithms is made clearly apparent. In our synthetic tests, five cameras are generated on a circle of radius 100 units looking approximately toward the origin (± 20 units). The angular separation between successive cameras is uniformly distributed in the range of $0.1^\circ - 10^\circ$, and camera focal length is uniformly distributed in the range of $600 - 800$ units.

For scene structure, 100 points are generated on the surface of a cube of width 100 units that is shifted some distance from the origin in the direction of the average camera principal ray. A top down view of a generic synthetic configuration is shown in Fig. 3. Correspondences are generated by projecting the true structure points onto the image plane and adding normally distributed noise to simulate measurement error in the correspondence finder.

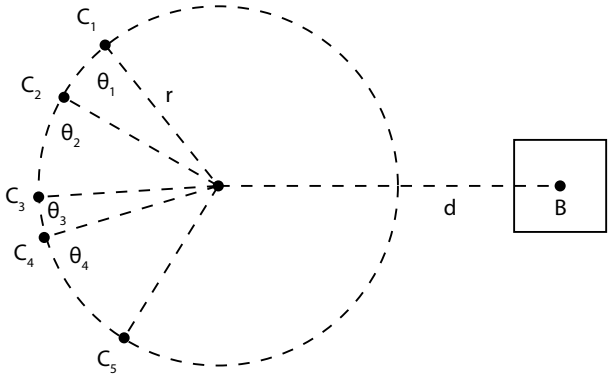

Figure 3. Top down view of a synthetic configuration. Cameras centers $\mathbf{C}_1, \ldots, \mathbf{C}_5$ are located on a circle of radius r with random angular separations of $\theta_1, \ldots, \theta_4$. The structure points are generated on the surface of a cube centered at $\mathbf{B}$, a distance d from the origin.

From these noisy correspondences we compute a robust estimate of the trifocal tensor for the first three views and the last three views (so that there is one view of overlap). We then attempt to merge these two reconstructions into a single reconstruction covering all five views using: (a) the optimal absolute orientation method of [10], after autocalibrating both partial reconstructions using the recent method of [30]; (b) Nister's forward linear method (Section 3.1); (c) our symmetric variation on Nister's method (Section 4); (d) the proposed Structure Invariant Maximum Likelihood (SIML) method (Section 5). We evaluate merging success for any particular trial by using the mean reprojection error of the merged result prior to bundle adjustment, because it is well known that the maximum likelihood projective reconstruction should minimize reprojection error.

In our first experiment we examined sensitivity to noise. This was done by generating noisy correspondences from 100 random configurations at each level of noise and then looking at the median of the mean reprojection error (see Fig. 4). We observe that all methods have zero median error in the absence of noise, but the absolute orientation method is extremely sensitive and produces high median errors even under low noise conditions. In contrast, the image-spaced approaches exhibit reconstruction error that is almost proportional to the measurement error. Our symmetric linear method has lower median error than Nister's method, and the proposed SIML improvement has lower median error still, although these reductions to *median* error are relatively marginal.

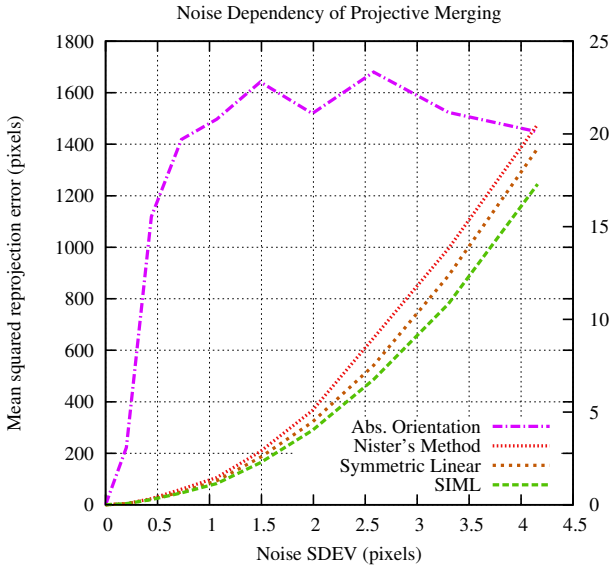

Figure 4. Comparison of reconstruction quality as a function of measurement noise. Scene distance is fixed at 500 units. Errors for the absolute orientation method are shown on the primary axis and errors for the other methods are shown on the secondary axis. The plotted curves are the median of 100 trials. All methods have zero median error in the absence of noise, but the absolute orientation method is extremely sensitive and produces high median errors even under low noise.

In our second experiment, we fix the noise level at $\sigma = 1$ pixels and examine the merged reconstruction quality as a function of scene distance (Fig. 5). Curiously, we observe that the reprojection error of the merged reconstruction is not a monotonic function of scene distance. This effect, while initially perplexing, can be attributed to two conflicting forces. On the one hand, the absolute (3D) error is increased for more distant reconstructions because measurement noise (which is added in image space) becomes relatively greater. On the other hand, more distant scenes tend to have lower reprojection error because the projection of the entire scene bounding box occupies a smaller portion of the image. Thus, as distance of the scene is increased, the merged reprojection error will gradually increase until it becomes no better than random guessing, at which point the reprojection error will gradually reduce and asymptote at some small value, although the absolute errors continue to increase.

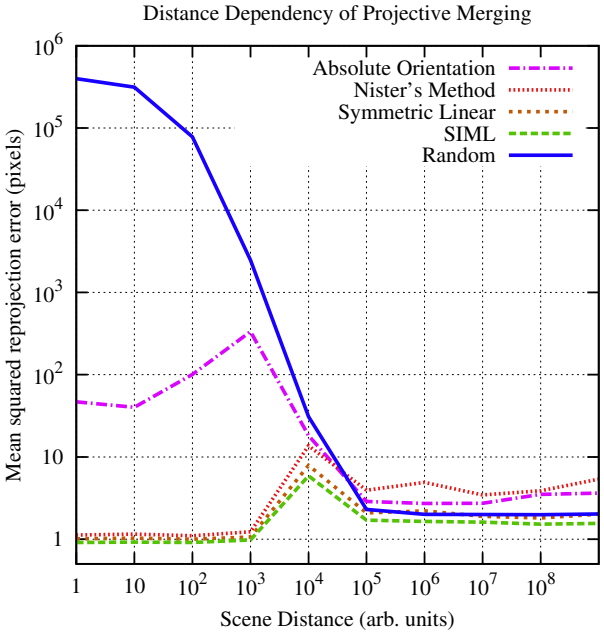

Figure 5. Comparison of reconstruction quality as a function of scene distance. Measurement noise is fixed at $\sigma = 1$ pixels. Interestingly, the reprojection error is not a monotonic function of scene distance. The plotted curves are the median of 100 trials. For this distribution of configurations, we see that the median absolute orientation method performs better than random when the scene distance is within 10^3, whereas the image-space methods perform better than random when the scene distance is less than 10^4, but they only provide accurate results out to 10^3. The absolute orientation method does not provide accurate results for any distance at this level of noise.

In order to approximate this overall downward trend in expected reprojection error for more distant geometry, we have plotted the median of the mean reprojection error obtained by randomizing the order of structure points in the true configuration and measuring the distance to the (incorrect) image projections after projecting those structure points by the true projection matrices. In other words, this shows the expected reprojection error if structure points were to be randomly chosen within the scene volume rather than being precisely triangulated.

We see from the graph that the median performance of the absolute orientation method is better than random only when the scene distance is within 10^3, whereas the image-space methods perform better than random when the scene distance is less than 10^4, but they only provide accurate results out to 10^3. Between these methods, the proposed SIML method has the lowest

median error, although the improvements to *median* performance are still only marginal. The absolute orientation method does not provide accurate results for any distance at this level of noise.

For extremely far distances (10^5 and beyond in this case), the random curve has lower error than some of the estimation methods. This is because at this extreme distance, the prior knowledge of the true scene bounding box that was assumed when generating the random curve becomes more informative than the image measurements, because the large relative noise causes the uncertainty ellipsoid of a triangulated structure point to become larger than the true scene bounding box.

It should be noted that there is nothing magic about the number 10^3, it is simply the point at which the signal to noise ratio becomes too small for accurate reconstruction, and this is dependent on the specific camera configuration (particularly the distance between cameras) as well as the correspondence measurement noise.

We compare the Empirical Cumulative Distribution Functions (ECDFs) of the mean squared reprojection error (MSE) for each method of merging using from a set of 1000 random configurations with noise fixed at $\sigma = 1$ pixels and scene 'distance' set to zero (i.e., the scene being centered at the origin) in Fig. 6.

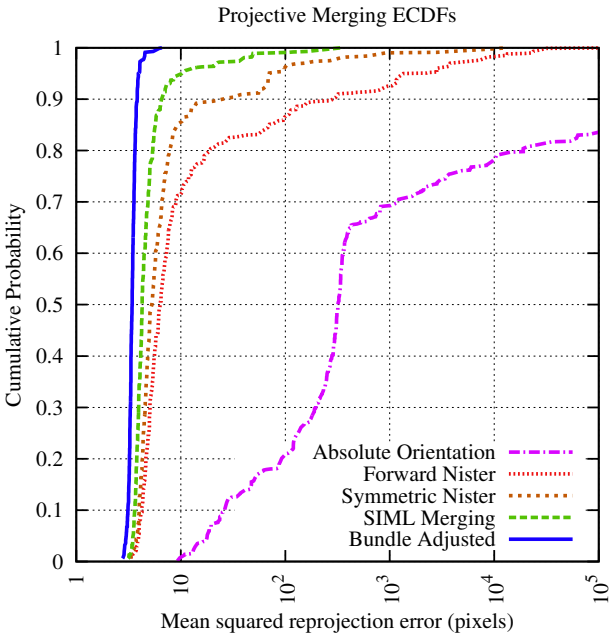

Figure 6. Comparison of the empirical cumulative distribution of mean squared reprojection error of the merged reconstructions using various merging methods, based on merging from 1000 randomly generated configurations.

By looking at the complete performance distribution, we finally see that the improvements offered by our symmetric and SIML improvements are significant when it comes to tail performance. This is expected, because the improvements are primarily designed to increase robustness under conditions of small baseline, but the majority of configurations that we randomly generate will not have the problem of small baseline.

Specifically, using Nister's original method the 90th percentile of mean squared reprojection error was 300 pixels, whereas the 90th percentile was reduced to 29 pixels after our symmetric modification, and further reduced down to just 6.5 pixels using the SIML method. Finally, after bundle adjusting the result, the 90th percentile error was reduced to 3.8 pixels, and was never worse than 6.6 pixels.

Finally, we demonstrate an example of the robust SIML merging method on some measurements gathered from real data (see Fig. 7). We started with a series of five sequential snapshots of a desk and then proceeded to find correspondences by matching corners. We computed two estimates of the trifocal tensor robustly and then merged them together using the proposed

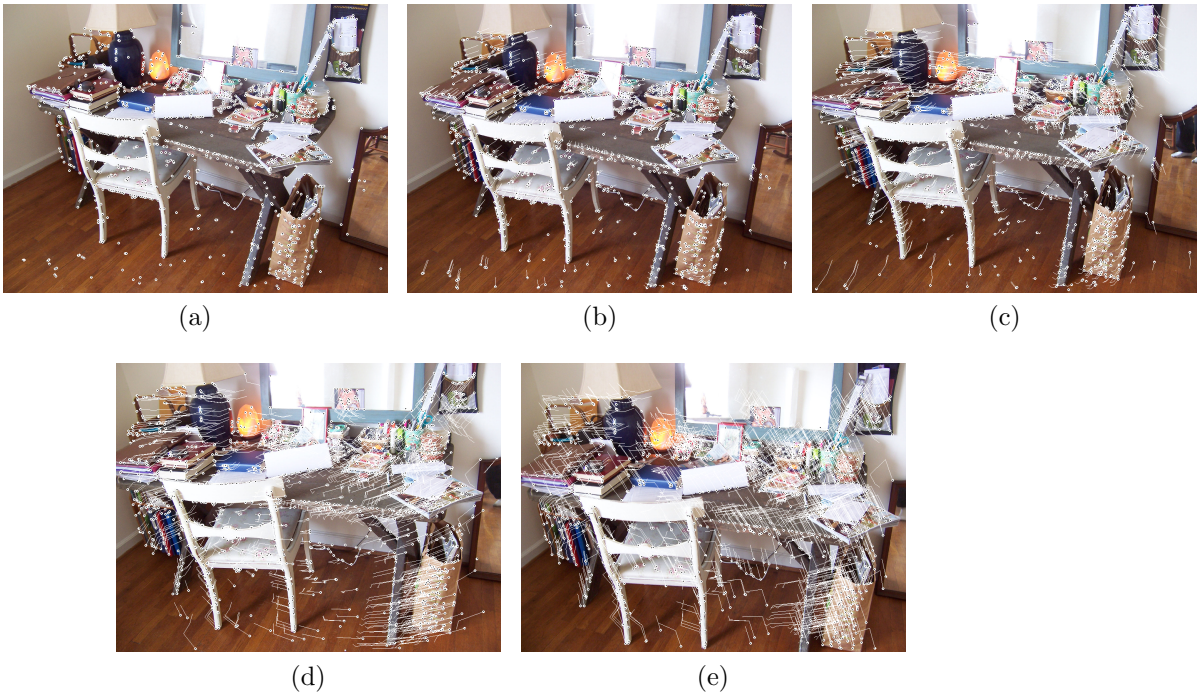

Figure 7. Projected structure from a reconstruction of five views that was formed by merging two triplets that overlap by one view using the proposed approach. The white tracks show image measurements and the black points are the reprojected structure points. The merged reconstruction consists of 3,367 structure points with a mean squared reprojection error of 0.51 pixels (the image width is 1000 pixels).

robust maximum likelihood method. The correspondences were merged as described in Section 7 and then bundle adjustment was used to nonlinearly improved the merged result.

In this reconstruction we found a total of 2,661 structure points with a mean squared reprojection error of 0.55 pixels, all of which were below the threshold of 2 pixels used within the RANSAC framework. We show a selection of three views from the merged result of five views in Fig. 7, where we have drawn the reprojected structure points in comparison to the original corner points to demonstrate the low reprojection error visually. To reduce visual clutter, we only draw the reprojections of points that were merged so that they have an image in all five views. Some views of the reconstructed point cloud are shown in Fig. 8.

(a) (b)

Figure 8. Two views of the structure points in the merged desk reconstruction. (a) from a side perspective, and (b) from a front perspective. The reconstructed cameras are shown as pyramids.

9. Conclusions

Merging of partial reconstructions requires the computation of the transformation matrix that aligns their respective spaces. This is commonly solved as an absolute orientation problem in metric space after autocalibration. However, autocalibration is an inherently sensitive procedure that we feel is best delayed until the reconstruction becomes larger and more precise in order to avoid instabilities. Moreover, the absolute orientation approach is very sensitive to the accuracy of structure points that have been already triangulated in the partial reconstructions.

By using Nister's method, errors can be measured in metric image space, thereby avoiding the need to perform premature autocalibration, which also has the advantage of canceling out the majority of structure point uncertainty, thereby reducing sensitivity to noise. However, this uncertainty is never fully canceled out. To solve this problem, we have proposed a way to apply Nister's method symmetrically and thereby provide a more reliable initialization. Most importantly, we have also proposed a maximum likelihood method that is completely invariant to the uncertainty in structure points, and shown how this can be used within a RANSAC framework to obtain truly robust results.

Thus, this new merging method can be used to increase the accuracy and

reliability of any projective structure from motion system that relies on a merging operation.

Acknowledgements

We thank Dr. Margaret J. Eppstein for her useful comments.

References

[1] R. I. Hartley, A. Zisserman, Multiple View Geometry in Computer Vision, Cambridge University Press, ISBN: 0521540518, second edn., 2004.

[2] M. Pollefeys, R. Koch, L. V. Gool, Self-Calibration and Metric Reconstruction in spite of Varying and Unknown Instrinsic Camera Parameters, Intl. Journal of Computer Vision .

[3] A. Fitzgibbon, A. Zisserman, Automatic camera recovery for closed or open image sequences, in: Proc. of the 5th European Conf. on Computer Vision, 311–326, 1998.

[4] D. Nister, Automatic dense reconstruction from uncalibrated video sequences, Ph.D. thesis, Royal Institute of Technology KTH, Stockholm, Sweden, 2001.

[5] M. Farenzena, A. Fusiello, R. Gherardi, Structure-and-motion pipeline on a hierarchical cluster tree, in: Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on, 1489 –1496, doi:10.1109/ICCVW.2009.5457435, 2009.

[6] J.-M. Frahm, P. Fite-Georgel, D. Gallup, T. Johnson, R. Ragauram, C. Wu, Y.-H. Jen, E. Dunn, B. Clipp, S. Lazebnik, M. Pollefeys, Building Rome on a Cloudless Day, in: ECCV 2010: Proceedings of the 11th European Conference on Computer Vision, 368–381, 2010.

[7] J. Repko, M. Pollefeys, 3D models from extended uncalibrated video sequences: addressing key-frame selection and projective drift, in: 3-D Digital Imaging and Modeling, 2005. 3DIM 2005. Fifth International Conference on, ISSN 1550-6185, 150 – 157, doi:10.1109/3DIM.2005.4, 2005.

[8] S. Laveau, Geometry of a system of N cameras. Theory, estimation and applications., Ph.D. thesis, INRIA, 1996.

[9] B. K. P. Horn, H. Hilden, S. Negahdaripour, Closed-Form Solution of Absolute Orientation using Orthonormal Matrices, Journal of the Optical Society of America 5 (7) (1988) 1127–1135.

[10] S. Umeyama, Least-Squares Estimation of Transformation Parameters Between Two Point Patterns, IEEE Trans. Pattern Anal. Mach. Intell. 13 (1991) 376–380, ISSN 0162-8828, doi:http://dx.doi.org/10.1109/34.88573, URL http://dx.doi.org/10.1109/34.88573.

[11] B. Matei, P. Meer, Optimal rigid motion estimation and performance evaluation with bootstrap, in: Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on., vol. 1, 2 vol. (xxiii+637+663), doi:10.1109/CVPR.1999.786961, 1999.

[12] P. H. S. Torr, Bayesian Model Estimation and Selection for Epipolar Geometry and Generic Manifold Fitting, Int. J. Comput. Vision 50 (2002) 35–61, ISSN 0920-5691, doi:10.1023/A:1020224303087.

[13] M. Pollefeys, L. V. Gool, M. Vergauwen, K. Cornelis, F. Verbiest, J. Tops, Video-to-3d, in: Proceedings of Photogrammetric Computer Vision, 2002.

[14] C. Beder, R. Steffen, Determining an Initial Image Pair for Fixing the Scale of a 3D Reconstruction from an Image Sequence, in: Pattern Recognition, vol. 4174 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, 657–666, 2006.

[15] R. I. Hartley, P. Sturm, Triangulation, Computer Vision and Image Understanding 68 (2) (1997) 146 – 157, ISSN 1077-3142.

[16] D. W. Marquardt, An Algorithm for Least-Squares Estimation of Nonlinear Parameters, SIAM Journal on Applied Mathematics 11 (2) (1963) 431–441, doi:10.1137/0111030, URL http://link.aip.org/link/?SMM/11/431/1.

[17] B. Triggs, P. McLauchlan, R. Hartley, A. Fitzgibbon, Bundle Adjustment – A Modern Synthesis, in: B. Triggs, A. Zisserman, R. Szeliski

(Eds.), Vision Algorithms: Theory and Practice, vol. 1883 of *Lecture Notes in Computer Science*, Springer-Verlag, 298–372, URL http://lear.inrialpes.fr/pubs/2000/TMHF00, 2000.

[18] M. Lourakis, A. Argyros, The Design and Implementation of a Generic Sparse Bundle Adjustment Software Package Based on the Levenberg-Marquardt Algorithm, Tech. Rep. 340, Institute of Computer Science - FORTH, Heraklion, Crete, Greece, available from http://www.ics.forth.gr/ lourakis/sba+, 2004.

[19] M. A. Fischler, R. C. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Commun. ACM 24 (1981) 381–395, ISSN 0001-0782, doi:http://doi.acm.org/10.1145/358669.358692, URL http://doi.acm.org/10.1145/358669.358692.

[20] P. H. S. Torr, A. Zisserman, MLESAC: a new robust estimator with application to estimating image geometry, Comput. Vis. Image Underst. 78 (1) (2000) 138–156, ISSN 1077-3142, doi: http://dx.doi.org/10.1006/cviu.1999.0832.

[21] C. Zach, D. Gallup, J.-M. Frahm, Fast gain-adaptive KLT tracking on the GPU, in: Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08. IEEE Computer Society Conference on, 1 –7, doi: 10.1109/CVPRW.2008.4563089, 2008.

[22] M. Hwangbo, J.-S. Kim, T. Kanade, Inertial-aided KLT feature tracking for a moving camera, in: Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on, 1909 –1916, doi: 10.1109/IROS.2009.5354093, 2009.

[23] J.-S. Kim, M. Hwangbo, T. Kanade, Realtime affine-photometric KLT feature tracker on GPU in CUDA framework, in: Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on, 886 –893, doi:10.1109/ICCVW.2009.5457608, 2009.

[24] J. Shi, C. Tomasi, Good features to track, in: Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on, 593 –600, doi:10.1109/CVPR.1994.323794, 1994.

[25] T. Tommasini, A. Fusiello, E. Trucco, V. Roberto, Making good features track better, in: Computer Vision and Pattern Recognition, 1998. Proceedings. 1998 IEEE Computer Society Conference on, ISSN 1063-6919, 178 –183, doi:10.1109/CVPR.1998.698606, 1998.

[26] A. Fusiello, E. Trucco, T. Tommasini, V. Roberto, Improving Feature Tracking with Robust Statistics, Pattern Analysis and Applications 2 (1999) 312–320, ISSN 1433-7541, URL http://dx.doi.org/10.1007/s100440050039, 10.1007/s100440050039.

[27] T. Thormahlen, N. Hasler, M. Wand, H.-P. Seidel, Merging of Feature Tracks for Camera Motion Estimation from Video, in: Visual Media Production (CVMP 2008), 5th European Conference on, ISSN 0537-9989, 1 –8, 2008.

[28] C. Harris, M. Stephens, A Combined Corner and Edge Detector, in: Proceedings of The Fourth Alvey Vision Conference, 147–151, 1988.

[29] J. L. Bentley, A survey of techniques for fixed radius near neighbor searching., Tech. Rep., Stanford University, Stanford, CA, USA, 1975.

[30] R. Gherardi, A. Fusiello, Practical Autocalibration, in: ECCV 2010: Proceedings of the 11th European Conference on Computer Vision, 2010.