# Robust Rotation and Translation Estimation in Multiview Reconstruction

Daniel Martinec and Tomáš Pajdla

{martid1, pajdla}@cmp.felk.cvut.cz

D. Martinec and T. Pajdla. Robust rotation and translation estimation in multiview reconstruction. In *Proceedings of the Computer Vision and Pattern Recognition conference 2007*, IEEE, Minneapolis, MN, USA, June 2007.

Center for Machine Perception, Department of Cybernetics
Faculty of Electrical Engineering, Czech Technical University
Technická 2, 166 27 Prague 6, Czech Republic
fax +420 2 2435 7385, phone +420 2 2435 7637, www: http://cmp.felk.cvut.cz

# Robust Rotation and Translation Estimation in Multiview Reconstruction

Daniel Martinec Tomáš Pajdla
Center for Machine Perception, Dept. of Cybernetics, Faculty of Elec. Eng.
Czech Technical University in Prague, Karlovo nám. 13, 121 35 Prague, Czech Rep.
{martid1,pajdla}@cmp.felk.cvut.cz

## Abstract

*It is known that the problem of multiview reconstruction can be solved in two steps: first estimate camera rotations and then translations using them. This paper presents new robust techniques for both of these steps. (i) Given pairwise relative rotations, global camera rotations are estimated linearly in least squares. (ii) Camera translations are estimated using a standard technique based on Second Order Cone Programming. Robustness is achieved by using only a subset of points according to a new criterion that diminishes the risk of chosing a mismatch. It is shown that only four points chosen in a special way are sufficient to represent a pairwise reconstruction almost equally as all points. This leads to a significant speedup. In image sets with repetitive or similar structures, non-existent epipolar geometries may be found. Due to them, some rotations and consequently translations may be estimated incorrectly. It is shown that iterative removal of pairwise reconstructions with the largest residual and reregistration removes most non-existent epipolar geometries. The performance of the proposed method is demonstrated on difficult wide base-line image sets.*

## 1. Introduction

This paper[1] makes a step towards automatic reconstruction procedure from a large number of images. This task is difficult and has been extensively studied for the last two decades [7]. In this paper, cameras are assumed to be calibrated [23]. In such a setup, pairwise Euclidean reconstructions can be estimated using RANSAC [18] up to similarities. Given these, reconstruction of the whole scene can be obtained by first registering all camera rotations and then translations using them [26, 15]. Mismatches, i.e. wrong point correspondences, cause several problems in such a two-step reconstruction procedure:

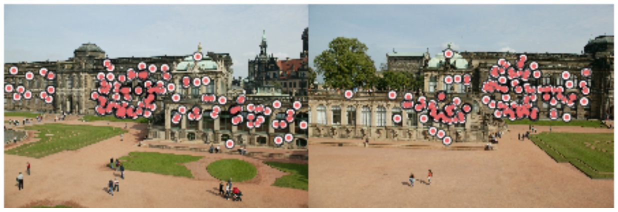

Figure 1. A non-existent epipolar geometry (EG) raised by matching similar structures on different buildings in the Zwinger scene. The shown image pair 37-70 has 163 inliers which are 45% of all tentative matches. It would be extremely difficult to find out that this EG does not exist based on the two images only.

1. A *few mismatches* which survived RANSAC cause no difficulty in rotation registration as the relative rotation is only slightly biased. On the other hand, a single mismatch may cause a complete failure of translation registration when minimizing the maximum reprojection error [10].

2. A *non-existent two-view geometry* may be found when similar or repetative structures appear on different objects, see figures 1 and 9. According to our knownledge, no attempt has been done to handle the presence of non-existent pairwise geometries in either rotation or translation registration.

### Previous Work

Enumeration of multiple view reconstruction methods can be started with factorization methods. First Tomasi & Kanade [25] used factorization on affine cameras. Jacobs [9] improved handling occlusions. Extension for perspective cameras was given in [24]. Projective depths of points, which correspond to the perspective effect, are iteratively improved in iterative factorization methods, e.g. [13].

[1] This research was supported by The Czech Academy of Sciences under project 1ET101210406, by the EU projects eTRIMS FP6-IST-027113 and DIRAC FP6-IST-027787, and by MSM6840770038 DMCM III. Jan Čech from the Czech Technical University provided routines for dense stereo [4]. Our bundle adjustment routine was based on publicly available software [11].

Martinec & Pajdla [14] reformulated Jacob's [9] approach while enhancing numerical stability and applied it for rotation registration using calibrated cameras [15]. Incremental structure from motion can perform in real-time [2].

Recently, methods minimizing the $\mathrm{L}_\infty$-norm appeared. Kahl's method [10] based on Second Order Cone Programming (SOCP), which is a standard technique in convex optimization, estimates both camera translations and point positions given rotations while keeping all points in front of cameras. In this paper, this problem is called *translation registration*. The problem is quasiconvex and thus can be solved via a series of SOCP problems using the bisection method [10]. While [10] may fail due to a single mismatch, method [20] is more robust as it relies on translation directions between camera pairs instead of on individual point correspondences. However, method [20] would probably fail when a non-existent epipolar geometry (EG) is included in the data as the maximum angle between the estimated translation vector and the desired one is minimized. The translation vector of the non-existent EG cannot fit the remaining ones, thus the solution can be obtained with very low precision. This problem cannot be solved by using the uncertainty information [20]. Note that camera translation in figure 1 is estimated with a low uncertainty due to a high number of inliers even when the EG does not exist.

For a wide class of $L_\infty$ problems, Sim & Hartley [21] proved that the set of measurements with the greatest residual must contain at least one outlier. Thus, one could keep throwing out the measurements with the greatest residual. However, it would be very time-consuming on large scenes with hundreds of images and hundreds thousands of points.

This paper proposes (i) a new method for rotation registration. Two variants are presented: using quaternions and using approximate rotations. The latter is simpler and more stable than [15]. (ii) In each pairwise reconstruction, a Gaussian is fitted in the rescaled image space and the most likely mismatches are removed. (iii) Only four points carefully chosen among the remaining points are used to represent them almost equally as all points, thus bringing large speedup and memory savings. (iv) It may happen that the rotation or the translation registration reveals that some EG does not exist. In case rotations were estimated using that EG, they should be reestimated without it as such estimate was biased. It is shown that iterative removal of EGs with the largest residual leads to the removal of most non-existent EGs even for the case of a combination of a least squares and an $L_\infty$ problem.

## 2. Rotation Registration

It is supposed that pair-wise Euclidean reconstructions given up to rotations, translations, and scales are provided. A brief description of how they were obtained for the data presented in this paper is given in section 6.

The pair-wise reconstruction between views $i$ and $j$ describes the relative rotation between the two cameras, $\mathtt{R}^{ij}$, $\mathtt{R}^{ij} \in \mathbb{R}^{3\times3}$, $\mathtt{R}^{ij}$ orthonormal. The problem of *rotation registration* can be formulated as a search for the registered rotations $\mathtt{R}^i$, $\mathtt{R}^i \in \mathbb{R}^{3\times3}$, $\mathtt{R}^i$ orthonormal, $i = 1, \ldots, m$, such that relations among them are given by the relative rotations:

$$\mathtt{R}^j = \mathtt{R}^{ij}\mathtt{R}^i \quad \text{for all } ij \tag{1}$$

$$\mathtt{R}^i \text{ orthonormal for } i = 1, \ldots, m \tag{2}$$

When $m - 1$ relative rotations are known such that they form a tree graph (with views as vertices connected by an edge whenever the relative rotation between the views is known), system of equations (1) is not overdetermined and can be easily solved by fixing the first rotation and chaining the remaining ones.

When at least $m$ relative rotations are given, system (1) becomes overdetermined and an exact solution may not exist due to noise in the data. Thus, we solve it in the least squares while satisfying the orthonormality conditions (2).

A straightforward solution can be obtained using quaternions. Using them, system (1) becomes

$$\dot{r}^j = \dot{r}^{ij}\dot{r}^i \quad \text{for all } ij \tag{3}$$

where $\dot{r}^i$ and $\dot{r}^j$ are the unknown quaternions of the $i^{\mathrm{th}}$ and $j^{\mathrm{th}}$ camera rotation, respectively, and $\dot{r}^{ij}$ is the known relative rotation between cameras $i$ and $j$. Each quaterion can be thought of as a four-vector, similarly as complex numbers can be thought of as two-vectors. Using known manipulations with quaternions [8], each equation in system (3) can be rewritten as

$$\begin{pmatrix} r_0^j \\ r_x^j \\ r_y^j \\ r_z^j \end{pmatrix} - \begin{bmatrix} r_0 & -r_x & -r_y & -r_z \\ r_x & r_0 & -r_z & r_y \\ r_y & r_z & r_0 & -r_x \\ r_z & -r_y & r_x & r_0 \end{bmatrix} \begin{pmatrix} r_0^i \\ r_x^i \\ r_y^i \\ r_z^i \end{pmatrix} = 0_{4\times1} \tag{4}$$

where $\dot{r}^i = r_0^i + \imath r_x^i + \jmath r_y^i + k r_z^i$ and $\dot{r}^{ij} = r_0 + \imath r_x + \jmath r_y + k r_z$ with $\imath$, $\jmath$ and $k$ as imaginary units. From now on, by the $i^{\mathrm{th}}$ quaternion we will mean the four-vector $(r_0^i, r_x^i, r_y^i, r_z^i)^\top$.

There are $4m$ unknowns $r_0^1$, $r_x^1$, $r_y^1$, $r_z^1$, $\ldots$, $r_0^m, r_x^m$, $r_y^m$, $r_z^m$ with constraints (4) for each camera pair $ij$ with a known rotation. System of all $ij$-constraints (4) is sparse, thus it can be solved using, e.g., MATLAB's EIGS. The solution is obtained as a unit vector. The quaternions, from whose parameters the unit vector is composed, are not unit. However, they can be easily made unit by dividing each by its Euclidean length. This conversion is needed as only a unit quaternion has a corresponding rotation. Then, the orthonormality conditions (2) are trivially satisfied.

Due to errors in relative rotations, the individual quaternions in the solution vector have different lengths. Because of this, each $ij$-constraint, i.e. the four equations (4) demanded by the $ij$-relative rotation, has a different influence (weight), which is approximately proportional to the lengths of the resulting $i$- and $j$-quaternions. The shorter are the two four-vectors, the smaller attention has to be given to the four equations. As a consequence, the difficult partial reconstructions, i.e. those which significantly differ from the remaining ones, are given small attention. They get weighted down to better fit the majority of constraints.

**Remark.** A solution would be to add the constraint on unit lengths of all resulting quaternions:

$$(r_0^i)^2 + (r_x^i)^2 + (r_y^i)^2 + (r_z^i)^2 \quad = \quad 1 \qquad \text{for all } i \quad (5)$$

Unfortunately, sofar no satisfactory way for solving a linear system with quadratic equations like (5) is known.

### 2.1. Registration using Approximate Rotations

An alternative way is to solve system (1) without satisfying the orthonormality constraints (2). In fact, system (1) consists of three smaller subsystems

$$\mathbf{r}_k^j - \mathtt{R}^{ij}\mathbf{r}_k^i \quad = \quad 0_{3\times 1} \qquad \text{for all } ij \qquad (6)$$

for $k = 1, 2, 3$, where $\mathbf{r}_k^i$ are columns of $\mathtt{R}^i$, $\mathtt{R}^i = [\mathbf{r}_1^i \mathbf{r}_2^i \mathbf{r}_3^i]$. The solution for approximate rotations can be found as the best three linearly independent least squares solutions to system (6). System (6) is sparse and thus can be solved, e.g., using MATLAB's EIGS. See [14] for details on a solution to a similar system. The orthonormality constraints (2) are enforced by projecting the approximate rotation to the closest rotation in the Frobenius norm using SVD [15].

Compared to [15], no auxiliary variables rotating the partial reconstructions to the global coordinate system are needed. Thus, this solution is simpler and faster. We observed that it is also more stable.

Results got improved when $ij$-equations (6) corresponding to the $ij$-EG were reweighted by $\min(a, 400)$, where $a$ is the number of inliers in the $ij$-EG. Solution to (6) can be found very efficiently. Rotation registration of 259 views using 2049 relative rotations in the Tête scene (see figure 7) using MATLAB's EIGS took only 0.37 seconds.

#### Comparison with Quaternions

On the Head scene [15] (not shown here due to the lack of space and its simplicity), the ratio between the maximum and minimum quaternion lengths from (4) was 5.04. On the other hand, the norms of the $3 \times 3$ matrices found by (6) were very close to each other (less than 1%). Norms of individual 3-vectors were even closer (less than 0.1%). The maximum Frobenius norm of the difference between the relative rotation and the relative rotation after registration, $||\mathtt{R}^{ij} - \mathtt{R}^j {\mathtt{R}^i}^\top||$, was 1.98 and 0.37 for quaternions and approximate rotations, respectively. The fact that the first number is very close to the maximum possible norm (which is 2) shows that the method using quaternions in not usable in practice. In the rest of the paper, only approximate rotations are used.

The reason why the least squares solution is worse for quaternions than for approximate rotations is perhaps the following. When searching for the most suitable rotations, it is easier to search in the space of approximate rotations (all $3 \times 3$ matrices) than in the space of rotations (quaternions). The latter is a small manifold included in the first space. In both cases, a solution that well satisfies all constraints on relative rotations is searched for.

The inconsistencies in constraints prove as (i) getting off the manifold and (ii) changing lengths of vectors representing individual rotations. The approximate rotations "use" both (i) and (ii) "effects" and thus are in higher accordance with all constraints as they can be off the manifold. (It is not far from the manifold, as will be shown on experiments.) For quaternions, (i) is not possible. This thus causes a bigger pressure on (ii), i.e. deformation of quaternion lengths. As a consequence, the constraints with very short quaternions are given a very low attention, which is the undesired side effect. This effect happens with approximate rotations as well but with a lower order differences, as shown above.

## 3. Data Compression and Clarification

We found out that it is possible to represent each partial reconstruction using four points only while capturing the overall geometry well. The idea comes from projective factorization using perspective cameras [24]. Projection matrices of a partial reconstruction, $\mathtt{P}$, multiplied with all points reconstructed in that partial reconstruction, $\mathbf{X}$, form so-called *rescaled measurement matrix* $\lambda\mathbf{x} = \mathtt{P}\mathbf{X}$, where the measured image points $\mathbf{x}$ are rescaled by depths $\lambda$ element-wise, $\lambda_p^i \mathbf{x}_p^i = \mathtt{P}^i \mathbf{X}_p$ [24]. Here we work with projected points $\mathtt{P}\mathbf{X}$ instead of the rescaled measured image points $\lambda\mathbf{x}$. It is equivalent when there is no noise in the data. Usage of the projected points has the advantage that the rescaled measurement matrix is less affected by noise when cameras are well estimated (which is often the case).

The desired four points are chosen so that the corresponding four columns in $\mathtt{P}\mathbf{X}$ represent the four dimensional subspace spanned by all columns of $\mathtt{P}\mathbf{X}$. Thus, the necessary condition is that the chosen four columns are linearly independent. There are many such quadruplets, therefore an additional criterion is needed. Before formulating it, a criterion for identifying mismatches will be given.

### 3.1. Identifying Mismatches

True matches connect one or several surfaces visible in an image pair. True matches connecting the same surface

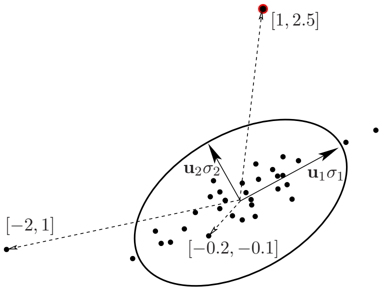

Figure 2. Each point represents a vector in a two-dimensional vector space (here plane). The ellipse characterizes the fitted Gaussian to the centered data. The ellipse center is in $[0, 0]$ and its half-axes are $\mathbf{u}_1\sigma_1$ and $\mathbf{u}_2\sigma_2$ where $[\mathbf{u}_1\mathbf{u}_2]\,\mathrm{diag}(\sigma_1, \sigma_2)\mathtt{V}^\top$ is the "economy size" SVD factorization of $\mathtt{P}\mathbf{X}$. It is drawn for the 2D case instead of 6D. The ellipse shape characterizes the most of the data mass. The ML mismatches are the most distant points from the ellipse center w.r.t. the coordinate system given by ellipse half-axes. The coordinates are drawn at three points. These are also rows of the $\mathtt{V}$ matrix. Although the leftmost point is the most distant from $[0, 0]$, the upmost point is a more likely mismatch as its distance is larger in the ellipse coordinate system: $||[1, 2.5]|| > ||[-2, 1]||$.

are (i) localized close to one another in the images and (ii) have similar depths. As a result, true matches form clusters in the rescaled image space while mismatches are far from the remaining data due to incorrect depths. To ensure that the clusters are formed, the images of the scene must contain sufficiently large surfaces on which multiple matches forming a cluster could be detected and matched. There are scenes which do not satisfy this assumption like, e.g., many tiny branches of a tree. However, such scenes would hardly be matched by any algorithm, thus the assumption on scenes containing sufficiently large surfaces is not so restrictive in practice. Any clustering algorithm could be used to find individual surfaces corresponding to the clusters. Matches contained in no or small clusters could be thrown away as most likely (ML) mismatches. Nevertheless, in this work we did something much simpler.

In this paper, the main purpose was to reliably remove all mismatches as the $L_\infty$-norm, i.e. the maximum reprojection error, is minimized in translation estimation [10], which may be hundreds of pixels due to a single mismatch. Thus, to get a reasonable estimate using [10], all (or at least most) mismatches have to be removed. We observed on the presented scenes that either an EG was non-existent or its inliers were contaminated by a low amount (less than $\epsilon = 25\%$) of mismatches.[2] A Gaussian was fitted to the data in the rescaled image space and a prescribed amount, $\epsilon$, of most distant points was thrown away as the ML mismatches, see figure 2. Localizing the largest cluster (or a set of large clusters) by a single Gaussian is justifiable when the inter-cluster distances are relatively small compared to the distances to mismatches. This simple way worked well on scenes presented in the paper and many others.

```
for k = 4:-1:1
  [U,s,v] = svd(R*R’,0);              % svd of a long matrix
  S = sqrt(diag(s(1:k,1:k)));   % using svd of a short one
  V = ((diag(1./S)*v(:,1:k)’)*R)’;  % R = U*diag(S)*V’
  len = V’.^2; if k > 1,
    len = sum(len); end  % squared lengths of rows of V
  best = find(len == max(len));
  p(k) = best(1);
  C = R(:,p(k));                         % the chosen column
  R = R - C*(pinv(C)*R);               % subtract its span
end
```

Figure 3. Choosing the four most different points representing a partial reconstruction. In MATLAB code, variable R contains the rescaled measurement matrix, $\mathtt{P}\mathbf{X}$. Indices of the chosen four points are stored in variable p.

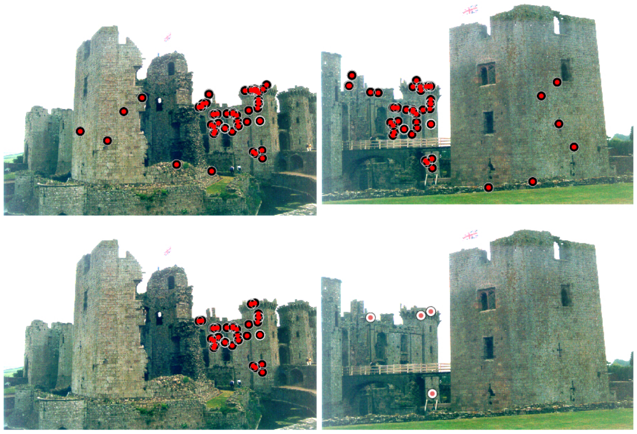

Figure 4. Image pair 19-22 in the Raglan scene. Points satisfying EG of this image pair (top row). Non-mismatch candidates identified before the multiview registration (bottom left). The four points used for translation registration (bottom right).

After estimating the data mean and subtracting it from all vectors, the covariance matrix of the Gaussian is obtained using SVD. The ML mismatch is the most distant point in the coordinate system given by the Gaussian covariance matrix. Its corresponding row in matrix $\mathtt{V} \in \mathbb{R}^{n\times 4}$ has the largest norm, where $\mathtt{P}\mathbf{X} = \mathtt{U}\,\mathrm{diag}(\sigma_1, ..., \sigma_4)\mathtt{V}^\top$ is the "economy size" SVD decomposition. It is illustrated on figure 2, see the explanation there.

All $\epsilon$ ML mismatches can be either (i) removed at once or (ii) one by one while refitting the Gaussian after each ML mismatch removal from $\mathtt{P}\mathbf{X}$. The latter way was used in this work as a higher stability can be expected. The SVD decomposition can be carried out efficiently, see lines 2–4 of the algorithm in figure 3. Note that the most time consuming operation is SVD applied to a $6 \times 6$ matrix irrespective of

[2] When more mismatches are present, such EG is likely to be detected and removed after translation registration, see section 5.

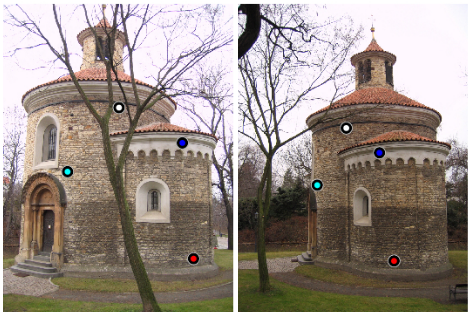

Figure 5. Four most different points chosen after the removal of $\epsilon = 25\%$ ML mismatches. Image pair 41-48 in the St. Martin rotunda is shown. The points lie in different depths and thus capture the 3D geometry of the image pair well.

the number of points.[3] An example of identified ML mismatches at $\epsilon = 25\%$ is shown in figure 4.

**Normalization.** As the procedure is done on the rescaled measurement matrix, i.e. on rescaled image data, the image coordinates should be normalized to be close to one [7] and the resulting $\mathtt{P}\mathbf{X}$ should be balanced by rescaling its columns and row triplets, as described in [24].

ML mismatches are identified prior to rotation registration. Doing it aferwards based on the partial reconstruction reestimated using the registered rotations might be incorrect as the estimate of the registered rotations may be severely corrupted due to non-existent EGs (see section 5).

As a side effect of the removal of all $\epsilon$ ML mismatches, many true matches are removed as well. Nevertheless, it is not a problem as the left data constrain the multiview reconstruction sufficiently, as will be shown in section 4.

### 3.2. Reconstruction Represented by Four Points

After $\mathtt{P}\mathbf{X}$ has been cleared of mismatches, the same Gaussian fitting technique is used for choosing the four points for representation of the partial reconstruction. If the data contains a mismatch, the most different point is the ML mismatch. However, after the data was cleared of mismatches, the most different point is the best inlier for representing the geometry. The four points are found in the following way. After identifying the most different point, the whole matrix is projected onto the span of the chosen column and subsequently subtracted from $\mathtt{P}\mathbf{X}$. This is repeated four times. The procedure is summarized in figure 3.

The chosen points lie in different depths as well as the ML mismatch does. However, here it is advantageous as the different depths capture the 3D geometry of the two images well, as can be seen in figure 5. Note that if the data contained any previously not removed mismatch, it would very likely appear among the chosen four points.

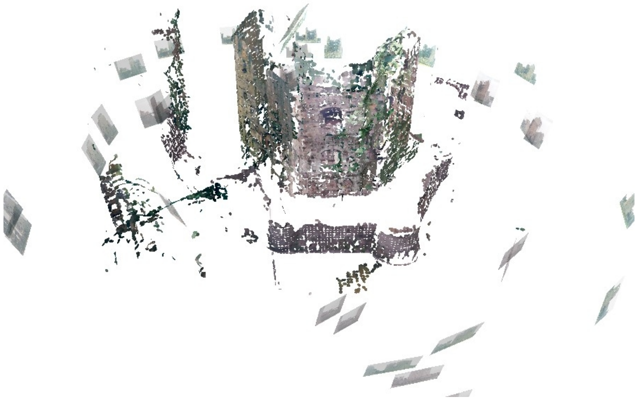

Figure 6. The Raglan scene: an overall view with a bridge on the left.

**Remark.** The ML mismatch identification and the choice of the four representative points work for projective reconstruction as well since product $\mathtt{P}\mathbf{X}$ depends on images and not on the choice of a reference frame. The depths do not have to be positive, nor the cameras calibrated. The only thing that matters is how the columns corresponding to points are situated in the subspace generated by normalized columns of $\mathtt{P}\mathbf{X}$ (cf. the note on normalization above).

## 4. Translation Registration

In [15], translations and points in each partial reconstruction were estimated using [10]. Then, all partial reconstructions were refined together using bundle adjustment (BA) while keeping rotations registered. Unlike in [15], in this work, no such intermediate BA is performed. The reason is that the precision of the proposed rotation registration is satisfactory when combined with the robust point sampling explained above.

Method [10] is applied only once on the data from all partial reconstructions. However, each partial reconstruction is represented by four points chosen as explained in section 3.2 instead of almost all points. Thus, it is much faster. After translation registration, BA on all data was done and dense reconstructions were obtained using [6, 4].

The Raglan scene [19] was captured on 46 images, 238 EGs were found (see details in section 6). When [10] was applied on all points in all partial reconstructions (186131 points in total), the maximum residual of 98.57 pixels was obtained in 3 hours and 6 minutes. When using only the four representative points, the maximum residual of 98.46 pixels was obtained in 4.68 seconds. This demonstrates that the four points represent geometry of the individual reconstructions well while achieving a huge speedup (of factor 2385 at this particular scene). When using quadruplets chosen from the non-mismatch candidates at $\epsilon = 25\%$, the obtained error decreased to 22.30 pixels. It was manually

[3] If needed, even a more efficient implementation is possible using incremental SVD [3] instead of the standard SVD.

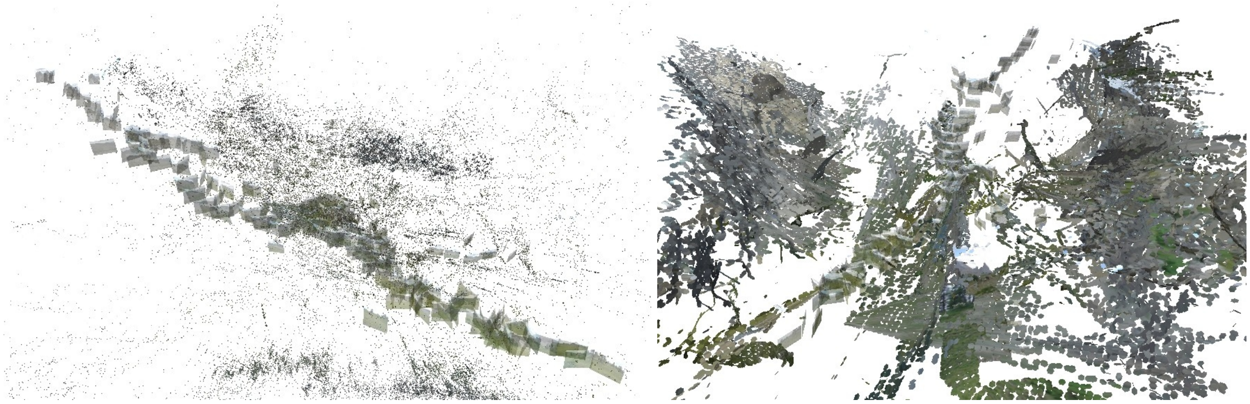

Figure 7. The Tête scene. The profile view - top of the mounain is on the left-hand side (left). View from the valley up (right).

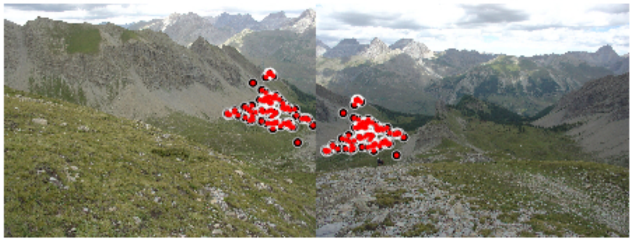

Figure 8. The Tête scene. View from top into the left (left) and right valley (right).

verified on several quadruplets with largest residuals [21] that none of them included a mismatch, although there were many in the data, see figure 4. When using the intermediate BA with rotations kept registered (see the beginning of this section) before applying [10] on all image pairs, the maximum error dropped to 12.09 pixels. The reconstruction is shown in figure 6.

The Tête de Plate Longe (shortly Tête) scene (259 images, 2049 EGs) was reconstructed with the largest residual of 38 pixels in 74 seconds. Manual inspection verified that no mismatch was present. We tried several strategies for reweighting equations (6) based on residuals in individual partial reconstructions, however no general strategy was found (the best trial dropped to 27). See figures 7 and 8.

## 5. Robust Rotation Estimation

It turned out that even if the found relative rotation is close to the desired one in the Frobenius norm, i.e. $||\mathtt{R}^{ij} - \mathtt{R}^{j}\mathtt{R}^{i^{\top}}||$ is small, the partial reconstruction with rotation replaced by the found rotation (and with translations reestimated [10]) may still produce large residua. This effect could be reduced by using rotation uncertainties [20]. However, a more serious problem is when the rotation registration is contaminated by some non-existent EG. Fortunately, it has been observed that the points from such an EG have large residua after the rotation and translation registration. Thus, it is straightforward to remove such partial reconstruction and reestimate rotations and translations.

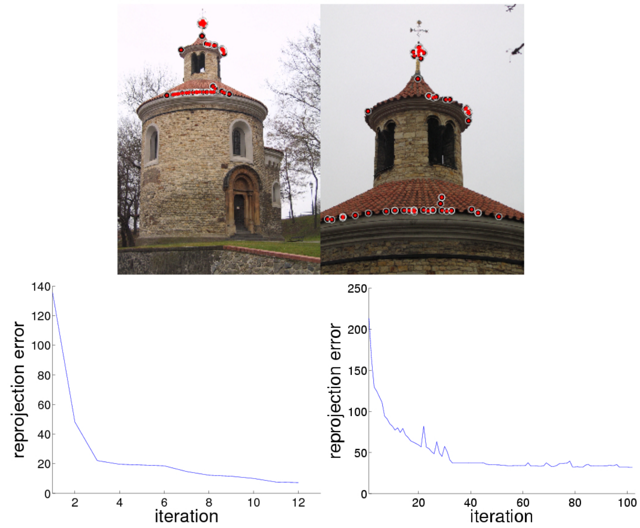

Figure 9. Iterative removal of EGs with the largest residual. One of 13 non-existent EGs in the St. Martin rotunda: image pair 4-119 (top row). Decreasing of the maximum residual is shown for the St. Martin (bottom left) and the Zwinger scene (bottom right).

As mentined in the introduction, it was proved [21] for a wide class of $L_\infty$ problems that the set of measurements with the greatest residual must contain at least one outlier. However, it is not the case of the least squares rotation estimate presented here. Nevertheless, it will be shown on two scenes that this property holds in practice even for the $L_\infty$ problem [10] initiated by the least squares solution to (6).

Our least squares solution to (6) provides quite a good estimate even when many relative rotations came from non-existent EGs (in the Zwinger scene, more than 156 (8%) EGs were non-existent). The reason why it works so well is perhaps that the existent EGs support each other while

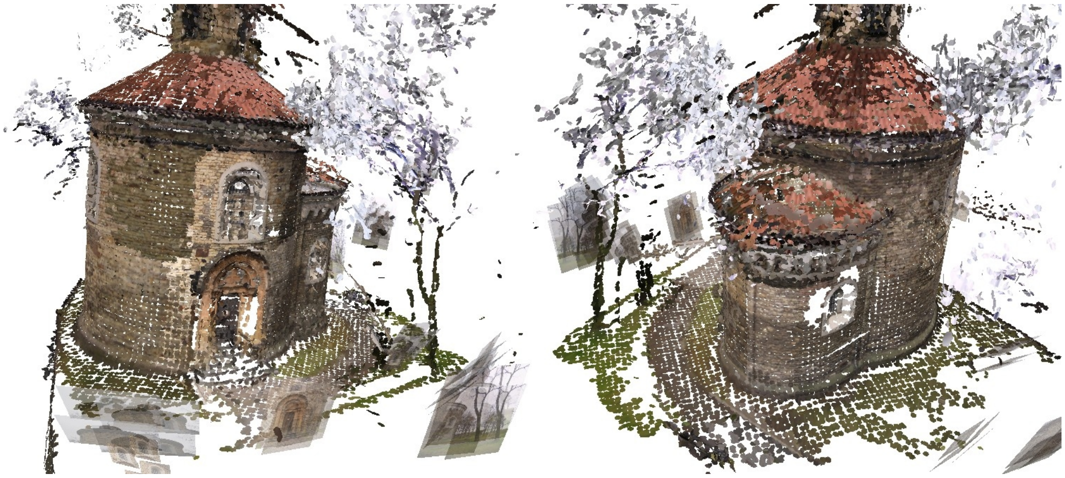

Figure 10. The St. Martin rotunda. Front and back view of the dense reconstruction with some cameras shown as image planes. Note the details as the tree and the footpath around the building. The clouds come from tiny branches.

the non-existent ones rather do not as they raised almost randomly and independently. However, each non-existent EG deteriorates the quality of the solution.

Unlike [21], we do not remove single points but whole partial reconstructions, in which some of the four points reached the maximum residual. This brings an additional speedup besides the compression to four points.

The St. Martin rotunda (124 images, 1670 EGs) was reconstructed with the mean/maximum residual of 1.5/7.66 pixels after 11 iterations of removing EGs with the largest residual and rotation and translation reestimation, see figure 9. There were 13 non-existent EGs detected (manually checked), one of which is shown in figure 9. In some iterations, more EGs with the same maximum residual (at some of the four points) were removed. The dense reconstruction using [6, 4] in figure 10 demonstrates that the proposed method reaches a high precision. The surface parts from different views shown in different colors due to varying lightning conditions fluently connect to each other.

On the Zwinger scene (199 images, 1954 EGs), method [10] produced error of 229 pixels in 51 seconds. There were many non-existent EGs, see figure 9. It seems that after the maximum residual dropped below 35 pixels (at iteration 51, 123 EGs removed), it was hard to improve the precision more. The reconstruction shown in figure 11 was done using the result of iteration 100 (156 EGs removed, error 31 pixels). It turned out after manual inspection that still some non-existent EGs remained in the data.

## 6. Experiments

In experiments reported here, pairwise image matching was done with Local Affine Frames [16] constructed on intensity and saturation MSER regions, LaplaceAffine and HessianAffine [17] interest points. Additional matches were found using SIFT features [12]. Only some image pairs were matched on the large Tête and Zwinger scenes. Details on the used heuristic will be published elsewhere.

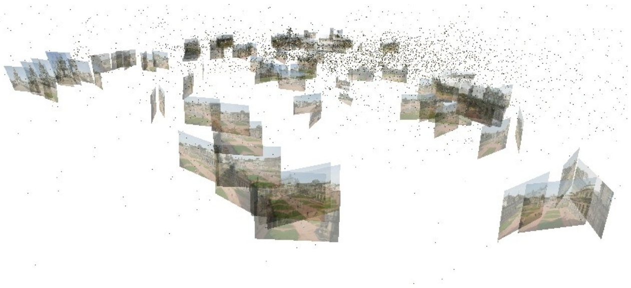

Figure 11. The Zwinger scene. Only quadruplets of points representing EGs are shown. Non-existent EGs with many mismatches between repetitive structures on building fasades are still present.

The six-point RANSAC [23] with plane detection [5] was run on the matched pairs and the focal length was calibrated as the mean of all estimates. Then, BA on all pairs with focal lengths kept equal (but varying) was run, followed by the five-point RANSAC [18] and track merging. Radial distortion was not removed from the images.

Due to a few repetitive structures in the Raglan scene and a huge amount of them in the Zwinger scene, RANSAC on many-to-many correspondences had to be used, details will be reported in another publication.

It was desired to forbid all pairs not suitable for dense stereo. These are especially pairs with (nearly) coinciding camera centers forming a panorama. If some pair should fit a panorama model, it must fit a weaker homography model at least so well. Thus, only pairs with 90% inliers lying on a (dominant) plane need to be checked for being a panorama, the remaining ones cannot be a panorama. Fitting the panorama model was started by making the two camera centers coincident by setting them to their mean. Then BA constrained to keep the camera centers equal was run. Many panoramas were successfully detected but some not, which can be seen on the Tête scene in figure 7right.

## 7. Summary and Conclusions

A practical method for automatic reconstruction was presented. It was shown to work on hundreds of images. 99.68% of the measurement matrix of the Tête scene were missing due to occlusions. There is no chance for any factorization method to deal with such a large amount of missing data. The whole algorithm uses only two-view correspondences except for the final BA, which starts with low errors (from 7 to 30 pixels) and thus can change the overall geometry only slightly. This means that the overall geometry is mostly determined by the rotation and translation registration. The rotation registration takes a fraction of a second on hundreds of images and the translation registration takes around a minute. Both should be repeated when the data is contaminated by non-existent EGs. Even in this case, the total running time is in the order of minutes, which is a fraction of the time spent by BA in the incremental structure from motion (SfM) [22]. Note that images of the presented scenes are very sparsely captured compared to [22].

Closed image sequence is a problem for any incremental SfM as the first and the last camera positions get misaligned. In our approach, using all EGs at once has the advantage that many closed loops among images can be handled.

It has been shown that the presented method is robust to some contamination by non-existent EGs. The contamination in the Zwinger scene is an extreme one: hundreds of non-existent EGs, most of the existent EGs have mismatches on repetitive structures. To reconstruct this scene better, detecting non-existent EGs prior to rotation registration seems to be needed.

More reconstructed scenes can be seen at [1].

## References

[1] http://cmp.felk.cvut.cz/˜martid1/demoCVPR07.

[2] A. Akbarzadeh et al. Towards urban 3D reconstruction from video. In *3DPVT*, University of North Carolina, Chapel Hill, USA, June 2006. CD-ROM.

[3] M. Brand. Incremental singular value decomposition of uncertain data with missing values. In *ECCV*, 2002.

[4] J. Čech and R. Šára. Efficient sampling of disparity space for fast and accurate matching. In *Proc. BenCOS Workshop CVPR*, 2007. To appear.

[5] O. Chum, T. Werner, and J. Matas. Two-view geometry estimation unaffected by a dominant plane. In *CVPR*, vol. 1, pp. 772–779, 2005.

[6] H. Cornelius, R. Šára, D. Martinec, T. Pajdla, O. Chum, and J. Matas. Towards complete free-form reconstruction of complex 3D scenes from an unordered set of uncalibrated images. In *SMVP/ECCV*, vol. LNCS 3247, pp. 1–12, Prague, Czech Republic, May 2004.

[7] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge University, 2nd edition, 2003.

[8] B. K. P. Horn. Closed form solution of absolute orientation using unit quaternions. *Journal of the Optical Society A*, 4(4):629–642, April 1987.

[9] D. Jacobs. Linear fitting with missing data: Applications to structure from motion and to characterizing intensity images. In *CVPR*, pp. 206–212, 1997.

[10] F. Kahl. Multiple view geometry and the $L_\infty$-norm. In *ICCV05*, pp. II: 1002–1009, 2005.

[11] M. Lourakis and A. Argyros. The design and implementation of a generic sparse bundle adjustment software package based on the levenberg-marquardt algorithm. Technical Report 340, Institute of Computer Science - FORTH, Heraklion, Crete, Greece, Aug 2004.

[12] D. Lowe. Distinctive image features from scale-invariant keypoints. In *International Journal of Computer Vision*, vol. 20, pp. 91–110, 2003.

[13] S. Mahamud and M. Hebert. Iterative projective reconstruction from multiple views. In *CVPR*, 2000.

[14] D. Martinec and T. Pajdla. 3D reconstruction by fitting low-rank matrices with missing data. In *Proc CVPR*, vol. I, pp. 198–205, San Diego, CA, USA, June 2005.

[15] D. Martinec and T. Pajdla. 3D reconstruction by gluing pairwise Euclidean reconstructions, or "how to achieve a good reconstruction from bad images". In *3DPVT*, University of North Carolina, Chapel Hill, USA, June 2006. CD-ROM.

[16] J. Matas, Š. Obdržálek, and O. Chum. Local affine frames for wide-baseline stereo. In *ICPR(4)*, pp. 363–366, 2002.

[17] K. Mikolajczyk et al. A Comparison of Affine Region Detectors. *IJCV*, 2005.

[18] D. Nistér. An efficient solution to the five-point relative pose. *PAMI*, 26(6):756–770, June 2004.

[19] F. Schaffalitzky and A. Zisserman. Multiview matching for unordered image sets, or, "how do i organize my holiday snaps?". In *ECCV*, 2002.

[20] K. Sim and R. Hartley. Recovering camera motion using $L_\infty$ minimization. In *CVPR*, vol. 1, pp. 1230–1237, New York , USA, June 2006.

[21] K. Sim and R. Hartley. Removing outliers using the $L_\infty$ norm. In *CVPR*, vol. 1, pp. 485–494, New York , USA, June 2006.

[22] N. Snavely, S. Seitz, and R. Szeliski. Photo tourism: Exploring photo collections in 3D. *ACM Transactions on Graphics (SIGGRAPH Proceedings)*, 25(3):835–846, 2006.

[23] H. Stewénius, D. Nistér, F. Kahl, and F. Schaffalitzky. A minimal solution for relative pose with unknown focal length. In *CVPR*, vol. 2, pp. 789–794, 2005.

[24] P. Sturm and B. Triggs. A factorization based algorithm for multi-image projective structure and motion. In *ECCV96(II)*, pp. 709–720, 1996.

[25] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *IJCV*, 9(2):134–154, November 1992.

[26] M. Uyttendaele et al. High-quality image-based interactive exploration of real-world environments. *CG&A*, 24(3):52–63, May/June 2004.