CENTER FOR
MACHINE PERCEPTION

CZECH TECHNICAL
UNIVERSITY

REPRINT

3D Reconstruction by Gluing Pair-wise Euclidean Reconstructions, or "How to Achieve a Good Reconstruction from Bad Images"

Daniel Martinec and Tomáš Pajdla

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

D. Martinec and T. Pajdla. 3D reconstruction by gluing pair-wise euclidean reconstructions, or "how to achieve a good reconstruction from bad images". In *Proceedings of the 3D Data Processing, Visualization and Transmission conference (3DPVT)* , University of North Carolina, Chapel Hill, USA, June 2006.

Available at
ftp://cmp.felk.cvut.cz/pub/cmp/articles/martinec/Martinec-3DPVT2006.pdf

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

3D Reconstruction by Gluing Pair-wise Euclidean Reconstructions, or "How to Achieve a Good Reconstruction from Bad Images"

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

This paper presents a new technique for estimating a multi-view reconstruction given pair-wise Euclidean reconstructions up to rotations, translations and scales. The partial reconstructions are glued by the following three step procedure: (i) Camera rotations consistent with all reconstructions are estimated linearly. (ii) All the pair-wise reconstructions are modified according to the new rotations and refined by bundle adjustment while keeping the corresponding rotations same. (iii) The refined rotations are used to estimate camera translations and 3D points using Second Order Cone Programming by minimizing the L_∞-norm. We introduce a new criterion for evaluating importance of an epipolar geometry in influence on the overall 3D geometry. The estimated importance is used to reweight data in the above algorithm to better handle unequiponderantly captured images. The performance of the proposed method is demonstrated on difficult wide base-line image sets.

1. Introduction

This paper makes a step towards automatic reconstruction procedure providing a high quality reconstruction from a difficult image set. This task is difficult and has been extensively studied for last two decades [8]. In this paper we particularly focus on the following problems:

- An extreme occurrence of the missing data. In practice it may happen that there are no points visible in more than two images in a (sub)set of images, see figure 1.
- Degenerate situations like large planes in the image may prevent RANSAC [8] from choosing the non-degenerate full 3D model but with smaller support than of the model given by homography [4]. Another degenerate situation arises when images are taken from one place just by zooming or rotating the camera.
- Some parts of the object may be captured on much more photographs than some other parts. We have observed that in such a case, without a good estimate of the focal length, the standard bundle adjustment [8] may break the reconstruction into discontinuous parts, see figures 2 and 4.

Figure 1: An extreme case of the missing data: (top) image correspondences shown in three colors corresponding to three image pairs. There is no point visible in three images. (bottom) Reconstruction of the boxes from two views

Note that one might argue that in order to obtain a good model, one should take appropriate images rather than accept bad ones, and that even a non-expert can be easily guided to take better images. Nevertheless, cases where bad images have to be accepted may indeed be relevant sometimes, e.g. when modeling a building or monument that can not be easily photographed from all around.

*This research was supported by The Czech Academy of Sciences under project 1ET101210406 and by the EU projects eTRIMS FP6-IST-027113 and DIRAC FP6-IST-027787. Richard Szeliski from Microsoft Research provided the ICCV'05 Contest data. Jana Kostková from the Czech Technical University provided routines for dense stereo. Our bundle adjustment routine was based on publicly available software [12].

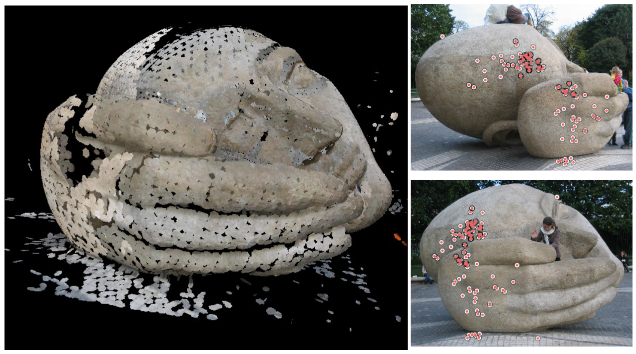


Figure 2: Incorrect reconstruction provided by the standard bundle adjustment [8] for non-equiponderantly captured images of the Head set (left) low amount of correspondences on the back of the hand. Only inliers w.r.t. the final 3D-model shown (right)

Previous Work

Enumeration of multiple view reconstruction methods can be started with factorization methods. First Tomasi & Kanade [24] used factorization on affine cameras. Jacobs [10] improved handling occlusions. Extension for perspective cameras was given in [23]. Projective depths of points, which correspond to the perspective effect, are iteratively improved in iterative factorization methods, e.g. [14]. Martinec & Pajdla [15] reformulated Jacob's [10] approach and enhanced numerical stability. In [23, 20, 15], the perspective effect is handled using epipolar geometry (EG) while other methods like [6] use trifocal tensors.

In all the above named methods, points visible in at least three images were used to glue partial reconstructions (estimating an EG can be viewed as equivalent to having the 3D reconstruction). For the case when each point is visible in just two images, like in fig. 1, it seems that only methods of Avidan & Shashua [3] and of Uyttendaele et al. [26] might be used. In [3], partial projective reconstructions represented by fundamental matrices were glued together while minimizing an algebraic error on parameters of the fundamental matrices.

It is not clear how well would method [26] work in a wide base-line setup as the image set in [26] was densely captured on a video enabling reliable detection of self-intersecting paths and vanishing points, which were used to make rotations consistent. Method [15] must fail because the depths estimated by individual EGs cannot be made consistent as this consistency can only be achieved via points shared by multiple EGs (i.e. points visible in at least three images). Gluing 3P3 problems is impossible when no point is visible in three views.

Recently, methods minimizing the L_∞-norm appeared in vision community. In this paper we use Kahl's method [11] based on Second Order Cone Programming (SOCP) which is a standard technique in convex optimization. Method [11] is capable of estimating both camera translations and point positions given rotations.

There are several methods on distinguishing between full 3D EG and degenerate cases like planes, see, e.g., [25, 4].

This paper proposes a new algorithm with the advantage that no point visible in three or more views is required. It consists of three steps: (i) a linear estimate of consistent rotations (ii) refinement of the estimate and (iii) camera translation and point recovery using SOCP [11]. Its first step is a variation on [15] for the Euclidean case. In our approach, the partial reconstructions are glued via cameras. All partial reconstructions are glued at the same time thus exploiting all data equiponderantly.

Our method differs from [15] in that the estimated transformation between the coordinate system of a partial reconstruction and the coordinate system of the reconstruction of all cameras is more simple: only rotation, translation and scale are needed instead of full projective 4×4 homographies in [15]. Second advantage of Euclidean over projective reconstruction is that no Euclidean upgrade [19, 9] is needed. Euclidean upgrade becomes a very difficult task on data like presented here. Even if the projective reconstruction is successfully transformed so that a reasonable subset of points gets in front of cameras, the internal camera parameters get rather far from what is desired (e.g. square pixel) making bundle adjustment prone to stucking in local minima. In [15], the algebraic (SVD) error is minimized instead of the reprojection error. Minimizing an algebraic error is also the case of [3]. Compared to these methods, minimizing inconsistency between rotations in the Euclidean space promises reaching a higher stability.

It is known that rotation can be estimated first and translation can be estimated using it afterwards [26]. In [26], differences between rotations parameterized using quaternions were non-linearly minimized while using some additional constraints like vanishing points. In our method, 3×3 matrices are used to parameterize rotations. Although these matrices describe the class of all homographies, the rotation obtained as the closest rotation to such homography in the least squares gives results sufficient for our task, as rotations are subsequently refined in step (ii).

The new method on linear estimation of consistent camera rotations will be explained in section 3.2 and rotation refinement in section 3.3. Obtaining consistent camera translations and scales comes in section 3.4. A new criterion on influence of an EG on the overall shape will be introduced in section 4 and its application will be shown in section 4.1. Experiments are reported in section 5. Discussion and summary in sections 6 and 7, respectively, conclude the paper.

2. Problem Formulation

Suppose m images captured using a standard camera with focal lengths known[1] up to an unknown overall scale factor. Points of interest are found in all images and matched between all image pairs using a similarity measure (see more details on our experiments in section 5) There are mismatches in image measurements. The goal is to recover cameras and 3D points.

3. The New Method

The 6-point RANSAC [22] is applied to all image pairs. When the two corresponding focal lengths differ, one of the images is rescaled so that the focal lengths become the same. The overall scale of the focal length is then estimated as the mean of the estimates given by the 6-point algorithm weighted by the square of the EG support. Then, the 5-point algorithm [18] is run on all image pairs.

3.1. RANSAC on EG and a Dominant Plane

An epipolar geometry unaffected by a dominant plane is found using [4]. The inliers are used as the pool for drawing samples in calibrated RANSACs. This scheme is applied to the 6-point algorithm [22] as well as to the 5-point algorithm [18]. Due to unstability of estimate of the focal length [22], the degenerate samples (all points in the dominant plane provided by [4][2]) should be detected and thrown out. If all points lie in a plane, e.g. two images in a panorama, the correct model cannot be estimated. Thanks to the small amount of outliers in the pool, the 5-point RANSAC has a bigger chance to find the correct non-degenerate EG, especially with a substantial error in the focal length. It can be seen in figure 3 that estimates of the overall scale of the focal length is quite unreliable (see also beginning of section 4).

3.2. Consistent Rotations

Let $\mathtt{A}^{\mathbf{i}}$ denote the submatrix of $\mathtt{A}$ composed of elements in rows $\mathbf{i}$. Omitting superscript means taking all rows. Suppose T pair-wise Euclidean reconstructions are given for camera indices $\mathbf{i}_1, \mathbf{i}_2, \ldots, \mathbf{i}_T$ where $\mathbf{i}_t \in \{1, 2, ..., m\}$, $|\mathbf{i}_t| = 2$ for each t.[3] Let the cameras of the t^{th} reconstruction be denoted as $\tilde{\mathtt{P}}_t$, $\tilde{\mathtt{P}}_t \in \mathbb{R}^{6\times 4}$. Each reconstruction is generally in a different coordinate system. It has been shown in [15, Eq. (4)] that the coordinate systems are related by homographies, $\mathtt{H}_t$, which can be linearly estimated

[1] e.g., from the EXIF header of the JPEG file

[2] One might check if all the points lie on another (smaller) plane. Note that such samples cannot win in RANSAC.

[3] Here, only equations for pair-wise reconstructions are shown. Equations for triples, etc. are similar.

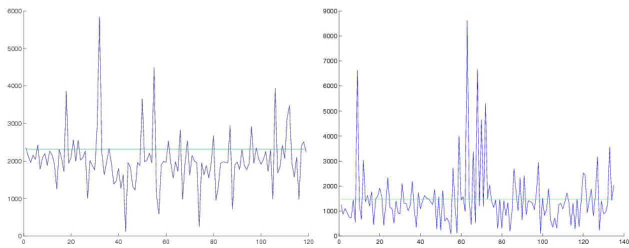


Figure 3: Focal length estimates from the 6-point algorithm used as described in section 3.1 on the Head and Contest sets. The horizontal line corresponds to the estimate of the overall focal length

together with a set of all cameras, $\mathtt{P} \in \mathbb{R}^{3m\times 4}$, in the same (global) coordinate system:

$$\begin{aligned} \tilde{\mathtt{P}}_1\mathtt{H}_1 &= \mathtt{P}^{\mathbf{i}_1} \\ &\vdots \\ \tilde{\mathtt{P}}_T\mathtt{H}_T &= \mathtt{P}^{\mathbf{i}_T}. \end{aligned} \tag{1}$$

For cameras are calibrated, after denoting indices of the t^{th} pair-wise reconstruction as i and j, $\{i, j\} = \mathbf{i}_t$, the t^{th} equation in (1) can be written as

$$\begin{bmatrix} \mathtt{K}^i \; [\; \mathtt{I} \mid \mathtt{0} \;] \\ \mathtt{K}^j \; [\; \mathtt{R}_t \mid \mathbf{t}_t \;] \end{bmatrix} \begin{bmatrix} \mathtt{h}_t & \mathbf{u}_t \\ \mathtt{0}_{1\times 3} & s_t \end{bmatrix} = \begin{bmatrix} \mathtt{K}^i \; [\; \mathtt{R}^i \mid \mathbf{t}^i \;] \\ \mathtt{K}^j \; [\; \mathtt{R}^j \mid \mathbf{t}^j \;] \end{bmatrix}. \tag{2}$$

Here, the matrix of internal parameters, $\mathtt{K}^i \in \mathbb{R}^{3\times 3}$, is known with focal length estimated as described in section 3.1, $\mathtt{R}$ is 3×3 rotation, $\mathbf{t} \in \mathbb{R}^3$ is translation, and $\mathtt{I} \in \mathbb{R}^{3\times 3}$ is an identity matrix. Homographies $\mathtt{H}_t$ simplified to rotations $\mathtt{h}_t$, translations $\mathbf{u}_t$ and scales s_t. To simplify notation, rotation and translation of the first camera in each partial reconstruction has been transformed to identity and zeros, respectively, by applying appropriate rotation and translation beforehand.

After multiplying each triple of rows by the corresponding ${\mathtt{K}^i}^{-1}$ from the left, system (2) becomes

$$\left[\begin{array}{c|c} \mathtt{I} & \mathtt{0} \\ \mathtt{R}_t & \mathbf{t}_t \end{array}\right] \begin{bmatrix} \mathtt{h}_t & \mathbf{u}_t \\ \mathtt{0}_{1\times 3} & s_t \end{bmatrix} = \left[\begin{array}{c|c} \mathtt{R}^i & \mathbf{t}^i \\ \mathtt{R}^j & \mathbf{t}^j \end{array}\right]. \tag{3}$$

By writing only the first three columns of eq. (3), one obtains:

$$\begin{aligned} \begin{bmatrix} \mathtt{I} \\ \mathtt{R}_1 \end{bmatrix} \mathtt{h}_1 &= \begin{bmatrix} \mathtt{R}^i \\ \mathtt{R}^j \end{bmatrix} \\ &\vdots \\ \begin{bmatrix} \mathtt{I} \\ \mathtt{R}_T \end{bmatrix} \mathtt{h}_T &= \begin{bmatrix} \mathtt{R}^i \\ \mathtt{R}^j \end{bmatrix} \end{aligned} \tag{4}$$

as translation $\mathbf{t}_t$ is multiplied with zeros in the middle matrix in (3).

A solution to system (4) in the least squares was published in [15, section 2.1]. See it for details on the solution using Matlab's EIGS and its numerical behaviour. It provides the closest solution to all partial rotations in the least squares. The found solution does not represent rotations (the 3x3 matrices are not orthonormal), but as it is close to the true rotations of partial reconstructions, it is close to some true rotations as well. The true rotations, $\bar{\mathtt{R}}^i$, can be found as the closest rotation in the least squares, e.g., as $\bar{\mathtt{R}}^i = \mathtt{UV}^\top$ where $\mathtt{R}^i = \mathtt{U}\,\mathrm{diag}(\sigma_1, \sigma_2, \sigma_3)\mathtt{V}^\top$ is the SVD factorization.

The t^{th} equation in (4) is weighted by the root of the EG support, as it was done in [15, Eq. (10)].

3.3. Refining Rotations

Rotations in the t^{th} partial reconstruction are replaced by the consistent rotations, $\bar{\mathtt{R}}^{\mathbf{i}_t}$, and translations are modified accordingly using $\mathtt{h}_t$[4]. Due to errors in image measurements, the rotation between consistent rotations $\bar{\mathtt{R}}^i$ and $\bar{\mathtt{R}}^j$, $\{i, j\} = \mathbf{i}_t$ for some t, does not equal the rotation between the two cameras in the t^{th} original partial reconstruction, $\bar{\mathtt{R}}^{i\top}\bar{\mathtt{R}}^j \neq \mathtt{R}_t$. As a consequence, reprojection errors grow after making rotations consistent. Hence, refinement using bundle adjustment is desired. Before that, it is necessary to either rotate points in the t^{th} partial reconstruction by $\mathtt{h}_t^{-t}$ or triangulate them using new rotations. The latter is preferred, as lower reprojection errors are achieved.

Triangulation using the Sampson's approximation [8] is computed. For points reconstructed behind at least one camera [27], the triangulation is computed again using SOCP in the L_∞-norm [11]. The reason for doing the Sampson's approximation first is that it is much faster. Then, BA was applied on the t^{th} partial reconstruction while keeping rotations constant. When some of the points got behind a camera during the BA, both camera translations and points were estimated using SOCP [11]. SOCP was used here only as an emergency solution as it is time consuming compared to BA.

Finally, full bundle adjustment on all partial reconstructions at the same time with corresponding rotation parameters kept the same is run. The overall scale of focal lengths is varied. Such a bundle adjustment (BA) is something in between T independent BAs of T partial reconstructions, each with two cameras, on one hand and standard BA with all m cameras in the same coordinate system on the other hand. Here, $2T$ cameras share only rotations while translations are still inconsistent. This way provides a higher flexibility allowing to change translation in the t^{th} partial reconstruction almost independently of translations in the remaining reconstructions. More precisely, translations influence themselves via shared rotations, which is however a more free connection than demanding consistency between translations. This approach should be thus less prone to stucking in a local minimum than the standard BA.

Some EGs with small support may be found even on image pairs with no overlap. These EGs are easily detected after several (20) steps of BA as those with no inliers with respect to the desired accuracy (1 pixel).

[4] Better results were achieved without projecting $\mathtt{h}_t$ onto space of rotations.

3.4. Consistent Translations and Scale

A straightforward way for obtaining global scales and translations is to estimate them together with rescaling and translation of each partial reconstruction in a similar way as in section 3.2. However, our results when using this approach were not satisfactory. The reason why this approach worked well for rotations is perhaps that there are no significant differences in magnitudes of the variables (there are just orthonormal 3×3 matrices in eq. (4)). On the other hand, translations can have large differences in magnitudes across partial reconstructions.

Therefore, the state-of-the-art SOCP method [11] was used to estimate both camera translations and points. It gives good results as the reprojection error is minimized while keeping all points in front of cameras. The fact that the L_∞-norm is minimized is not a problem as final bundle adjustment minimizing the L_2-norm is run anyway.

Handling Mismatches

To add robustness to mismatches, only some points (from more than 10^5 in our experiments) are sampled for bundle adjustment and SOCP in section 3.4. To capture the overall geometry, points are sampled so that from each image pair having nonzero points in common, 90% of points with lowest reprojection errors are chosen. By this simple way most mismatches are removed while removing only a reasonable amount of inliers and capturing the overall geometry without a complicated threshold estimation.

4. Handling Unequiponderant Data

In our setup, the focal length for camera calibration was estimated from the data not accurately (there was about 5% error on the Head scene). If images of the scene were captured equiponderantly so that all sides of the object occupied roughly the same amount of the image data, the standard bundle adjustment converged to a satisfactory minimum. However, this is not the case of the data used here. All parts of the head sculpture were captured on quite many images except the back of the hand where only six images (11-13, 20-22, see fig. 4) were taken with large changes in base-line causing fewer matches. The error in the estimate of the focal length (and perhaps also the radial distortion) caused in the standard bundle adjustment that the well covered data overweighted the small contribution from the back of the hand. As a result, the hand and the head in place above it are split into two discontinuous surfaces, see fig. 2.

Our approach to avoid such failures is to find which EGs are more important for the geometry of the overall 3D model and to support such EGs more. Distinguishing which EGs are more important for 3D is hard even if some rough 3D model is given because there may be various degeneracies like dominant planes and camera rotations. To do it thoroughly, one should consider that on one hand wide base-line pairs provide better conditioned 3D estimates but on the other hand have smaller support due to large camera movement.

Two terms should be distinguished: *importance* of an EG and *quality / reliability* of an EG. An EG will be called *weak* if it is not reliable. Even a weak EG can be important, thus if it was removed, the overall geometry would change much. On the other hand, a strong EG does not have to be necessarily important, consider for instance two same images. Here by an EG it is meant relative position and orientation of the camera pair, which is equivalent to the epipolar constraint plus camera calibration.

The larger the support is, the more reliable the EG could seem to be. However, if many correspondences lie on a dominant plane, the constraint on the overall geometry provided by the EG is rather weak. Quality / reliability of an EG can be evaluated by perturbation of EG parameters [7, 13].

One way for determining importance of an EG could be to reconstruct / refine the 3D reconstruction without the EG and observe how much the reconstruction bends. If it bends much, the EG is important for the overall geometry. This would be repeated for all EGs. Note that this approach is very computationally demanding.

Our method is rather simple but worked well on our data. It is based on finding shortest (and slightly longer) paths in a graph induced by known EGs. Each vertex of the graph stands for a camera. Two vertices are connected by an edge iff there is a known EG between the corresponding cameras. We have observed that

Principle 1 *For estimating relative positions of any two cameras, the most important EGs are those which lie on the shortest paths between the two cameras.*

Shorter paths seem to be more important than longer ones because noise in each additional camera along the path increases uncertainty in the 3D geometry between the two cameras.

Let the *graph of known EGs*, $G = (V, \mathtt{E})$, be defined as a set of vertices, V, $V = \{1 \dots m\}$, and adjacency matrix, $\mathtt{E} \in \mathbb{R}^{m \times m}$, where V corresponds to cameras and $\mathtt{E}(i, j) = 1$ when an EG is defined between cameras i and j, otherwise $\mathtt{E}(i, j) = 0$. In our current implementation, an EG is defined if it has at least some minimal support (30 inliers).

The task is to estimate importance of all EGs. It will be stored in *EG importance* matrix $\mathtt{S} \in \mathbb{R}^{m \times m}$. Between each pair of vertices, all shortest (and slightly longer) paths will be found in a breadth-first-search manner as will be explained below. All such paths contribute to the importance of EGs (associated with the edges) through which they pass. All contributions are summed up in the EG importance $\mathtt{S}$.

It is not sufficient to find just one shortest path between two vertices in the graph. The reason is that if more shortest paths exist, all participate in constraining the 3D geometry between the two cameras. Thus, Floyd-Warshall's algorithm [21] is not usable as it finds just one shortest path between two vertices, although it has low complexity $O(m^3)$.

Finding All Shortest Paths

A *path* means here a sequence of adjacent vertices and edges where both can appear multiple times. In a *simple path*, all vertices and edges are distinct.

It is well known in graph theory that $\mathtt{E}^k(i, j)$ equals the number of all paths of length k between i and j where $\mathtt{E}^k$ is the k^{th} power of $\mathtt{E}$. On a complete graph, i.e. $\mathtt{E}(i, j) = 1$ iff $i \neq j$, it can be easily shown that $\mathtt{E}^k(i, j) \geq (m - 2)^{k-1}$ for $i \neq j$. Due to the exponential growth of the number of paths with their length, finding all paths between two vertices followed by adding some weight to the $\mathtt{S}$ matrix on edges along the shortest paths is infeasible.

Our strategy is not to track all paths one by one (they are too many) but to track all paths of length k from vertex f to the remaining vertices at the same time. For each vertex, t, all paths of length k from f to t are registered. As the only desired output is the $\mathtt{S}$ matrix, i.e. some weights on graph edges, it is sufficient to register not all particular paths but only the number of paths leading via each edge. At each vertex, $t \in \{1 \dots m\}$, matrix $\mathtt{A}_t^k \in \mathbb{R}^{m \times m}$ is stored. Entry $\mathtt{A}_t^k(i, j)$ equals the number of all paths of length k between vertices f and t leading via edge (i, j).

Our algorithm for finding all paths from a given vertex, f, to the remaining vertices works as follows. Paths of length one are registered, i.e. $\mathtt{A}_i^1(f, i) = \mathtt{A}_i^1(i, f) = 1$ for $i \in$ neighbors(f). At step k, paths of length k are prolonged and stored in matrices $\mathtt{A}_i^{k+1}$. The shortest paths between f and i are in $\mathtt{A}_i^k$ where

$$k = \min\{k \mid \mathtt{A}_i^k \text{ is not all zeros}\}. \tag{5}$$

Proposition 1 *Matrix $\mathtt{A}_i^k$ corresponds to shortest paths from f to i for k defined in eq. (5) (which also means they are simple paths).*

Proof. If there was any shorter path of length l, $l < k$, matrix $\mathtt{A}_i^l$ would have some non-zero element. Contradiction with definition of k. □

The algorithm is summarized in algorithm 1. Here, norm $|\cdot|$ of a matrix denotes the sum of its elements, $|\mathtt{A}| = \sum_{i,j} \mathtt{A}(i,j)$. The upper bound on complexity of algorithm 1 is $O(m^3 E)$ where E is the number of graph edges when using sparse matrix representation. It is run for all vertices:

initiate matrix $\mathtt{S} \in \mathbb{R}^{m \times m}$ to zeros
for $f \in \{1 \dots m\}$
  $\mathtt{S} = \mathtt{S} + \mathtt{S}_f$ //contribution by paths from f

Thus, the overall complexity is at most $O(m^4 E)$. It results in a fraction of time spent in the reconstruction pipeline.

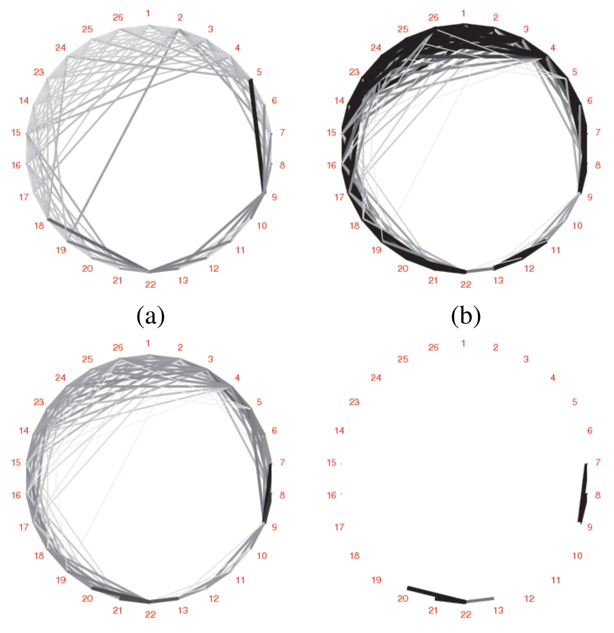
Input: A graph and a vertex, f. EG reliability matrix, $\mathtt{w}$.
Output: Contribution, $\mathtt{S}_f$, to the EG importance matrix by all shortest and slightly longer paths from f to the remaining vertices. Similarly for contribution, $\mathtt{T}_f$, to the EG reliability-importance matrix.

initiate $\mathtt{A}_i^k, \mathtt{W}_i^k \in \mathbb{R}^{m \times m}$ to zeros for $i, k \in \{1 \dots m\}$
for $i \in$ neighbours(f)
  $\mathtt{A}_i^1(f,i) = \mathtt{A}_i^1(i,f) = 1$
  $\mathtt{W}_i^1(f,i) = \mathtt{W}_i^1(i,f) = \mathtt{w}(i,f)$
for $k \in \{1 \dots m-2\}$ //prolong paths of length k
  for $p \in \{i \mid \mathtt{A}_i^k$ is not all zeros$\}$
    for $t \in$ neighbours(p)
      $\mathtt{B} = \mathtt{A}_p^k$ //take all paths from f to p
      $\mathtt{V} = \mathtt{W}_p^k$
      $\mathtt{B}(t,p) = \mathtt{B}(p,t) = \mathtt{B}(p,t) + \frac{|\mathtt{A}_p^k|}{2k}$ //prolong to t
      $\mathtt{V}(t,p) = \mathtt{V}(p,t) = \mathtt{V}(p,t) + \frac{|\mathtt{A}_p^k|}{2k}$
      $\mathtt{A}_t^{k+1} = \mathtt{A}_t^{k+1} + \mathtt{B}$
      $\mathtt{W}_t^{k+1} = \mathtt{W}_t^{k+1} + \mathtt{V} \cdot \mathtt{w}(p,t)$
initiate $\mathtt{S}_f, \mathtt{T}_f \in \mathbb{R}^{m \times m}$ to zeros
for $t \in \{1 \dots m\} \backslash f$
  $l = \min\{l \mid \mathtt{A}_t^l$ is not all zeros$\}$ //shortest path to t
  initiate $\mathtt{B}, \mathtt{V} \in \mathbb{R}^{m \times m}$ to zeros
  for $k \in \{l \dots \lceil \frac{3}{2} l \rceil\}$ //+ slightly longer
    $\mathtt{B} = \mathtt{B} + \mathtt{A}_t^k \frac{2}{|\mathtt{A}_t^k|}$
    $\mathtt{V} = \mathtt{V} + \mathtt{W}_t^k \frac{2}{|\mathtt{A}_t^k|}$
  $\mathtt{S}_f = \mathtt{S}_f + \mathtt{B} \frac{2}{|\mathtt{B}|}$
  $\mathtt{T}_f = \mathtt{T}_f + \mathtt{V} \frac{2}{|\mathtt{V}|}$

Algorithm 1: Algorithm for finding all shortest and slightly longer paths from a given vertex to the remaining ones

Note on algorithm 1. The formula for the number, N, of paths of length k leading from f to p, $N = \frac{|\mathtt{A}_p^k|}{2k}$, can be easily found by induction. It also holds $N = \mathtt{E}^k(f,p)$. □

The EG importance, $\mathtt{S}$, found using algorithm 1 on the Head set is shown in figure 4a. It turns out that edges close to articulations (here vertices 9 and 22) in the graph gather up most importance, which is what is desired. However, it turns also out that shortest paths tend to include weak EGs, which are prone to be completely wrong, like EG 2-19 between images with no overlap, see fig. 4a and 4b. Therefore, two extensions are made:



(a) (b) (c) (d)

Figure 4: Scoring of EGs using all shortest and slightly longer paths on the Head set. (a) EG importance (b) EG reliability (c) EG reliability-importance (d) 4% of the most important EGs. More important/reliable EGs are drawn darker and thicker. The most shortest paths lead via articulations (images 9 and 22). Images are reorded due to visualization

1. So called *EG reliability-importance matrix*, $\mathtt{T}$, is estimated similarly to the EG importance matrix $\mathtt{S}$ by reweighting edges along each path by the corresponding EG reliability. The *EG reliability matrix*, $\mathtt{w} \in \mathbb{R}^{m \times m}$, holds the ij-EG support in $\mathtt{w}(i,j)$. See alg. 1.

2. Slightly longer (by factor of $\frac{3}{2}$ in our implementation) than the shortest paths are used also.

Both these extensions lead to suppressing of weak EGs which lie along short paths (see fig. 4c), thus providing higher robustness to mismatches and less sensitivity to the threshold on an acceptable EG.

4.1. Using the EG importance

The estimated EG reliability-importance can be used in a two ways: (i) First, it can be used to weight the corresponding equation in (4) instead of EG support (see end of sec. 3.2) and to weight the data in the BA in sec. 3.3 as well as in the final BA. (ii) Second, the most important EGs can be strengthened by adding appropriate image triples.

Unfortunately, important EGs have often small support, thus triples containing them have even smaller (a point visible in three images must be visible in each image pair). Our solution is to add triples containing only image i and triples containing image j. In experiments shown in this paper, 4% of the EGs with the highest reliability-importance were chosen (see fig. 4d). All triples containing at least one of the images associated with these EGs were taken if all the three EGs were defined. In the Head set, 69 triples were chosen.

One might use the three-view matches in BA to refine the initial estimate obtained using pair-wise matches as described in section 3. A better way is to exploit the data at the very early stage for obtaining the consistent rotations in section 3.2. This is very usefull as image triples provide stronger constraints on 3D geometry. Thus, a better initial estimate of the reconstruction may lead to avoiding some local minima in BA.

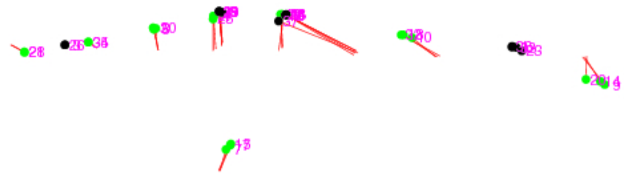
Importance of a triple was estimated as the mean of the reliability-importance of the three associated EGs weighted by the number of the three-view inliers. Partial reconstructions of the chosen image triplets were obtained in the following way. (i) An initial estimate of the camera triple was estimated from pair-wise reconstructions using algorithm described in section 3. (ii) A four-point RANSAC was run on three-view matches. For each sample, BA was run on the four points to get the model (three camera matrices). Support was obtained using triangulation. (iii) The reconstuction was refined by BA on two- and three-view inliers.

5. 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.

The Head sculpture was captured on 26 images. A similar image set of 10 images was used in [5] but covering only cca 120 degrees of the circular path around the statue. 91% data is missing. Figure 6 captures the all-around reconstruction with correct surface and surrounding buildings obtained using the EG reliability-importance from sec. 4.1.

The proposed algorithm has been tested on the image set from the final round of the ICCV Computer Vision Contest [1]. This difficult data set contains several panoramas with many camera rotations and dominant planes. Our method achieved mean / maximum reprojection error of 3.01 / 4.87 meters evaluated on the GPS ground truth available at the contest page. Our result with average score 4.09 outperformed the best team in the contest. Cameras' focal length has not been calibrated using calibration grids available at the contest page. No radial distortion removal has been applied. The results can be seen in figure 5. The bending of the reconstruction is caused by imprecise focal length estimation and perhaps also radial distortion. This scene has a linear structure without any cycle around an object like in the Head set which could enforce strong constraints on focal length. Two most distant cameras with known GPS positions were aligned by a similarity to the ground truth.



Avg. Score	< 2 m (5 pts)	< 4 m (4 pts)	< 8 m (3 pts)
4.09	4	16	2

Figure 5: Reconstruction of the final round image set from the ICCV05 Vision Contest: (top) top view. The black points correspond to cameras with known GPS positions used to transform the reconstruction into the world coordinate system. Lines join the estimated cameras with the ground truth. (bottom) Score is counted on cameras with unknown GPS positions, see [1]

To demonstrate quality of the reconstructions, the estimated cameras were used by method [5] to produce dense reconstructions. The results can be seen in figures 6 and 2.

Besides scenes shown here, our method was tested on other scenes including the Dinosaur sequence with similar results as in [15]. See more reconstructed scenes at [2].

6. Discussion

To handle degenerate situations, one might detect panoramas. However, decision if two images are related by a camera rotation is difficult especially with an unreliable estimate of focal length. All steps of our algorithm are suited for both degenerated pairs and pairs describing full 3D geometry, which was demonstrated on the ICCVC05 data.

If all camera centers are collinear, 3D reconstruction obtained using points visible in two images only will not be unique. Then triples are needed as well.

Another possible application of the EG importance is detection of most important image pairs for guided matching.

7. Summary

A new method for muttiple-view reconstruction based on making rotations consistent using a linear formulation was proposed. It can be used for an extreme case of the missing data, i.e. when each point is visible in two images only. The method is capable of dealing with degenerate situations like dominant planes and camera rotation and zooming.

Figure 6: All-around reconstruction of the Head statue (from left): front view using fish-scales, side and overall top views with reconstructed buildings around using point clouds (10% points shown)

It has been shown that standard bundle adjustment fails on unequiponderantly obtained data with an imprecise estimate of the focal length, but when importance of the data is examined from a global view, correct reconstruction can be obtained. For this purpose, a new criterion of importance of an EG on the overall 3D geometry has been formulated using shortest paths in a graph.

References

[1] http://research.microsoft.com/iccv2005/Contest.

[2] http://cmp.felk.cvut.cz/˜martid1/demo3DPVT06.

[3] S. Avidan and A. Shashua. Threading fundamental matrices. In *IEEE Trans. on PAMI*, vol. 23(1), pp. 73–77, 2001.

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

[5] H. Cornelius et al. Towards complete free-form reconstruction of complex 3D scenes from an unordered set of uncalibrated images. In *SMVP/ECCV*, pp. 1–12, Prague, Czech Republic, May 2004.

[6] A. W. Fitzgibbon and A. Zisserman. Automatic camera recovery for closed or open image sequences. In *Proc. ECCV*, vol. I, pp. 311–326, June 1998.

[7] D. Forsyth, S. Ioffe, and J. Haddon. Bayesian structure from motion. In *ICCV*, pp. 660–665, 1999.

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

[9] R. I. Hartley, E. Hayman, L. d. Agapito, and I. D. Reid. Camera calibration and the search for infinity. In *Proc. ICCV*, vol. 1, pp. 510–517, 1999.

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

[11] F. Kahl. Multiple view geometry and the l_∞-norm. In *ICCV05*, pp. II: 1002–1009, 2005.

[12] 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. Available from `http://www.ics.forth.gr/˜lourakis/sba`.

[13] Q.-T. Luong and O. D. Faugeras. A stability analysis of the fundamental matrix. In *ECCV (1)*, pp. 577–588, 1994.

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

[15] 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.

[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] D. Nistér. Untwisting a projective reconstruction. *IJCV*, 60(2):165–183, November 2004.

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

[21] R. Sedgewick. *Algorithms in C : Part 5 : Graph Algorithms*. Addison-Wesley, Reading, Massachusetts, 3rd edition, 2002.

[22] 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.

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

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

[25] P. H. S. Torr. Model selection for two view geometry: A review. In *Shape, Contour and Grouping in Computer Vision*, pp. 277–301, 1999.

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

[27] T. Werner and T. Pajdla. Oriented matching constraints. In *BMVC*, pp. 441–450, Manchester, UK, September 2001.