CENTER FOR
MACHINE PERCEPTION

CZECH TECHNICAL
UNIVERSITY

REPRINT

3D Reconstruction by Fitting Low-rank Matrices with Missing Data

Daniel Martinec and Tomáš Pajdla

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

D. Martinec and T. Pajdla. 3D reconstruction by fitting low-rank matrices with missing data. In *Proceedings of the Computer Vision and Pattern Recognition conference 2005*, San Diego, CA, USA, June 2005.

Available at
ftp://cmp.felk.cvut.cz/pub/cmp/articles/martinec/Martinec-CVPR2005.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 Fitting Low-rank Matrices with Missing Data

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

A technique for building consistent 3D reconstructions from many views based on fitting a low rank matrix to a matrix with missing data is presented. Rank-four submatrices of minimal, or slightly larger, size are sampled and spans of their columns are combined to constrain a basis of the fitted matrix. The error minimized is expressed in terms of the original subspaces which leads to a better resistance to noise compared to previous methods. More than 90% of the missing data can be handled while finding an acceptable solution efficiently. Applications to 3D reconstruction using both affine and perspective camera models are shown. For the perspective model, a new linear method based on logarithms of positive depths from cheirality is introduced to make the depths consistent with an overdetermined set of epipolar geometries. Results are shown for scenes and sequences of various types. Many images in open and closed sequences in narrow and wide base-line setups are reconstructed with reprojection errors around one pixel. It is shown that reconstructed cameras can be used to obtain dense reconstructions from epipolarly aligned images.

1. Introduction

Problem of fitting a low-rank matrix to a matrix with missing data appears in 3D reconstruction of rigid [14, 7, 9, 4] and non-rigid scenes [2]. Several attempts to provide reconstruction from many images in a one-step algorithm have been made. However, none of these methods succeeded to process more than a few tens of images when the amount of missing elements reaches 90% of the measurement matrix and cameras have large field of view or are in a wide base-line setup. In this paper we present a technique that builds a consistent reconstruction (1) for scenes of various types: open and closed sequences in both narrow and wide base-line setups, and (2) for various camera models: affine and perspective, which can model also omnidirectional cameras.

Figure 1. Reconstruction from a wide base-line scenario after Euclidean BA: the St. Martin rotunda on 24 images, 89% data missing, top view. Some cameras are positioned very close to each other while some are distant making the SFM problem difficult

Our algorithm has the following advantages: (i) it provides an overall scene structure and motion in a single step without requirements such as linear ordering of images in a sequence (ii) the solution is obtained as a global optimum of a reasonable cost function defined on an approximation to the original SFM (structure-from-motion) problem. The obtained projective reconstruction can be easily upgraded to the metric one, see figure 1. The result can be used for dense reconstruction, see figure 6.

Relevant methods can be divided into two groups. (i) So called factorization methods, e.g. [14, 13, 7, 9], try to fill the unknown elements of the matrix of all measurements. (ii) In

*This research was supported by the The Czech Academy of Sciences under project 1ET101210406 and by the EU project IST-2001-39184. Andrew Zisserman from the University of Oxford kindly provided the Dinosaur data, Tomáš Werner from the Czech Technical University (CTU) provided the routine for the bundle adjustment and Jana Kostková from CTU kindly made the dense reconstructions.

Guilbert's method [4], affine fundamental matrices are estimated from the image measurements, and affine camera matrices are estimated from the fundamental matrices. Our new method cannot be classified as a factorization method, although factorization on small complete matrices appears inside. Similarly to [4], it produces a direct solution on camera matrices. In contrast to [4], the minimized error is expressed in terms of image data and not elements of fundamental matrices as in [4].

Our method does not try to fill any elements and even does not hallucinate them as, e.g., in [2][1]. The missing data are not modelled in the algorithm at all. Only known data are exploited while minimizing their distance to the fitted matrix. The method bootstraps from rank-four submatrices of minimal or slightly larger size. Linear spaces generated by their rows or columns are combined to constrain basis of the whole measurement matrix. This idea has already appeared in [7]. However, the way it is realized in this paper is novel.

The most crucial difference from [7] is that the solved problem is formulated in terms of the original subspaces, and not the complementary ones as in [7]. Therefore, error due to noise is corrected where it was physically caused, i.e. in the spaces generated by image measurements and not in their complements. Our formulation is equivalent to [7] in case there is no noise in the data. However, as a reasonable error is minimized, it has much better behaviour when noise is present which enables handling lots of missing data. Moreover, it leads to precise and fast algorithms as only a small system of equations with a sparse design matrix has to be solved. In application to 3D reconstruction, large data compression is reached by taking only the four singular vectors best explaining the submatrix of (possibly) large amount of points seen in an image pair or triple.

This paper brings the following contributions: (i) a new technique for fitting a low-rank matrix to a matrix with missing data is introduced (section 2). (ii) Two ways of its application to the structure-from-motion problem are given for both affine and perspective camera models (sections 3 and 4), (iii) a new method for estimating projective depths consistent with an overdetermined system of epipolar geometries is introduced (section 4.1).

Basic Concepts

Consider a set of n 3D points, some of which are visible in some of m images. There may be mismatches in image measurements. The goal is to reject mismatches and to recover the 3D structure (point locations) and motion (camera locations) from the remaining image measurements.

Let $\mathbf{X}_p$ be the unknown homogeneous coordinate vectors of the 3D points, $\mathtt{P}^i$ the unknown 3×4 projection matrices, and $\mathbf{x}_p^i$ the measured homogeneous coordinate vectors of the image points, where $i = 1, \ldots, m$ labels images and $p = 1, \ldots, n$ labels points. Due to occlusions and misdetections, $\mathbf{x}_p^i$ are unknown for some i and p.

The basic image projection equation says that $\mathbf{x}_p^i$ are the projections of $\mathbf{X}_p$ up to unknown scale factors λ_p^i called (*projective*) *depths*:

$$\lambda_p^i \mathbf{x}_p^i = \mathtt{P}^i \mathbf{X}_p \tag{1}$$

The complete set of image projections can be gathered into a matrix equation. The $3m \times n$ matrix $[\mathbf{x}_p^i]_{p=1\ldots n}^{i=1\ldots m}$ will be called the *measurement matrix* (MM).

2. Fitting Matrices with Missing Data

Let $\mathbf{y} \in \mathbb{R}^{3m \times n}$ be some matrix with missing elements. The method will be explained on rank-four matrices, but it can be used for any rank. Matrix $\mathbf{y}$ can represent, e.g., the MM rescaled by the projective depths, $[\lambda_p^i \mathbf{x}_p^i]_{p=1\ldots n}^{i=1\ldots m}$. The task is to find the best rank-four fit $\mathtt{PX}$ to known elements of $\mathbf{y}$ in the least squares where $\mathtt{P} \in \mathbb{R}^{3m \times 4}$, $\mathtt{X} \in \mathbb{R}^{4 \times n}$. A description of a good suboptimal solution follows.

Rank-four submatrices of $\mathbf{y}$ will be used to constrain the column basis $\mathtt{P}$ of $\mathbf{y}$. Let their number be denoted by T. Let $\mathbf{i}_t$ and $\mathbf{p}_t$ denote sets of row and column indices of the t^{th} submatrix within $\mathbf{y}$, respectively, $3|\mathbf{i}_t| \geq 4$, $|\mathbf{p}_t| = 4$. Notation $\mathtt{A}_{\mathbf{p}}^{\mathbf{i}}$ will denote the submatrix of $\mathtt{A}$ composed of elements in rows $\mathbf{i}$ and columns $\mathbf{p}$. Omitting superscript or subscript means taking all rows or columns, respectively. Let only submatrices with (i) all its elements known and (ii) linearly independent columns be chosen. Let the t^{th} submatrix be denoted by $\tilde{\mathtt{P}}_t$, $\tilde{\mathtt{P}}_t = \mathbf{y}_{\mathbf{p}_t}^{\mathbf{i}_t}$.[2] Then,

$$\begin{aligned} \tilde{\mathtt{P}}_1 &= \mathtt{P}^{\mathbf{i}_1} \mathtt{X}_{\mathbf{p}_1} \\ &\vdots \\ \tilde{\mathtt{P}}_T &= \mathtt{P}^{\mathbf{i}_T} \mathtt{X}_{\mathbf{p}_T}. \end{aligned} \tag{2}$$

From linear independency of columns of $\tilde{\mathtt{P}}_t$ and (2),

$$\operatorname{rank} \tilde{\mathtt{P}}_t = \operatorname{rank} \mathtt{P}^{\mathbf{i}_t} = \operatorname{rank} \mathtt{X}_{\mathbf{p}_t} = 4. \tag{3}$$

Jacobs [7] used the fact that $\operatorname{span} \tilde{\mathtt{P}}_t = \operatorname{span} \mathtt{P}^{\mathbf{i}_t}$ to constrain $\mathtt{P}$ by $\operatorname{span} \mathtt{P} \subseteq \operatorname{span} \mathbf{y}_{\mathbf{p}_t}$ where $\operatorname{span} \mathbf{y}_{\mathbf{p}_t}$ can be interpreted as the linear hull of all possible fillings of $\mathbf{y}_{\mathbf{p}_t}$ (see the interpretation in [9, fig. 1]). He formulated the constraint using complementary subspaces $\mathtt{N}_t = (\operatorname{span} \mathbf{y}_{\mathbf{p}_t})^{\perp}$ as $\mathtt{P} \subseteq \mathtt{N}^{\perp}$ where $\mathtt{N}$ is the union of the complementary subspaces, $\mathtt{N} = \bigcup_{t=1,\ldots,T} \mathtt{N}_t$.

However, this formulation does not treat noise well. Small changes in $\mathtt{N}_t$ (caused by noise in $\mathbf{y}_{\mathbf{p}_t}$) are accumulated in their union $\mathtt{N}$ and may result into a large change

[1] Method [2] was not usable on the data presented in this paper at all

[2] $\tilde{\mathtt{P}}_t$ can be viewed as cameras in a projective reconstruction of $\mathbf{y}_{\mathbf{p}_t}^{\mathbf{i}_t}$ with points $\mathtt{I}_{4\times4}$, $\mathbf{y}_{\mathbf{p}_t}^{\mathbf{i}_t} = \tilde{\mathtt{P}}_t \mathtt{I}_{4\times4}$, where $\mathtt{I}$ denotes the identity matrix.

in $\mathtt{N}^\perp$. The reason is that the noise is physically caused in the original subspaces (on image data) where it should also be corrected, as our method does, which will be shown below. In fact, [7] corrects the error in the complementary subspaces ($\mathtt{N}_t$, $\mathtt{N}$) which have an unclear connection to the original noise.[3] We observed that [7] breaks down when noise is present and the number of images, over which the partial reconstructions are glued, reaches some limit. E.g., [7] could reconstruct only a subsequence of at most 22 images of the Dinosaur sequence shown in figure 2. We have not observed any such limit at our method even when hundreds of images were used.

Our new approach exploits the fact that $\operatorname{rank} \mathtt{X}_{\mathbf{p}_t} = 4$ thanks to which the inverse to $\mathtt{X}_{\mathbf{p}_t}$ exists, $\mathtt{H}_t = \mathtt{X}_{\mathbf{p}_t}^{-1}$. The t^{th} equation in (2) can be now multiplied by $\mathtt{H}_t$ from the right:

$$\begin{aligned} \tilde{\mathtt{P}}_1 \mathtt{H}_1 &= \mathtt{P}^{\mathbf{i}_1} \\ &\vdots \\ \tilde{\mathtt{P}}_T \mathtt{H}_T &= \mathtt{P}^{\mathbf{i}_T}. \end{aligned} \tag{4}$$

Although equations 2 are bilinear in unknowns $\mathtt{P}$ and $\mathtt{X}$, equations 4 are linear in all unknowns, $\mathtt{P}$ and $\mathtt{H}_t$, and thus, given sufficiently many equations, solvable uniquely up to an overall projective transformation. Due to the bilinearity of the original problem (2), only its approximation is found by solving the transformed problem (4). However, it is a good approximation, as will be demonstrated.

2.1. Solving System 4

Denoting $\mathtt{P}^{\mathbf{i}_t} = [\mathbf{q}_1^{\mathbf{i}_t} \mathbf{q}_2^{\mathbf{i}_t} \mathbf{q}_3^{\mathbf{i}_t} \mathbf{q}_4^{\mathbf{i}_t}]$ and $\mathtt{H}_t = [\mathbf{h}_{t,1} \mathbf{h}_{t,2} \mathbf{h}_{t,3} \mathbf{h}_{t,4}]$, the t^{th} equation in (4) can be rewritten as

$$\begin{aligned} \tilde{\mathtt{P}}_t \mathbf{h}_{t,1} - \mathbf{q}_1^{\mathbf{i}_t} &= 0 \\ \tilde{\mathtt{P}}_t \mathbf{h}_{t,2} - \mathbf{q}_2^{\mathbf{i}_t} &= 0 \\ \tilde{\mathtt{P}}_t \mathbf{h}_{t,3} - \mathbf{q}_3^{\mathbf{i}_t} &= 0 \\ \tilde{\mathtt{P}}_t \mathbf{h}_{t,4} - \mathbf{q}_4^{\mathbf{i}_t} &= 0. \end{aligned} \tag{5}$$

Denoting $\mathbf{z}_c = (\mathbf{h}_{1,c}, \ldots, \mathbf{h}_{T,c}, \mathbf{q}_c^{\mathbf{i}_1}, \ldots, \mathbf{q}_c^{\mathbf{i}_T})^\top$, the whole system 4 can be rewritten as

$$\underbrace{\begin{bmatrix} \mathtt{A} & \mathtt{0} & \mathtt{0} & \mathtt{0} \\ \mathtt{0} & \mathtt{A} & \mathtt{0} & \mathtt{0} \\ \mathtt{0} & \mathtt{0} & \mathtt{A} & \mathtt{0} \\ \mathtt{0} & \mathtt{0} & \mathtt{0} & \mathtt{A} \end{bmatrix}}_{\mathtt{B}_{4f \times 4g}} \begin{pmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \\ \mathbf{z}_3 \\ \mathbf{z}_4 \end{pmatrix} = \mathtt{0}_{4f \times 1} \tag{6}$$

where $\mathtt{A}$ and all $\mathtt{0}$ matrices are of size $f \times g$, $f = \sum_{t=1}^{T} 3|\mathbf{i}_t|$, $g = 4|T| + f$. Matrix $\mathtt{A}$ is composed of T sub-blocks in the form $\mathtt{A}^{\mathcal{I}(\mathbf{i}_t)}_{[4t-3:4t, 4|T|+\mathcal{I}(\mathbf{i}_t)]} = \mathtt{C}_t = [\tilde{\mathtt{P}}_t \mid -\mathtt{I}_{3|\mathbf{i}_t| \times 3|\mathbf{i}_t|}]$ corresponding to one equation in (5) where $\mathcal{I}(\mathbf{i})$ returns indices of rows in the MM corresponding to images $\mathbf{i}$. Matrix $\mathtt{C}_t$ is of size $3|\mathbf{i}_t| \times (4+3|\mathbf{i}_t|)$, $\operatorname{rank} \mathtt{C}_t = 3|\mathbf{i}_t|$, $\dim \operatorname{null} \mathtt{C}_t = 4$. If the partial reconstructions have sufficient overlaps in cameras for ensuring that all cameras $\mathtt{P}^i$ have consistent projective frames[4], $\dim \operatorname{null} \mathtt{A} = 4$ in case there is no noise in the data. Consequently, $\dim \operatorname{null} \mathtt{B} = 16$, see (6). The number sixteen corresponds to the freedom for the sixteen parameters of the overall projective transformation. However, not all solutions from this sixteen-dimensional space are acceptable. It is required that the solution satisfies $\operatorname{rank} \mathtt{H}_t = \operatorname{rank} \mathtt{P}^{\mathbf{i}_t} = 4$, see (3), which is equivalent a conjunction:

$$\left.\begin{array}{r} \mathbf{h}_{t,a} \text{ and } \mathbf{h}_{t,b} \text{ are linearly independent} \\ \mathbf{q}_a^{\mathbf{i}_t} \text{ and } \mathbf{q}_b^{\mathbf{i}_t} \text{ are linearly independent} \end{array}\right\} \text{ for } a \neq b. \tag{7}$$

It might be possible to solve the large system 6, e.g., by Matlab's EIGS on $\mathtt{B}^\top \mathtt{B}$, and to choose (we do not know how) some appropriate vector from its sixteen-dimensional solution space. Nevertheless, a more simple and efficient way is to find the best four linearly independent solutions to system

$$\mathtt{A}\mathbf{z} = \mathtt{0}_{f \times 1} \tag{8}$$

in the least squares. These solutions satisfy properties (7):

Proposition 1 *Let $\mathbf{z}_1$, $\mathbf{z}_2$, $\mathbf{z}_3$, $\mathbf{z}_4$ be four linearly independent solutions to system 8. Then, (7) holds.*

Proof. Due to the limited space, only the idea of the proof for $T = 1$ is shown. Let the assumption hold. For contradiction, let, e.g., $\mathbf{h}_{t,2} = \alpha \mathbf{h}_{t,1}$ for some $\alpha \in \mathbb{R}$. Then, using (5), columns of

$$\begin{bmatrix} \mathbf{h}_{t,1} & \mathbf{h}_{t,2} \\ \mathbf{q}_1^{\mathbf{i}_t} & \mathbf{q}_2^{\mathbf{i}_t} \end{bmatrix} = \begin{bmatrix} \mathbf{h}_{t,1} & \mathbf{h}_{t,2} \\ \tilde{\mathtt{P}}_t \mathbf{h}_{t,1} & \tilde{\mathtt{P}}_t \mathbf{h}_{t,2} \end{bmatrix} = \begin{bmatrix} \mathbf{h}_{t,1} & \alpha \mathbf{h}_{t,1} \\ \tilde{\mathtt{P}}_t \mathbf{h}_{t,1} & \alpha \tilde{\mathtt{P}}_t \mathbf{h}_{t,1} \end{bmatrix}$$

are linearly dependent. Contradiction.

In case of noisy data, $\tilde{\mathtt{P}}_t$ in the eq. above can be replaced by another rank-four matrix, $\dot{\mathtt{P}}_t$, close to $\tilde{\mathtt{P}}_t$ such that equations 5 hold for $\dot{\mathtt{P}}_t$ exactly. □

It turned out in our experiments that transforming $\tilde{\mathtt{P}}_t$ into an orthonormal basis by $\tilde{\mathtt{P}}_t \mapsto \tilde{\mathtt{P}}_t \mathtt{G}_t$ is a good choice, as in [7]. Note that $\mathtt{G}_t$ is absorbed by the estimated $\mathtt{H}_t$ matrix achieving well conditioning of the system.

2.2. What Is Being Minimized in System 4

The best approximate solution to the original problem (2) in the least square sense minimizes the error

$$e_{orig} = \min_{\operatorname{rank} \mathtt{P}^{\mathbf{i}_t} = \operatorname{rank} \mathtt{X}_{\mathbf{p}_t} = 4} \sum_t \left\| \tilde{\mathtt{P}}_t - \mathtt{P}^{\mathbf{i}_t} \mathtt{X}_{\mathbf{p}_t} \right\|^2$$

[3] Subspace $\mathtt{N}_t$ is represented in [7] using an orthonormal basis of the complementary subspace to $\operatorname{span} \mathbf{y}_{\mathbf{p}_t}$.

[4] Let the projective frame of cameras $\mathbf{i}_t$ be chosen as a reference frame for some t. Camera a has a consistent frame if there is (i) one image triple $\mathbf{i}_t = \{a, b, c\}$ or (ii) two image pairs $\mathbf{i}_t = \{a, b\}$ and $\mathbf{i}_{t'} = \{a, c\}$ such that b and c have a consistent frame.

where $\|\cdot\|$ denotes the Frobenius norm. In (4), error

$$e_{min} = \min_{\operatorname{rank} \mathtt{P}^{\mathbf{i}_t} = \operatorname{rank} \mathtt{H}_t = 4} \sum_t \left\| \tilde{\mathtt{P}}_t \mathtt{H}_t - \mathtt{P}^{\mathbf{i}_t} \right\|^2 \quad (9)$$

is minimized. Remember that t goes over all sampled rank-four submatrices thus in a typical situation the same elements of $\mathtt{P}$ appear many times in the previous formula. Therefore, in presence of noise it is impossible to reach zero value for e_{min}. Although e_{min} differs from e_{orig}, it is still reasonable to minimize such error, as will be shown in experiments.

Remind that factorization minimizes exactly the reprojection error when the affine camera model is used. Using the perspective camera model, if all the depths are close to equal, then it minimizes a good approximation to the reprojection error [5, p. 446]. Factorization searches for such a four-dimensional subspace that best approximates each data column. In our framework, an extensive sampling of matrices $\mathbf{y}^{\mathbf{i}_t}_{\mathbf{p}_t}$ would be necessary for reaching a similar effect at least for that all the data is used.

Such sampling would be computationally expensive and it is not clear to us how successful could be such an attempt in, e.g., equiponderant exploitation of all the data. Nevertheless, there is a way of simplifying the sampling. Its idea comes out from that whatever data are contained in matrices $\tilde{\mathtt{P}}_t$, their columns are always transformed (close) to $\mathtt{P}^{\mathbf{i}_t}$, see (4). So if two matrices $\tilde{\mathtt{P}}_t$ and $\tilde{\mathtt{P}}_s$, $t \neq s$, share the same row (or column) indices, $\mathbf{i}_t = \mathbf{i}_s$, it makes sense to use the least squares approximation to the two subspaces instead. The effect is not only reduction of the number of unknowns $\mathtt{H}_t$ but foremost suppression of noise. The least squares approximation can be obtained by factorization using SVD. By this, the powerful feature of factorization of the optimal propagation of error is adopted. It is good to use factorization globally. In our framework, factorization is used only locally due to the missing data but it does not matter much because a global propagation of error is done in (4). This propagation is done very finely as the solution is searched for in a high dimensional space thanks to many auxiliary variables $\mathtt{H}_t$, see (9). Thus, our model is very rich compared to [7] but it does not overfit. Consistent cameras $\mathtt{P}$ and homographies $\mathtt{H}_t$ are searched for so that projections of the four points $\mathtt{I}_{4\times4}$ best fit all partial reconstructions, $\tilde{\mathtt{P}}_t \mathtt{H}_t = \mathtt{P}^{\mathbf{i}_t} \mathtt{I}_{4\times4}$.

2.3. Aligning Partial Reconstructions

It would be best to estimate the subspaces from as large submatrices as possible. However, finding the largest complete submatrix in a matrix with missing elements is known to be NP-hard [7]. Moreover, in vision applications, image measurements often originate from image pairs or image triples. Therefore, we use all known measurements in an image pair or triple, similarly to [7]. Thus, for given image indices $\mathbf{i}$, submatrices $\mathbf{y}^{\mathbf{i}}_{\mathbf{p}}$ are as wide as possible.

Outlier rejection from such matrices can be easily done using, e.g., RANSAC or iterative factorization with rejecting points with reprojection errors above some threshold after each iteration (both ways lead to similar results in our experiments with 1 pxl threshold). The column basis of $\operatorname{span} \mathbf{y}^{\mathbf{i}_t}_{\mathbf{p}_t}$, denote it by $\hat{\mathtt{P}}_t$, is estimated as $\hat{\mathtt{P}}_t = \mathtt{U}_{1,2,3,4}$ where $\mathbf{y}^{\mathbf{i}_t}_{\mathbf{p}_t} = \mathtt{U} \operatorname{diag}(\sigma_1, \ldots, \sigma_z) \mathtt{V}^\top$ is the SVD factorization. The row basis of $\operatorname{span} \mathbf{y}^{\mathbf{i}_t}_{\mathbf{p}_t}$, $\hat{\mathtt{X}}_t$, is estimated as $\hat{\mathtt{X}}_t = {\mathtt{V}_{1,2,3,4}}^\top$.

Partial reconstructions are aligned via cameras using:

$$\begin{aligned} \omega_1 \hat{\mathtt{P}}_1 \mathtt{H}_1 &= \omega_1 \mathtt{P}^{\mathbf{i}_1} \\ &\vdots \\ \omega_T \hat{\mathtt{P}}_T \mathtt{H}_T &= \omega_T \mathtt{P}^{\mathbf{i}_T}. \end{aligned} \quad (10)$$

Here, ω_t denotes the weight of the t^{th} partial reconstruction taking into consideration belief of the estimate of the reconstruction expressed in terms of the number of correspondences consistent with it:

$$\omega_t = \sqrt{\frac{n_t}{\bar{n}}}$$

where $n_t = |\mathbf{p}_t|$ and $\bar{n}$ is the average number of correspondences. Normalization by $\bar{n}$ gets all weights close to one and is done due to conditioning. System 10 is solved in the least squares, thus the square root from ω_t disappears in the minimized error, see (9), which thanks to this well approximates the reprojection error measured on the whole MM.

Aligning reconstructions via points, the transposed problem, is solved similarly using the row bases:

$$\begin{aligned} \omega_1 \hat{\mathtt{X}}_1^\top \tilde{\mathtt{H}}_1 &= \omega_1 \mathtt{X}^\top_{\mathbf{p}_1} \\ &\vdots \\ \omega_T \hat{\mathtt{X}}_T^\top \tilde{\mathtt{H}}_T &= \omega_T \mathtt{X}^\top_{\mathbf{p}_T}. \end{aligned} \quad (11)$$

Here, n_t used to compute ω_t denotes the number of cameras, $n_t = |\mathbf{i}_t|$.

Indeed, system 10 can be interpreted as aligning or gluing partial reconstructions, each represented in a different projective coordinate frame by at least two cameras $\hat{\mathtt{P}}_t$, $|\mathbf{i}_t| \geq 2$. Homography $\mathtt{H}_t$ maps the coordinate system of the t^{th} partial reconstruction to the global coordinate system of the reconstruction of all data. Similarly, in system 11 each partial reconstruction is represented by at least four points $\hat{\mathtt{X}}_t$, $|\mathbf{p}_t| \geq 4$. Systems 10 and 11 are special cases of (4).

To demonstrate the quality of approximation to (2) by (10), comparison with factorization by SVD on a complete MM was done. Our method gave only 0.7% worse mean reprojection error than SVD of the complete MM obtained by multiplying cameras $\mathtt{P}$ and points $\mathtt{X}$ from the reconstruction of the Dinosaur sequence with added 1 pxl

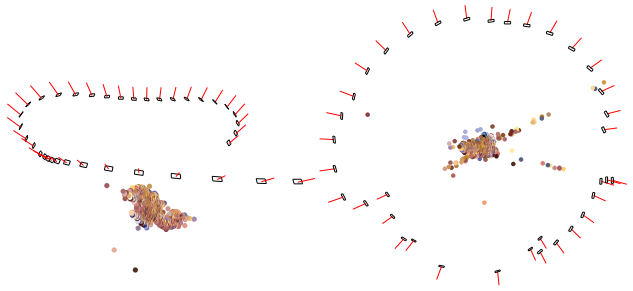

Figure 2. Initial reconstruction of the Dinosaur sequence of 35 images using affine camera model and gluing via points: (left) Open sequence with mean reprojection error 2.57 pxl. 72 image triple constraints were used. (right) Closed sequence with mean reprojection error 2.65 pxl. 74 image triple constraints were used

noise.[5] Note that in contrast to factorization our approach can handle the missing data.

It turned out in our experiments that adding also constraints from image four, five, ...-tuples did not improve results much. The reasons are the following: (i) the less images, the more matches are in them and thus the better estimation of $\hat{\mathtt{P}}_t$ and $\hat{\mathtt{X}}_t$ and (ii) the more images, the more it is likely that an outlier appears in a track (column of $\mathbf{y}^{\mathbf{i}}_{\mathbf{p}}$) regardless of the type of the used outlier rejection scheme.

3. Affine Camera Model

In the affine model, camera centers are considered to be infinitely distant from the scene structure, hence (i) all depths are equal and can be set to one and (ii) the last row of all camera matrices is $[\,0\,0\,0\,1\,]$. Therefore, image projection equation 1 can be rewritten as

$$1\begin{pmatrix}\bar{\mathbf{x}}^i_p\\1\end{pmatrix} = \left[\begin{array}{c|c}\bar{\mathtt{P}}^i & \mathbf{t}^i\\ 0\,0\,0 & 1\end{array}\right]\begin{pmatrix}\bar{\mathbf{X}}_p\\1\end{pmatrix}$$

and simplified to

$$\bar{\mathbf{x}}^i_p = \left[\,\bar{\mathtt{P}}^i\,\mathbf{t}^i\,\right]\begin{pmatrix}\bar{\mathbf{X}}_p\\1\end{pmatrix}. \tag{12}$$

Let $\bar{\mathbf{x}}^{\mathbf{i}_t}_{\mathbf{p}_t} = \mathtt{U}\,\mathrm{diag}(\sigma_1,\ldots,\sigma_z)\mathtt{V}^\top$ be the SVD factorization.

(i) *Gluing via cameras.* Equations 10 are of the following form due to the special structure of affine homographies $\mathtt{H}_t$:

$$[\hat{\bar{\mathtt{P}}}_t\hat{\mathbf{t}}_t]\begin{bmatrix}\mathtt{A}_t & \mathbf{b}_t\\ \mathtt{0}_{1\times3} & 1\end{bmatrix} = [\bar{\mathtt{P}}^{\mathbf{i}_t}\mathbf{t}^{\mathbf{i}_t}] \quad \mathtt{t} = 1,\ldots,\mathtt{T}.$$

From this, $\hat{\bar{\mathtt{P}}}_t\mathtt{A}_t = \bar{\mathtt{P}}^{\mathbf{i}_t}$, which allows to estimate $\bar{\mathtt{P}}$ up to translations as a rank-three fit similarly to (10), provided $\hat{\bar{\mathtt{P}}}_t$ have been estimated as $\hat{\bar{\mathtt{P}}}_t = \mathtt{U}_{1,2,3}$. Then, translations of all cameras $\mathbf{t}^i$ and all points $\bar{\mathbf{X}}_p$ can be estimated from the non-homogenous system 12 written for all projections.

(ii) *Gluing via points.* $\hat{\mathtt{X}}_t$ is estimated as $\hat{\mathtt{X}}_t = \mathtt{V}_{1,2,3,4}{}^\top$. The fourth coordinates of points $\mathtt{X}$ estimated using (11) are not exactly one. Therefore, the closest vector from $\mathrm{span}\,\mathtt{X}^\top$ to vector $[1\,1\ldots1]^\top$ is found in the least squares as $\mathbf{v} = \mathtt{X}^\top(\mathtt{X}^\top)^+[1\,1\ldots1]^\top$ where $+$ stands for pseudoinverse. Then, $\mathbf{v}$ is replaced in $\mathrm{span}\,\mathtt{X}^\top$ by $[1\,1\ldots1]^\top$ as $\mathtt{X} := [\mathtt{U}(:,1:3),[1\,1\ldots1]^\top]^\top$ where $\mathtt{X}^\top - \mathbf{v}\mathbf{v}^+\mathtt{X}^\top = \mathtt{U}\,\mathrm{diag}(\sigma_1,\ldots,\sigma_z)\mathtt{V}^\top$ is the SVD factorization. Cameras are estimated as $\left[\,\bar{\mathtt{P}}^i\,\mathbf{t}^i\,\right] = (\mathtt{X}^\top_{\mathbf{p}^i})^+\bar{\mathbf{x}}^i_{\mathbf{p}^i}$ where $\mathbf{p}^i$ denotes points observed by the i^{th} camera. Finally, points are estimated as $\bar{\mathbf{X}}_p = \bar{\mathtt{P}}^{i\,+}(\bar{\mathbf{x}}^{\mathbf{i}_p}_p - \mathbf{t}^i)$ where $\mathbf{i}_p$ denotes cameras observing the p^{th} point.

Results of gluing via cameras on open and closed Dinosaur sequence were 3.85 and 3.68 pxl, respectively. Results of gluing via points were better, see figure 2 and its caption. Euclidean update was done using Guilbert's method [4] and his code downloadable from his web-page (see [4]). Focal length was set to 2000 as in Guilbert's code. Reprojection errors of the initial reconstruction on both open and closed sequences are below 2.7 pxl, which is twice lower than 5.4 pxl of the state-of-the art technique [4].

Our technique using the affine model on a wider field of view resulted into mean reprojection errors of hundreds of pixels. In the St. George rotunda, a significant perspective effects are present as, for instance, cameras 22 and 65 are very close to the object, see figure 4 right. Our conclusion is that this method can be used with the affine model for a narrow field of view only. It will be shown in the next section that the perspective model can be successfully used to model a wide field of view in this method.

4. Perspective Camera Model

In factorization using perspective camera model, if all the depths are close to equal, then an approximation to the reprojection error scaled by the common value of projective depths is minimized [5, p. 446]. Depths in an image pair can be estimated using method [13] from the epipolar geometry (EG). If the image pairs form a graph without cycles (tree), depths from individual image pairs can be easily chained and the result is known to be a set of depths consistent with all used EGs [13] even in case of missing data [3]. Nevertheless, in practical situations, many more EGs are available than the $m-1$ ones exploitable in an acyclic graph. Using overdetermined constraints on depths from all (reliable) EGs would naturally (i) result in better depth estimates and (ii) allow to relate data in image pairs within cycles, which

[5] The perspective camera model was used. It was used in this paper with Hartley's normalization of the image measurements. Here, the projective depths were obtained from the reconstruction. Partial reconstructions from image triplets 1-2-3, 2-3-4, ..., $(m-2)-(m-1)-m$ were used.

concerns not only closed sequences but any wide base-line setup.

Cycles appear often in practice. For example in a closed sequence taken around an object, there is typically no point visible in all the images, as can be seen in figure 4. Although all subsequent cameras are close to each other in the graph of EGs, whatever tree is chosen, some cameras get always located at large distance in the tree graph. Particularly, the larger is the amount of images of a scene available, the more cycles are likely to appear in the data.

Let $[\lambda_p^i \mathbf{x}_p^i]_{p\in\mathbf{p}_t}^{i\in\mathbf{i}_t} = \mathtt{U}\,\mathrm{diag}(\sigma_1,\ldots,\sigma_z)\mathtt{V}^\top$ be the SVD factorization. $\hat{\mathtt{P}}_t$ from (10) are estimated as $\hat{\mathtt{P}}_t = \mathtt{U}(:,1:4)$. $\hat{\mathtt{X}}_t$ from (11) are estimated as $\hat{\mathtt{X}}_t = \mathtt{V}(:,1:4)^\top$.

4.1. Overdetermined Depths

Consider EG between images i and j. Then, the corresponding image points can be scaled by $\gamma_p^{ij,k}$ as

$$\begin{bmatrix}\gamma_1^{ij,1}\mathbf{x}_{p_1}^i \ldots \gamma_z^{ij,1}\mathbf{x}_{p_z}^i \\ \gamma_1^{ij,2}\mathbf{x}_{p_1}^j \ldots \gamma_z^{ij,2}\mathbf{x}_{p_z}^j\end{bmatrix} = \begin{bmatrix}\hat{\mathtt{P}}^{ij,1} \\ \hat{\mathtt{P}}^{ij,2}\end{bmatrix}\left[\,\hat{\mathbf{X}}_{p_1}\ \ldots\ \hat{\mathbf{X}}_{p_z}\,\right] \quad (13)$$

where the right-hand side of the equation is the structure and motion in some projective frame. Depths in system 13 can be arbitrarily row- and column-wise rescaled [13]. However, whatever scaling is chosen, the connection to the scaling of the overall system of depths for all the data, λ_p^i, can be written as

$$\begin{bmatrix}r^{ij}[\lambda_{p_1}^i \ldots \lambda_{p_z}^i] \\ s^{ij}[\lambda_{p_1}^j \ldots \lambda_{p_z}^j]\end{bmatrix} = \left[c_1^{ij}\begin{pmatrix}\gamma_1^{ij,1}\\ \gamma_1^{ij,2}\end{pmatrix}\ldots c_z^{ij}\begin{pmatrix}\gamma_z^{ij,1}\\ \gamma_z^{ij,2}\end{pmatrix}\right] \quad (14)$$

where r, s and c are some non-zero scalars defined for each image pair ij individually. Eq. 14 relates all equivalent scalings corresponding to one class of projective reconstructions. System 14 consists of 2 by z equations. c's can be eliminated by dividing one row by the other:

$$r^{ij}/s^{ij}\left[\begin{array}{ccc}\frac{\lambda_{p_1}^i}{\lambda_{p_1}^j} & \frac{\lambda_{p_2}^i}{\lambda_{p_2}^j} & \ldots\ \frac{\lambda_{p_z}^i}{\lambda_{p_z}^j}\end{array}\right] = \left[\begin{array}{ccc}\frac{\gamma_1^{ij,1}}{\gamma_1^{ij,2}} & \ldots & \frac{\gamma_z^{ij,1}}{\gamma_z^{ij,2}}\end{array}\right].$$

After substituting unknowns r^{ij} and s^{ij} by $\alpha^{ij} = r^{ij}/s^{ij}$ and knowns γ's by $g_p^{ij} = \frac{\gamma_p^{ij,1}}{\gamma_p^{ij,2}}$, the equations can be rewritten as

$$\alpha^{ij}\left[\,\lambda_{p_1}^i\ \ldots\ \lambda_{p_z}^i\,\right] = \left[\,g_1^{ij}\lambda_{p_1}^j\ \ldots\ g_z^{ij}\lambda_{p_z}^j\,\right]. \quad (15)$$

These z equations are bilinear in unknowns α^{ij} and λ's. They can be "linearized" by applying logarithm to both sides of the equations, which is a reasonable operation because both α and λ's can be expected to be (i) positive due to oriented projective geometry (cheirality) [15] and (ii) close to one, see figure 3a, where the log function well approximates function $x-1$, see figure 3b:

$$\log\alpha^{ij} + \left[\,\log\lambda_{p_1}^i\ \ldots\ \log\lambda_{p_z}^i\,\right] = \left[\,\log g_1^{ij} + \log\lambda_{p_1}^j\ \ldots\ \log g_z^{ij} + \log\lambda_{p_z}^j\,\right]. \quad (16)$$

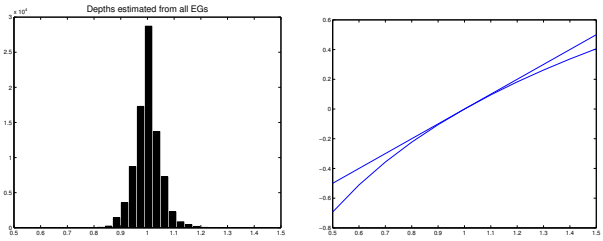

Figure 3. Depths estimated from all EGs: (left) depths $\gamma_p^{ij,k}$ in the St. Martin rotunda balanced to be close to one (right) Change in scaling after applying logarithm: $\log x$ well approximates $x-1$

After substituting

$$\overline{\alpha} = \log\alpha, \overline{\lambda} = \log\lambda, \overline{g} = \log g, \quad (17)$$

(16) can be rewritten to

$$\overline{\alpha}^{ij} + \left[\,\overline{\lambda}_{p_1}^i\ \ldots\ \overline{\lambda}_{p_z}^i\,\right] = \left[\,\overline{g}_1^{ij} + \overline{\lambda}_{p_1}^j\ \ldots\ \overline{g}_z^{ij} + \overline{\lambda}_{p_z}^j\,\right].$$

Let all unknowns be rearranged to the left-hand side:

$$\begin{aligned}\overline{\alpha}^{ij} + \overline{\lambda}_{p_1}^i - \overline{\lambda}_{p_1}^j &= \overline{g}_1^{ij}\\ &\vdots\\ \overline{\alpha}^{ij} + \overline{\lambda}_{p_z}^i - \overline{\lambda}_{p_z}^j &= \overline{g}_z^{ij}.\end{aligned} \quad (18)$$

After solving system 18, both $\overline{\lambda}$'s and $\overline{\alpha}$'s can be computed and back-substituted using (17). System 18 is sparse and hence can be solved efficiently by a sparse solver.

The MM of all scenes in this paper except the Dinosaur sequence were obtained from pair-wise matches satisfying EGs between distinguished regions of various types detected in image pairs in a way similar to [3]. The threshold on distance to the epipolar lines was set to one pixel. All image triple and image pair constraints with more than some given number of points were used. We tried also using only some of them with similar results. However, triple constraints turned out to be necessary for reaching a precise reconstruction, which is essential for metric upgrade, as the partial reconstruction from an image triple is better constrained.

In this paper, only results of the gluing via cameras are shown for the perspective model. It seems that gluing via points cannot be used in conjunction with the perspective model. At least we did not achieve any reasonable result using our implementation. Reconstructions from some minimal set of 58 image triple constraints (i.e. $m-2$) and from 166 triple constraints are shown in figure 4. In fig. 4 left, cameras 21 and 22 are reconstructed very far from each other compared to the surrounding cameras. This is because

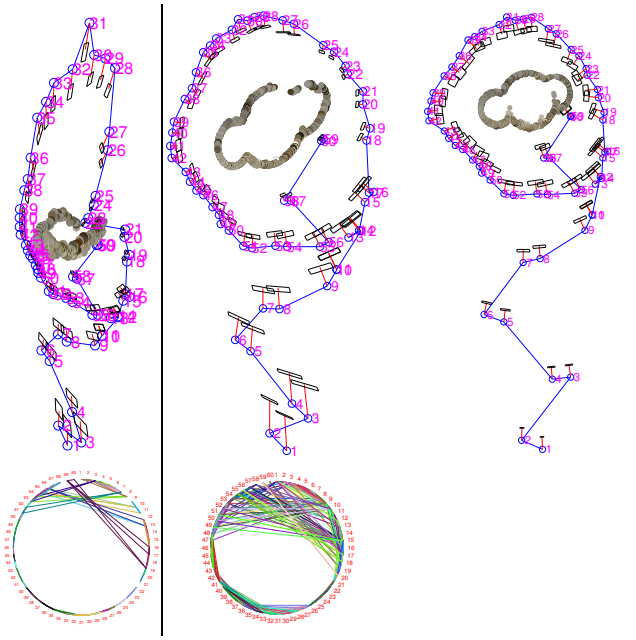

Figure 4. Reconstruction of the St. George rotunda captured on 60 images, 92% data missing: (left) from a minimal set of 58 triple constraints (right) from 166 triples and after Euclidean BA

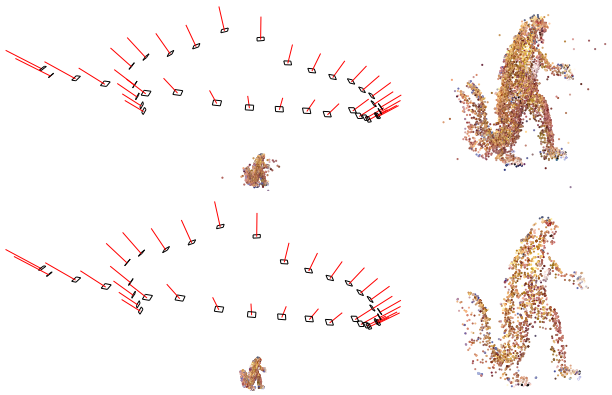

Figure 5. Metric reconstruction of the Dinosaur open sequence using the perspective camera model. Points reconstructed from the whole tracks (top) and from the tracks via images in the used image triple constraints (bottom). The mean reprojection errors are 0.50 pxl (top) and 0.25 pxl (bottom). Note that the outlying points in (top) could prevent converging BA to the global minimum

no constraint on camera pair 21 and 22 was used. Thus, it is clear that exploiting the cyclical structure of the data helps much in constraining the reconstruction. Recall, this would not be possible without depth consistency with EGs in a graph with cycles. Reconstruction using depths consistent only with EGs in an acyclic graph would look much worse than that in fig. 4 left.

An example of a wide base-line scene can be seen in fig. 1. The St. Martin rotunda is very difficult to reconstruct because (i) both overview and detailed images are present (see top of fig. 1a), (ii) some cameras are positioned very close to each other while some are very distant with wide base-lines (see middle of fig. 1a), (iii) it is a closed sequence around an object but at the same time there are many additional cycles (see bottom of fig. 1a), making the task perhaps unsolvable for sequential algorithms. The strong perspective effects make the task perhaps unsolvable for batch method [4] as it assumes affine cameras and slow motion.

4.2. Metric Reconstruction

Robust state-of-the art metric upgrade [11] was applied. However, if some cameras did not move along a fluent path with roughly the same distances between the consecutive frames, see first 8 cameras in figure 4 right, the Nister's preconditioning based on this assumption could not provide a starting point sufficient for his optimization process to reach a good minimum. Thus, for non video-sequences, exhaustive search of the plane at infinity by sampling the space of its possible positions [6] was used instead. Even better results were achieved when exploiting the knowledge of ratios of focal lengths in the criterion function.

After the metric upgrade, most 3D points had positive the fourth coordinate. Only these were used in Euclidean bundle adjustment. Intrinsic parameters of all cameras were set to square pixel, principle point at image center and focal lengths to known ratios. The BA was done on a few points from each sampled submatrices $\mathbf{y}^{\mathbf{i}_t}_{\mathbf{p}_t}$ with 3D points parameterized so that the fourth coordinate equals one. Because each bundled point was visible in two or three images only, there could be no outliers across many images (see figure 5) which could significantly obstruct converging to the global minimum. Results of the Euclidean BA can be seen in figures 1 and 4 right.

This method provides a complete internal and external camera calibration and a sparse set of reconstructed points. Cameras can be used for dense reconstruction as in [3]. Fig. 6 shows examples of disparity maps computed by method [8] on the Dinosaur and the St. George rotunda. The density approximation to point clouds, so called "fish-scales", shows that point clouds from individual image pairs fluently fade one into another thanks to correct gluing of partial reconstructions. For the Dinosaur, absence of any rough transition suggests reaching the global minimum since we did not use the constraint that the sequence was closed but the result is a closed camera trajectory.

Solving (10) using Matlab 6.5's EIGS took 0.25 seconds for 60 images of the St. George rotunda (PentiumIV@2.8GHz). Solving (18) using Matlab's QMR took about one minute even for about 100 000 unknown projec-

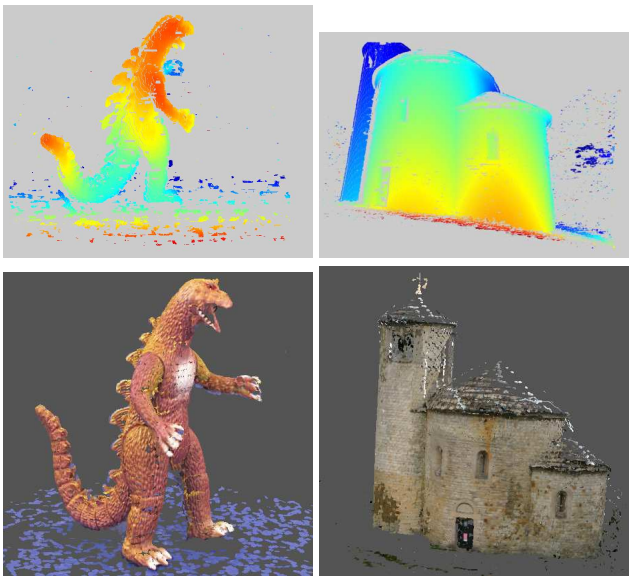

Figure 6. Application to dense matching (top) and dense reconstruction (bottom) on the Dinosaur sequence and the St. George rotunda: density of disparity map shows that the epipolar lines are correct

tive depths. This time was reduced to seconds by sampling only a few points from each submatrix while achieving similar results.

5. Discussion and Conclusions

A new method for fitting a low-rank matrix to a matrix with missing data was presented. Its correctness was demonstrated on an application to 3D reconstruction. In this approach, both affine a perspective camera models can be used. Affine model has the advantage of simplicity and stability if used on images taken by a distant camera. A linear method [4] is sufficient to get internal parameters close to the real ones to initiate the Euclidean BA. On the other hand, the model gives high reprojection errors for a wider field of view. This does not happen when using the perspective camera model. Even very wide base-line scenes are reconstructed with reprojection errors around one pixel already by the linear method. Although using projective depths in the richer perspective model brings necessity to estimate them, we showed that it is possible to estimate them reliably and consistently with all used EGs. Moreover, it has been shown that the richer perspective camera model does not overfit when used in our method.

There is a certain similarity between our method and Locally Linear Embedding (LLE) [12], although the tasks substantially differ. Our method is global in the same sense as LLE. Once the local structures (partial reconstructions) are chosen and fixed, they are combined by solving one optimization problem which has a global minimum as the eigenvalue problem is solved.

The perspective model can model omnidirectional cameras once points in omnidirectional images are attached to rays in space [10]. Other applications with missing data are possible, e.g. 3D reconstruction of non-rigid scenes. See more reconstructed scenes at [1].

References

[1] http://cmp.felk.cvut.cz/~martid1/demoCVPR05.

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

[3] 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 *Proc ECCV Workshop Statistical Methods in Video Processing*, volume LNCS 3247, pages 1–12, Prague, Czech Republic, May 2004.

[4] N. Guilbert and A. Bartoli. Batch recovery of multiple views with missing data using direct sparse solvers. In *Proceedings of the British Machine Vision Conference*, 2003.

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

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

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

[8] J. Kostková and R. Šára. Stratified dense matching for stereopsis in complex scenes. In *BMVC 2003: Proceedings of the 14th British Machine Vision Conference*, volume 1, pages 339–348, Norwich, UK, September 2003.

[9] D. Martinec and T. Pajdla. Structure from many perspective images with occlusions. In *Proc. of the European Conference on Computer Vision (ECCV)*, volume II, pages 355–369, Copenhagen, Denmark, May 2002.

[10] B. Mičušik, D. Martinec, and T. Pajdla. 3d metric reconstruction from uncalibrated omnidirectional images. In *Proc. of the Asian Conference on Computer Vision (ACCV)*, volume 1, pages 545–550, Jeju Island, Korea, January 2004.

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

[12] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*. To appear.

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

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

[15] T. Werner and T. Pajdla. Oriented matching constraints. In *British Machine Vision Conference 2001*, pages 441–450, Manchester, UK, September 2001.