

# Structure from Many Perspective Images with Occlusions

Daniel Martinec and Tomáš Pajdla\*

Center for Machine Perception  
Department of Cybernetics  
Czech Technical University in Prague  
Karlovo nám. 13, 121 35 Praha, Czech Republic  
{martid1, pajdla}@cmp.felk.cvut.cz

**Abstract.** This paper proposes a method for recovery of projective shape and motion from multiple images by factorization of a matrix containing the images of all scene points. Compared to previous methods, this method can handle perspective views and occlusions jointly. The projective depths of image points are estimated by the method of Sturm & Triggs [11] using epipolar geometry. Occlusions are solved by the extension of the method by Jacobs [8] for filling of missing data. This extension can exploit the geometry of perspective camera so that both points with known and unknown projective depths are used. Many ways of combining the two methods exist, and therefore several of them have been examined and the one with the best results is presented. The new method gives accurate results in practical situations, as demonstrated here with a series of experiments on laboratory and outdoor image sets. It becomes clear that the method is particularly suited for wide base-line multiple view stereo.

**Keywords:** projective reconstruction, structure from motion, wide base-line stereo, factorization

## 1 Introduction

In the past geometric and algebraic relations among uncalibrated views up to four in number have been described [5]. Various algorithms for scene reconstruction with both orthographic and perspective camera have been proposed [5,12,8,11,6,9,13,3,10]. The reconstruction problem from orthographic camera is conceptually satisfactorily solved but this could not be claimed for the case of a perspective camera. The biggest problem that remained to be solved was dealing consistently with scene occlusions.

---

\* This research was supported by the grants GACR 102/00/1679, MSMT KONTAKT 2001/09 and ME412, and MSM 212300013. Andrew Zisserman from the University of Oxford kindly provided the Dinosaur data, Marc Pollefeys from K.U.Leuven the Temple data, and Tomáš Werner from the University of Oxford provided the routine for the bundle adjustment.

**Table 1.** Comparison of some 3D reconstruction methods. Lexicographical ordering was used so that (i) the importance of a criterion decreases from the first to the last column and (ii) the quality of the method decreases from top to down

Algorithm	views	camera	occlusions	privileged data	depends on im. ordering
<b>the new algorithm</b>	<b>N</b>	<b>persp.</b>	<b>yes</b>	<b>no</b>	<b>no</b>
Fitzgibbon & Zisserman [3]	N	persp.	yes	no	yes
Avidan & Shashua [10]	N	persp.	yes	no	yes
Urban et al. [13]	N	persp.	yes	central view	no
Heyden [6]	N	persp.	no	no	no
Mahamud & Hebert [9]	N	persp.	no	weak persp.	no
Sturm & Triggs [11]	N	persp.	no	no	yes
Jacobs [8]	N	orthog.	yes	no	no
Tomasi & Kanade [12]	N	orthog.	yes	initial submatrix	no
Hartley & Zisserman [5]	2,3,4	orthog. persp.	no	no	no

This paper offers a linear method which extends and suitably combines previous methods so that the reconstruction in an entirely general situation, i.e. many images with perspective camera and occlusions, is possible. A review of previous works follows.

The situation is similar for two, three, and four uncalibrated images. 3D structure of a scene can be recovered up to an unknown projective transformation, where the camera geometry can be represented by the fundamental matrix, the trifocal, and the quadrifocal tensor respectively [5].

For any number of images, image coordinates of the projections of 3D points can be combined into a so called *measurement matrix*. Tomasi and Kanade [12] developed a factorization method of the measurement matrix for scene reconstruction with an orthographic camera and Sturm and Triggs [11] extended this method from affine to perspective projections. Heyden's method [6] uses a different approach. It relies only on subspace methods instead of multilinear constraints. Similarly, Mahamud & Hebert proposed a method [9], which computes projective depths iteratively but it can only be used for weak perspective or for full perspective with a good initial depth estimate.

Occlusions present a significant problem for reconstruction. The above mentioned Tomasi and Kanade's method solves this problem under the orthographic projection but the result depends on the choice of some initial submatrix of the measurement matrix. The method is iterative and errors may increase gradually with the number of iterations. Jacobs' method [8] improves the above approach so that no initial submatrix is needed. He combines constraints on the reconstruction derived from small submatrices of the full measurement matrix. It treats all data uniformly and is independent of image ordering.

Under perspective projection, the occlusion problem has not yet been generally solved. Method [13] by Urban et al. is dependent on the choice of a central image, which is combined with other images in a so called “cake” configuration. Only points whose projections are contained in the central image can be reconstructed. Method [3] by Fitzgibbon & Zisserman and [10] by Avidan & Shashua compute reconstruction from a sequence of images using trifocal tensors and fundamental matrices. Subsequent images are taken one after another and used to extend and improve actual reconstruction. Table 1 summarizes the differences among the mentioned methods.

Jacobs [8] solves reconstruction with occlusions for orthographic camera, Sturm & Triggs [11] solve reconstruction without occlusions for perspective camera. We present a novel method that builds on these two methods so that scene reconstruction from many perspective images with occlusions is obtained. Our method is independent of image ordering and treats all data uniformly up to a certain level of missing data. Above this level, the reconstruction process may prefer some data depending on the method of computing the projective depths.

The paper is organized as follows. The reconstruction problem is formulated in Section 2. In Section 3.1 and 3.2, algorithms [11] and [8] are reviewed, respectively. In 3.3, the new filling algorithm is presented. In 3.4, the new reconstruction method is proposed. Experiments with artificial and real data are presented in sections 5 and 6. Section 7 gives suggestions for future work.

## 2 Problem Formulation

Suppose a set of  $n$  3D points and that some of them are visible in  $m$  perspective images. The goal is to recover 3D structure (point locations) and motion (camera locations) from the image measurements. This recovery will be called *scene reconstruction*. No camera calibration or additional 3D information will be assumed, so it will be possible to reconstruct the scene up to a projective transformation of the 3D space.

Let  $\mathbf{X}_p$  be the unknown homogeneous coordinate vectors of the 3D points,  $\mathbf{P}^i$  the unknown  $3 \times 4$  projection matrices, and  $\mathbf{x}_p^i$  the measured homogeneous coordinate vectors of the image points, where  $i = 1, \dots, m$  labels images and  $p = 1, \dots, n$  labels points. Due to occlusions,  $\mathbf{x}_p^i$  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  $\lambda_p^i$ , which will be called (*projective*) *depths*:

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

The complete set of image projections can be gathered into a matrix equation:

$$\underbrace{\begin{bmatrix} \lambda_1^1 \mathbf{x}_1^1 & \lambda_2^1 \mathbf{x}_2^1 & \dots & \lambda_n^1 \mathbf{x}_n^1 \\ \times & \lambda_2^2 \mathbf{x}_2^2 & \dots & \times \\ \vdots & & \ddots & \vdots \\ \lambda_1^m \mathbf{x}_1^m & \times & \dots & \lambda_n^m \mathbf{x}_n^m \end{bmatrix}}_{\mathbf{R}} = \underbrace{\begin{bmatrix} \mathbf{P}^1 \\ \vdots \\ \mathbf{P}^m \end{bmatrix}}_{\mathbf{P}} \underbrace{\begin{bmatrix} \mathbf{X}_1 & \dots & \mathbf{X}_n \end{bmatrix}}_{\mathbf{X}}$$

where marks  $\times$  stand for unknown elements which could not be measured due to occlusions,  $\mathbf{X}$  and  $\mathbf{P}$  stand for structure and motion, respectively. The  $3m \times n$  matrix  $[\mathbf{x}_p^i]_{i=1..m, p=1..n}$  will be called the *measurement matrix* whereas  $\mathbf{R}$  will be called the *partially rescaled measurement matrix*, shortly PRMM, because  $\mathbf{R}$  will be used even with some unknown depths. Both measurement matrix and PRMM may have (and in most cases do have) some missing elements.

### 3 The Main Idea of the New Reconstruction Algorithm

A complete rescaled measurement matrix has rank four and therefore a projective reconstruction can be obtained by its factorization. However, from measurements in perspective images with occlusions, we can only compose a measurement matrix which is neither complete nor rescaled. When it is at all possible to compute projective depths of some known points in  $\mathbf{R}$ , e.g. via multi-view constraints, some missing elements of  $\mathbf{R}$  can often be filled using the knowledge that every five columns of complete rescaled  $\mathbf{R}$  are linearly dependent.

It would be ideal to first compute the projective depths of all known points in  $\mathbf{R}$  and then to fill all the missing elements of  $\mathbf{R}$  by finding a complete matrix of rank four that would be equal (or as close as possible) to the rescaled  $\mathbf{R}$  in all elements where  $\mathbf{R}$  is known. Such a two-step algorithm is almost the ideal linearized reconstruction algorithm, which uses all data and has a good statistical behavior. We have found that many image sets, in particular those resulting from wide base-line stereo, can be reconstructed in such two steps.

Of course, there are image sets, e.g. sets with the structure of missing data on the borderline of reconstructibility or long sequences with very fractionalized tracks, which cannot be solved in the above two steps. Instead, the two steps have to be repeated while the measurement matrix  $\mathbf{R}$  is not complete. If the correspondences between the images are such that the measurement matrix is large and diagonally dominant, then it is possible to use another reconstruction technique, e.g. to fuse partial consecutive reconstructions [3,10]. However, if there is no clear sequence of images or central image like in [13], the proposed algorithm has a clear advantage. It can handle arbitrary scenes in pseudo-optimal manner without a priori preferring any particular image. It provides a unique solution and thus is suited for the initialization of bundle adjustment optimizations.

In what follows, we shall describe the two steps of the algorithm. Let us first review the two steps we build on and their respective extensions. Later we will describe how to combine the two steps.

#### 3.1 Estimating the Projective Depths

Many works dealt with estimating the projective depths. In this work, we used Sturm & Triggs' method [11] exploiting epipolar geometry but other methods, e.g. [6,9,5], can be applied also. The method [11] was proposed in two alternatives. The alternative with a central image is more appropriate for wide base-line stereo while the alternative with a sequence is more appropriate for video-sequences. The former will be denoted as  $\omega_{cent,c}$  where  $c$  denotes the number

of a central image while the latter will be denoted as  $\omega_{seq}$ . Thus, we have altogether the totality  $\Omega = \{\omega_{seq}, \omega_{cent,1} \dots \omega_{cent,m}\}$  of alternatives for computing the projective depths. Also, the method from [11] has to be furthermore slightly modified on account of missing data. The complete algorithm is summarized in Algorithm 1. The  $p$ -th track there denotes a subsequence of known points in sequence  $\mathbf{x}_p^1 \dots \mathbf{x}_p^m$ .

1. Set  $\lambda_p^j = 1$  for all  $p$  corresponding to known points  $\mathbf{x}_p^j$  in view  $j = \begin{cases} 1 : & \text{for } \omega_{seq} \\ c : & \text{for } \omega_{cent,c} \end{cases}$
2. For  $\begin{cases} j = 1 \dots m - 1, i = j + 1 : & \text{for } \omega_{seq} \\ j = c, i \neq j : & \text{for } \omega_{cent,c} \end{cases}$  do the following. If images  $i$  and  $j$  have enough points in common to compute a fundamental matrix uniquely<sup>a</sup> then compute their fundamental matrix  $F^{ij}$ , epipole  $\mathbf{e}^{ij}$ , and depths  $\lambda_p^i$  according to
 
$$\lambda_p^i = \frac{(\mathbf{e}^{ij} \wedge \mathbf{x}_p^i) \cdot (F^{ij} \mathbf{x}_p^j)}{\|\mathbf{e}^{ij} \wedge \mathbf{x}_p^i\|^2} \lambda_p^j$$

if the right side of the equation is defined, where  $\wedge$  stands for the cross-product.

For  $\omega_{seq}$ : if the  $p$ -th track ( $p = 1 \dots n$ ) is discontinuous, start with  $j = b(p)$  where  $b(p)$  denotes the initial image of the longest continuous subtrack of the  $p$ -th track.

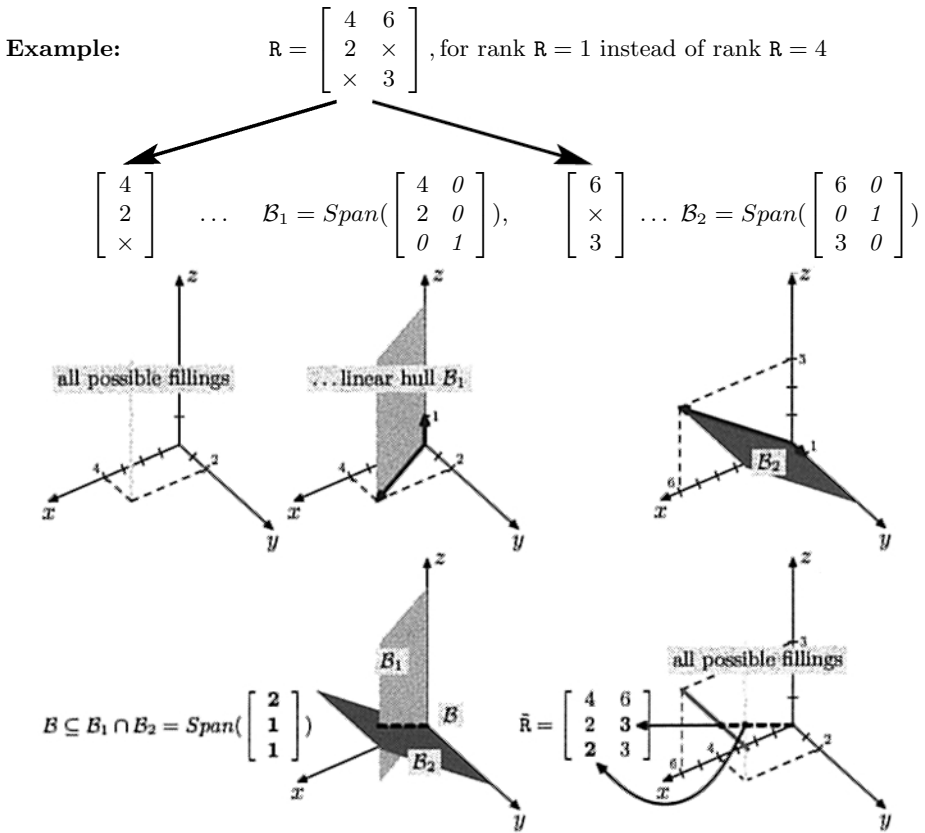
<sup>a</sup> See Section 3.4.

**Algorithm: 1.** Estimating the depths: alternatives  $\omega_{seq}$  and  $\omega_{cent,c}$

### 3.2 Filling of Missing Elements in R

Filling of missing data was first realized by Tomasi & Kanade [12] for orthographic camera. D. Jacobs [8] improved their method and we use our extension of his method for the perspective case. Often, not all depths can be computed because of missing data. Therefore, we extend the method from [8] so that also points with unknown depths are exploited. Moreover, the extension is independent of how depths are estimated and thus any method for estimating the depths could be used. Before describing our modification for the perspective camera, the original Jacobs' algorithm for the orthographic case has to be explained.

D. Jacobs treated the problem of missing elements in a matrix as fitting an unknown matrix of a certain rank to an incomplete noisy matrix resulting from measurements in images. Assume noiseless measurements for a while to make the explanation simpler. Assuming perspective images, an unknown complete  $3m \times n$  matrix  $\tilde{\mathbf{R}}$  of rank 4 is fitted to PRMM  $\mathbf{R}$ . Technically, a basis of the linear vector space that is spanned by the columns of  $\tilde{\mathbf{R}}$  is searched for. Thus, when there are 4 complete linearly independent columns in  $\mathbf{R}$ , then they form the desired basis. When no such 4-tuple of columns exists, the basis has to be constructed from incomplete columns. Fortunately, some 4-tuples of incomplete columns provide constraints on the basis and a sufficient number of such constraints determine it.



**Fig. 1.** Forming constraints on the basis and filling the matrix. For  $R$  is of rank 1, constraints on  $\mathcal{B}$  are formed by single columns

Let us explain what we mean by saying that an incomplete column  $c$  of  $R$  spans (generates) a subspace. Every complete column of  $R$  generates a one-dimensional subspace of  $\mathbb{R}^{3m}$ . Thus, an incomplete  $c$  generates a subspace  $V$ , as the smallest linear space containing all one-dimensional subspaces generated by  $c$  after replacing unknown elements by some arbitrary real numbers. Linear subspaces form a complete lattice [2] and therefore such smallest linear space  $V$  exists. It is a subspace of  $\mathbb{R}^{3m}$  and equals the linear hull of all one-dimensional subspaces. The generators of  $V$  can be obtained by constructing the column containing the known elements of  $c$  and zeros instead of the unknown ones and augmenting it with the standard basis spanning the dimensions of the unknown elements (see Fig. 1 and the example in Section 3.3).

Let the space generated by the columns of  $\tilde{R}$  be denoted by  $\mathcal{B}$ . Let  $\mathcal{B}_t$  denotes the span of the  $t$ -th 4-tuple of columns of  $R$  which are linearly independent in coordinates known in all four columns.  $\mathcal{B}$  is included in each  $\mathcal{B}_t$  and thus also

in their intersection i.e.  $\mathcal{B} \subseteq \bigcap_{t \in T} \mathcal{B}_t$ , where  $T$  is some set of indices. When the intersection is 4D,  $\mathcal{B}$  is known exactly. If it is of a higher dimension, only an upper bound on  $\mathcal{B}$  is known and more constraints from 4-tuples must be added. Any column in  $\tilde{\mathbf{R}}$  is a linear combination of vectors of a basis of  $\tilde{\mathbf{R}}$ . Thus, having a basis  $\mathbf{B}$  of  $\tilde{\mathbf{R}}$ , any<sup>1</sup> incomplete column  $c$  in  $\mathbf{R}$  can be completed by finding the vector  $\tilde{c}$  generated by  $\mathbf{B}$  which equals  $c$  in the elements where  $c$  was known in  $\mathbf{R}$  (see Fig. 1).

Linear independency of the 4-tuple of columns is crucial to obtain a valid constraint on the basis. Consider, e.g., a 4-tuple consisting of four equal columns, thus spanning only a 1D space. Even if three coordinates in one of its columns are made unknown, and thus a 4D space is spanned,  $\mathcal{B}$  does not have to be included in the span. A row with some missing coordinates can be ignored because the entire corresponding dimension is spanned and the constraint on  $\mathcal{B}$  is always satisfied in the dimension, meaning such a row contains no information. This is the reason to use just the 4-tuples of columns linearly independent in coordinates known in all four columns.

Because of noise in real data, the intersection  $\bigcap_{t \in T} \mathcal{B}_t$  quickly becomes empty. This is why  $\mathcal{B}$  is searched for as the closest 4D space to spaces  $\mathcal{B}_t$  in the sense of the minimal sum of square differences of known elements. Denoting complement of a linear vector space by  $\perp$ ,  $\bigcap_{t \in T} \mathcal{B}_t$  can be expressed according to the well known De Morgan rule as  $(Span_{t \in T} \mathcal{B}_t^\perp)^\perp$ . The generators of  $\mathcal{B}_t^\perp$  can be found as  $\mathbf{B}_t^\perp = \mathbf{u}(:, d + 1 : \text{end})$ , where  $[\mathbf{u}, \mathbf{s}, \mathbf{v}] = \text{svd}(\mathbf{B}_t)$  and  $d$  is the dimension of  $\mathcal{B}_t$ .  $Span_{t \in T} \mathcal{B}_t^\perp$ , where  $T$  is of cardinality  $z$ , is generated by  $[\mathbf{B}_1^\perp \mathbf{B}_2^\perp \dots \mathbf{B}_z^\perp]$ .  $(Span_{t \in T} \mathcal{B}_t^\perp)^\perp$  is generated by  $\mathbf{u}(:, \text{end} - 3 : \text{end})$ , where  $[\mathbf{u}, \mathbf{s}, \mathbf{v}] = \text{svd}([\mathbf{B}_1^\perp \mathbf{B}_2^\perp \dots \mathbf{B}_z^\perp])$ .

### 3.3 Filling of Missing Elements for Perspective Cameras

Jacobs' method [8] cannot use image points with unknown depths. But, PRMM constructed from measurements in perspective images often has many such points where the corresponding depths cannot be computed. Therefore, we extended the method to exploit also points with unknown depths. It brings two advantages: (i) because the actual iteration of the two-step algorithm exploits more information, the number of iterations may decrease and consequently more accurate results may be obtained; (ii) it is possible to reconstruct more scene configurations. See Section 8 in [1] for more details about this. It is important that the proposed extension is still a linear method as was the Jacobs' method [8].

Let us first explain the extension for two images. Suppose that  $\lambda_p^i$  and  $\mathbf{x}_p^i$  are known for  $i = 1, 2$ , and for  $p = 1 \dots 4$  except  $\lambda_4^2$ . Then, consider the first four columns of  $\mathbf{R}$  to be the  $t$ -th 4-tuple of columns,  $\mathbf{A}_t$ . A new matrix  $\mathbf{B}_t$ , whose span will be denoted by  $\mathcal{B}_t$ , can be defined using known elements of  $\mathbf{A}_t$  as

$$\mathbf{A}_t = \begin{bmatrix} \lambda_1^1 \mathbf{x}_1^1 & \lambda_2^1 \mathbf{x}_2^1 & \lambda_3^1 \mathbf{x}_3^1 & \lambda_4^1 \mathbf{x}_4^1 \\ \lambda_1^2 \mathbf{x}_1^2 & \lambda_2^2 \mathbf{x}_2^2 & \lambda_3^2 \mathbf{x}_3^2 & ? \mathbf{x}_4^2 \end{bmatrix} \longrightarrow \mathbf{B}_t = \begin{bmatrix} \lambda_1^1 \mathbf{x}_1^1 & \lambda_2^1 \mathbf{x}_2^1 & \lambda_3^1 \mathbf{x}_3^1 & \lambda_4^1 \mathbf{x}_4^1 & 0 \\ \lambda_1^2 \mathbf{x}_1^2 & \lambda_2^2 \mathbf{x}_2^2 & \lambda_3^2 \mathbf{x}_3^2 & 0 & \mathbf{x}_4^2 \end{bmatrix}$$

<sup>1</sup> containing at least four known elements, which in practice means six elements resulting from two known points

It can be proved (see Corollary 1 in Appendix A in [1]) that if  $B_t$  is of full rank (i.e. five here) then  $\mathcal{B} \subseteq Span(B_t)$ , which is exactly the constraint on  $\mathcal{B}$ .

In a general situation there are also some missing elements in  $R$ . Then, the matrix  $B_t$  is constructed from the  $t$ -th 4-tuple  $A_t$  of columns of  $R$  as follows:

1. Set  $B_t$  to  $A_t$ .
2. Replace all unknown points and points with unknown depth by zero in  $B_t$ .
3. For each unknown depth  $\lambda_p^i$  in  $A_t$ , add a column with  $\mathbf{x}_p^i$  and zeros everywhere else to  $B_t$ .
4. For each triple of rows in  $A_t$  containing some unknown point, add to  $B_t$  the standard basis spanning the dimensions of the unknown point.

The following example demonstrates the construction of  $B_t$  from a 4-tuple  $A_t$ :

$$\begin{array}{ccc}
 A_t = \begin{bmatrix} ? & \mathbf{x}_1^1 & \lambda_2^1 \mathbf{x}_2^1 & \lambda_3^1 \mathbf{x}_3^1 & \lambda_4^1 \mathbf{x}_4^1 \\ \lambda_1^2 \mathbf{x}_1^2 & \lambda_2^2 \mathbf{x}_2^2 & \lambda_3^2 \mathbf{x}_3^2 & \lambda_4^2 \mathbf{x}_4^2 \\ \lambda_1^3 \mathbf{x}_1^3 & \lambda_2^3 \mathbf{x}_2^3 & \lambda_3^3 \mathbf{x}_3^3 & \times \end{bmatrix} & \xrightarrow{2} & \begin{bmatrix} \mathbf{0} & \lambda_2^1 \mathbf{x}_2^1 & \lambda_3^1 \mathbf{x}_3^1 & \lambda_4^1 \mathbf{x}_4^1 \\ \lambda_1^2 \mathbf{x}_1^2 & \lambda_2^2 \mathbf{x}_2^2 & \lambda_3^2 \mathbf{x}_3^2 & \lambda_4^2 \mathbf{x}_4^2 \\ \lambda_1^3 \mathbf{x}_1^3 & \lambda_2^3 \mathbf{x}_2^3 & \lambda_3^3 \mathbf{x}_3^3 & \mathbf{0} \end{bmatrix} \\
 & & \downarrow 3 \\
 B_t = \begin{bmatrix} \mathbf{0} & \mathbf{x}_1^1 & \lambda_2^1 \mathbf{x}_2^1 & \lambda_3^1 \mathbf{x}_3^1 & \lambda_2^1 \mathbf{x}_2^1 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \lambda_1^2 \mathbf{x}_1^2 & \mathbf{0} & \lambda_2^2 \mathbf{x}_2^2 & \lambda_3^2 \mathbf{x}_3^2 & \lambda_4^2 \mathbf{x}_4^2 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \lambda_1^3 \mathbf{x}_1^3 & \mathbf{0} & \lambda_2^3 \mathbf{x}_2^3 & \lambda_3^3 \mathbf{x}_3^3 & \mathbf{0} & \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} & \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} & \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \end{bmatrix} & \xleftarrow{4} & \begin{bmatrix} \mathbf{0} & \mathbf{x}_1^1 & \lambda_2^1 \mathbf{x}_2^1 & \lambda_3^1 \mathbf{x}_3^1 & \lambda_4^1 \mathbf{x}_4^1 \\ \lambda_1^2 \mathbf{x}_1^2 & \mathbf{0} & \lambda_2^2 \mathbf{x}_2^2 & \lambda_3^2 \mathbf{x}_3^2 & \lambda_4^2 \mathbf{x}_4^2 \\ \lambda_1^3 \mathbf{x}_1^3 & \mathbf{0} & \lambda_2^3 \mathbf{x}_2^3 & \lambda_3^3 \mathbf{x}_3^3 & \mathbf{0} \end{bmatrix}
 \end{array}$$

If  $B_t$  is of full rank, its span  $\mathcal{B}_t$  includes  $\mathcal{B}$  (this can be proved by induction from Corollary 1 in Appendix A in [1]). By including also image points with unknown projective depths the spaces  $\mathcal{B}_t$  spanned by 4-tuples of columns become smaller, thus solving the complete problem becomes more efficient.

It can be seen that the concept of generating constraints on the basis for the orthographic case is only a special case of generating constraints for the perspective case. The former is equivalent to the latter having all depths set to the same number thus corresponding to the perspective camera with the projection center at infinity and looking at a finite scene.

### 3.4 Combining the Filling Method with Estimating the Depths

Due to occlusions, the computation of projective depths can be carried out in various ways depending on which depths are computed first and if and how those already computed are used to compute the others. One way of depth estimation will be called a *strategy*. Depending on the chosen strategy, different subsets of depths are computed and different submatrices of PRMM are filled. It may happen when some strategy exploiting e.g. epipolar geometry of some pair of images is used that the fundamental matrix cannot be computed due to occlusions. Consequently, depths needed to form a constraint on the basis of PRMM in one of the images cannot be estimated, thus the missing data in the image cannot be filled and the two steps of depth estimation and filling has to be repeated.

For accurate data, all strategies should be equivalent. It is not so if the data is noisy. In such case, the task is to choose the strategy which results in the smallest



error. It would be unrealistically costly to compute all possibilities (although there is “only” a finite number of them) and to choose the best one. Fortunately, we do not have to compute all of them in order to find some good one. From the structure of missing data, it is possible to predict a good strategy for depth estimation that results in a good reconstruction. Some criterion deciding which strategy is good is needed. For scenes reconstructible in more steps, such criterion also determines which subset of depths is better to be computed first.

The following two observations have been made. First, the more iterations are performed, the less accurate results are obtained because the error from the former iteration spreads in subsequent iterations as was also mentioned in [8]. Secondly, unknown elements should not be computed from fewer data when they can be computed from more data, and thus more accurately due to the law of big numbers and supposition of random noise. Both these observations support the following.

**Principle 1** *The more image points that are filled in one step, the smaller the expected error.*

This principle leads to a pseudo-optimal number of iterations that need to be performed.<sup>2</sup> Practically, however, it is not crucial problem that such obtained strategy is only pseudo-optimal because, as will be seen later, it is possible to realize Principle 1 so that, for many scenes, only one iteration is performed. The following proposition holds.

**Proposition 1** *The more depths known before the filling, the smaller the expected error.*

Proof of Proposition 1 inheres in our extension of Jacob’s method (see Appendix B in [1]). Usage of Principle 1 and Proposition 1 in order of their designation proved to be a good criterion. We choose the set of strategies which fill the most points, and from this set, we choose those which scale the most points. From the resulting set, an arbitrary strategy can be used.

The criterion will now be described formally. Let  $\omega$  denote some strategy for estimating the depths and  $\Omega$  denote some set of strategies. Let  $\mathcal{F}(\omega)$  denote the predicted number of newly filled unknown image points during one iteration when  $\omega$  is used. The strategy, for which  $\mathcal{F}(\omega)$  is maximal, is the best strategy according to Principle 1. More such strategies often exist. Let  $\mathcal{S}(\omega)$  denote the predicted number of estimated depths when  $\omega$  is used. According to Proposition 1,  $\mathcal{S}(\omega)$  is maximal for the best strategy. The complete new method for scene reconstruction is summarized in Algorithm 2.

The usefulness of the concept of predictor functions  $\mathcal{F}, \mathcal{S} : \Omega \rightarrow 0..mn$  consists in their ability to be evaluated without neither estimating the depths

---

<sup>2</sup> An optimal strategy would have to be searched for as the shortest branch in the tree graph of all partial solutions. Partial solutions can be ordered into a tree graph. Edges in this graph correspond to chosen strategies and vertices correspond to the partial solutions obtained after one iteration. The root of the tree corresponds to the initial PRMM.

1. Estimate depths using an arbitrary strategy  $\omega^* \in \Omega^*$  where

$$\Omega_{\mathcal{F}} = \left\{ \omega \in \Omega \mid \mathcal{F}(\omega) = \max_{\tau \in \Omega} \mathcal{F}(\tau) \right\}$$

$$\Omega^* = \left\{ \omega \in \Omega_{\mathcal{F}} \mid \mathcal{S}(\omega) = \max_{\tau \in \Omega_{\mathcal{F}}} \mathcal{S}(\tau) \right\}$$

2. Fill the missing data.

Repeat steps 1. and 2. until  $\mathbf{R}$  is complete or no data can be filled in. Then factorize a maximal complete submatrix of  $\mathbf{R}$ .

**Algorithm: 2.** Estimating the depths: alternatives  $\omega_{seq}$  and  $\omega_{cent,c}$

nor data filling. The knowledge of which image points are known or unknown is the only information for the evaluation of  $\mathcal{F}$  and  $\mathcal{S}$ . It is very simple (and fast) but it cannot detect degenerate configurations of points because, in fact, the multi-view tensors are not computed. If it then, when the tensor is computed, turns out that the configuration is degenerate, the second best strategy is used, etc.

To define  $\mathcal{F}$  and  $\mathcal{S}$ , a few symbols have to be introduced. Let logical variable  $x_p^i$  be true if and only if the image point  $\mathbf{x}_p^i$  is known. Let  $i$  and  $j$  be as in step 2 of Algorithm 1. Let  $\mathcal{I}^{ij}$  be true if and only if the data of image  $i$  can be used by the filling method consistently with other images [11]. It is only possible if  $i = j$  or if images  $i$  and  $j$  have enough (at least seven) points in common, which are necessary to compute a fundamental matrix uniquely, thus

$$\mathcal{I}^{ij} \equiv |\{p \mid x_p^i \wedge x_p^j\}| \geq 7 \quad \vee \quad i = j \quad (1)$$

The uniqueness is demanded for the depths consistency with other images. All available points are used for the fundamental matrix estimation. (i) If there are only 7 points, the 7-point algorithm [5] is performed. If it provides three real solutions, the fundamental matrix is not unique. (ii) If there are 8 points or more, the 8-point algorithm [5] is performed. In this case, degenerate configurations can easily be detected.

The predictor functions depend on the way how projective depths are computed. Let us first define the predictor functions for the alternative  $\omega_{cent,c}$  when the depths are computed using a central image  $c$ . Let  $\mathcal{P}_p^c$  be true if and only if the  $p$ -th 3D point can be filled in by the filling method when depths were estimated using strategy  $\omega_{cent,c}$ . To recover a 3D point uniquely from known basis of PRMM, at least two its images are needed. Moreover, it can be proved (see Theorem 4 in Appendix A in [1]) that at least two known depths in each image are needed for the constraints on  $\mathcal{B}$ . It means that  $\mathcal{P}_p^c$  is true if and only if the  $p$ -th 3D point is seen in at least 2 images and the corresponding fundamental matrices, which are needed for estimating at least some two depths in the images, can be computed:

$$\mathcal{P}_p^c \equiv |\{i \mid \mathcal{I}^{ic} \wedge x_p^i\}| \geq 2 \quad (2)$$

Now, predictor functions  $\mathcal{F}$  and  $\mathcal{S}$  can be defined as follows

$$\begin{aligned} \mathcal{F}(\omega_{cent,c}) &= |\{ \langle i, p \rangle \mid \mathcal{I}^{ic} \wedge \mathcal{P}_p^c \wedge \neg x_p^i \}| \\ \mathcal{S}(\omega_{cent,c}) &= |\{ \langle i, p \rangle \mid \mathcal{I}^{ic} \wedge \mathcal{P}_p^c \wedge x_p^i \wedge x_p^c \}| \end{aligned}$$

Term  $\mathcal{I}^{ic} \wedge \mathcal{P}_p^c$  says whether point  $\mathbf{x}_p^i$  can be reconstructed.

Similarly, the predictor functions for alternative  $\omega_{seq}$  when the depths are computed for a sequence are defined as

$$\begin{aligned} \mathcal{P}_p &\equiv |\{i \mid x_p^i\}| \geq 2 \\ \mathcal{F}(\omega_{seq}) &= |\{ \langle i, p \rangle \mid \mathcal{P}_p \wedge \neg x_p^i \}| \\ \mathcal{S}(\omega_{seq}) &= \sum_{p \in 1..n} \maxarg_{k \in b(p)..m} \bigwedge_{i \in b(p)..k} x_p^i \end{aligned} \tag{3}$$

Eq. (3) simply says that the points in the longest continuous subtracks have known depths (See Algorithm 1).

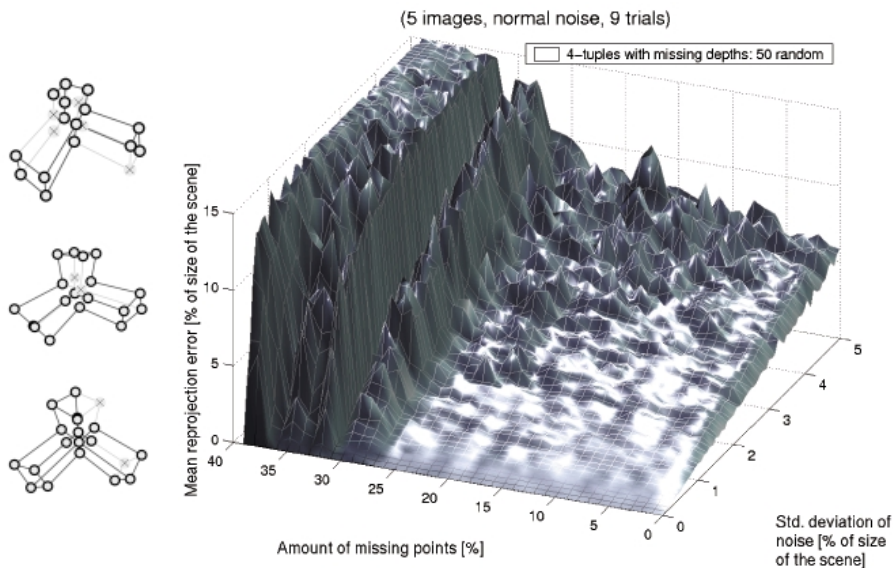
## 4 Implementation Details

On account of good numerical conditioning, several normalizations of the data and balancing similar to those in [11] need to be performed. Choosing of 4-tuples of columns is implemented so that almost each chosen 4-tuple gives the constraint on the basis of PRMM. This is aimed so that columns are chosen one after another. The columns, which cannot provide the constraint with already chosen ones, are temporarily removed from PRMM until the next 4-tuple is chosen. By this way, a good efficiency is achieved.

## 5 Experiments with Artificial Scenes

For experiments with artificial scenes, a simulated scene with cubes was used. The scene models a real scene, hence it represents a generic situation. Twenty points in space were projected by perspective cameras into several images from different locations and directions. Some image points were made unknown to simulate scene occlusions, see the left-hand side of Experiment 1.

Points were taken out from the scene randomly but in a uniform fashion so that, first, the numbers of missing points in each image differed maximally by one, and secondly, the numbers of images of each point differed maximally by one. Points were only removed as long as the whole scene could still be reconstructed. The necessary condition for a complete reconstruction is that each image contains at least 7 points and each point has at least 2 images (see (1) and (2)). The more data available, the higher the percentage of missing data permissible. For this specific experiment, i.e. 20 points in 5 images, 65 % of missing data is the upper bound allowable to get a complete reconstruction. But because of randomly spread holes in data, the actual level of the maximum amount of missing data




Experiment 1: Dependency of reprojection error on noise and missing data

for the complete reconstruction is lower. Experiment 1 shows the dependency of the reprojection error of the reconstruction using Alg. 2 on noise and missing data. Along the left horizontal axis, the amount of the missing data grows while along the right horizontal axis, standard deviation of Gaussian noise of zero mean value added to image points increases. The standard deviation of the added noise as well as the reprojection error is displayed in percentage of the scene size.

If no noise is present, the reconstruction is precise. The reprojection error grows linearly with noise with slope approximately equal one and is almost constant in the direction of missing points up to the level of missing data above which the reconstruction fails. To conclude, the new algorithm is accurate and robust with respect to noise as well as missing data.

## 6 Experiments with Real Scenes


For each experiment, one image, an error table, and the structure of PRMM are provided. The correspondences across the images have been detected either manually or by the Harris interest operator [4]. Besides the scene name and point detection, the table includes the chosen strategy for estimating the depths, the amount of missing data, the number of images used, image sizes, the number of known points in each image, and reprojection errors for our method Algorithm 2 and bundle adjustment initialized by the output of our method. The structure of PRMM shows the exploitation of image points with known ("●") and unknown ("○") projective depths. Empty places stand for unknown points. All scenes have been reconstructed in one iteration of Algorithm 2.

Method	LM = linear method, BA = bundle adj.												
	Scene name	<i>House</i>											
	Point detection	manual											
	Depth estimation	$\omega_{cent,1}$											
	Amount of missing data			<b>47.83 %</b>									
LM	Mean error per image point [pxl]		<b>3.91</b>										
LM + BA	Mean error per image point [pxl]		<b>1.44</b>										
	Image No. [2952×2003]	1	2	3	4	5	6	7	8	85	10		
	Number of corresp.	116	112	857	112	851	785	130	126	101	855		
LM	Maximal error [pxl]	11.0	36.6	12.1	85.3	25.8	15.5	13.6	8.9	14.7	13.4		
LM + BA	Maximal error [pxl]	4.3	6.6	4.5	4.4	5.8	8.3	7.5	6.3	10.7	10.1		
LM	Mean error	2.3	6.8	3.2	2.3	8.1	5.0	2.5	2.3	3.3	4.8		
LM + BA	Mean error	1.1	1.8	1.5	1.2	1.5	1.6	1.2	1.4	1.5	1.8		



size = 10 × 203, " " missing (47.83 %), "•" scaled (75.7 %), "o" not scaled (24.3 %)

Experiment 2: House

Method	LM = linear method, BA = bundle adj.												
	Scene name	<i>Dinosaur (Oxford)</i>											
	Point detection	Harris' operator											
	Depth estimation	$\omega_{seq}$											
	Amount of missing data			<b>90.84 %</b>									
LM	Mean error per image point [pxl]		<b>1.76</b>										
LM + BA	Mean error per image point [pxl]		<b>0.64</b>										
	Image No. [720×576]	1	5	9	13	17	21	25	29	33	36		
	Number of corresp.	257	318	322	516	535	568	602	459	464	381		
LM	Maximal error [pxl]	18.4	16.3	29.5	56.4	46.9	73.9	44.1	28.5	19.4	33.9		
LM + BA	Maximal error [pxl]	10.9	12.7	7.8	41.5	25.7	13.1	13.4	17.3	17.9	21.4		
LM	Mean error	0.6	0.7	2.3	2.0	3.8	1.7	1.4	1.6	1.3	1.0		
LM + BA	Mean error	0.3	0.5	0.6	1.0	1.0	0.4	0.3	0.5	0.9	0.7		



size = 36 × 4983, " " missing (90.84 %), "•" scaled (100.0 %)

Experiment 3: Dinosaur (Oxford)

Method	LM = linear method, BA = bundle adj.					
	Scene name	<i>Temple (Leuven)</i>				
	Point detection	Harris' operator				
	Depth estimation	$\omega_{seq}$				
	Amount of missing data	<b>46.32 %</b>				
LM	Mean error per image point [pxl]	<b>0.49</b>				
LM + BA		<b>0.23</b>				
	Image No. [867×591]	1	2	3	4	5
	Number of corresp.	456	456	297	374	285
LM	Maximal error [pxl]	3.0	2.3	2.8	2.5	3.0
LM + BA		2.5	1.5	2.4	1.8	2.5
LM	Mean error	0.4	0.5	0.6	0.5	0.5
LM + BA		0.3	0.2	0.2	0.2	0.2




---



---

size = 5 × 696, " " missing (46.32 %), "●" scaled (100.0 %)

#### Experiment 4: Temple (Leuven)

The “House” scene (see Experiment 2) was captured on 10 images at high resolution. Approximately 100 points were manually detected in each image. Although 47.83 % data was missing, the reprojection error, given in pixels, is low considering the image sizes. It can be seen that our algorithm could have exploited all known data including 24.3 % unscaled points.

The “Dinosaur” scene (see Experiment 3) was captured on 36 images. Points were detected automatically by the Harris operator. Although the amount of missing data is high (90.84 %), the mean error per image point was lower because of more precise point detection and since 100 % of points were scaled.

The data in Experiment 4 contained outliers, that were removed one after another in the following manner. The scene was first reconstructed with all the data including outliers. Then, the column of PRMM, which contained the point with the highest reprojection error, was discarded. Afterwards, the scene was again reconstructed, another column discarded etc. These two steps were repeated till the highest reprojection error was significant. For the “Temple” scene in Experiment 4, the threshold was set to 4 pixels which lead to discarding 23 out of 719 columns.

To conclude, the new algorithm is enough accurate on real scenes to provide a good initial solution for bundle adjustment.

## 7 Summary and Conclusions

A new linear method for scene reconstruction has been proposed and tested on artificial and real scenes. The method extends and suitably combines previous

methods so that the reconstruction in an entirely general situation, i.e. many images with perspective camera and occlusions, is possible.

A new way of exploiting points with unknown depth was developed. Correctness of this way was proved as well as its abilities and limitations were studied in [1]. Its theoretical asset is the ability to reconstruct linearly some very small scene configurations, which can be reconstructed by other methods only non-linearly (see Theorem 3 in [1]), cannot be reconstructed at all (see Theorem 2 in [1]), or cannot exploit all known data (see Theorem 1 in [1]). Moreover, it gives good results in practical situations as presented here.

The proposed method was intended to deal with several problems in 3D reconstruction. These were the perspective projection, many images, and occlusion. However, one problem was not taken into account explicitly and that is the problem of outliers in correspondences. Although the method was not intended to deal with outliers, it was observed that it can deal with them if they are few compared to the number of inliers (see Experiment 4). To deal well with a bigger amount of outliers, extension [7] of factorization handling outliers can be added.

## References

1. D. Martinec and T. Pajdla. Structure from Many Perspective Images with Occlusions. Research Report CTU-CMP-2001-20, Center for Machine Perception, K333 FEE Czech Technical University, Prague, Czech Republic, July 2001. <ftp://cmp.felk.cvut.cz/pub/cmp/articles/martinec/Martinec-TR-2001-20.pdf>.
2. M. K. Bennett. *Affine and Projective Geometry*. John Wiley and Sons, New York, USA, 1995.
3. A. W. Fitzgibbon and A. Zisserman. Automatic camera recovery for closed or open image sequences. In *Proc. European Conference on Computer Vision*, pages 311–326. Springer-Verlag, June 1998.
4. C. J. Harris and M. Stephens. A combined corner and edge detector. In *Proc. of Alvey Vision Conference*, pages 147–151, 1988.
5. R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, Cambridge, UK, 2000.
6. A. Heyden. Projective structure and motion from image sequences using subspace methods. In *Proc. 10th SCIA*, pages 963–968, June 1997.
7. D. Q. Huynh and A. Heyden. Outlier Detection in Video Sequences under Affine Projection. In *Proc. of CVPR*, 2001.
8. D. Jacobs. Linear fitting with missing data: Applications to structure from motion and to characterizing intensity images. In *CVPR*, pages 206–212, 1997.
9. S. Mahamud and M. Hebert. Iterative projective reconstruction from multiple views. In *CVPR*, 2000.
10. S. Avidan and A. Shashua. Threading Fundamental Matrices. In *IEEE Trans. on PAMI*, Vol. 23(1), pp. 73–77, 2001.
11. P. Sturm and B. Triggs. A factorization based algorithm for multi-image projective structure and motion. In *ECCV96(II)*, pages 709–720, 1996.
12. C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: A factorization method. In *IJCV(9)*, No. 2, pages 137–154, November 1992.
13. M. Urban, T. Pajdla, and V. Hlaváč. Projective reconstruction from N views having one view in common. In *Vision Algorithms: Theory & Practice*. Springer LNCS 1883, pages 116–131, September 1999.