

OUTLIER DETECTION FOR FACTORIZATION-BASED RECONSTRUCTION FROM 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

KEY WORDS: Vision, Detection, Reconstruction, Structure, Reliability.

ABSTRACT

This paper proposes a method for outlier detection in 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, occlusions, and outliers in image correspondences jointly. The main novelty of this paper is the method for outlier detection whereas the proper reconstruction was described in (Martinec and Pajdla, 2002). In this work we assume that the amount of inliers is significantly larger than the amount of outliers. The main idea is that minimal configurations of points in triples of images are sufficient to validate inliers reliably. The RANSAC paradigm is used. Trifocal tensors are computed from randomly selected minimal n-tuples of points in triples of images. After the tensor estimation, the number of points consistent with the tensor is counted. If there are sufficiently enough consistent points, those not used to estimate the trif. tensor receive one positive vote. The voting is repeated until points in the measurement matrix are sufficiently sampled. The points that obtain zero or a very small number of votes are rejected as outliers. Inliers are used by the method described in (Martinec and Pajdla, 2002) to obtain a projective reconstruction. The set of inliers can be further enlarged by an iterative process. The new method is demonstrated here by experiments with laboratory and outdoor image sets.

1 INTRODUCTION

Tomasi & Kanade (Tomasi and Kanade, 1992) developed a factorization method of the measurement matrix for scene reconstruction with an orthographic camera. This method as well as Jacobs' method (Jacobs, 1997) can handle occlusions. Sturm and Triggs (Sturm and Triggs, 1996) extended this method from affine to perspective projections but without occlusions. Martinec & Pajdla (Martinec and Pajdla, 2002) solved reconstruction with both perspective projections and occlusions. Heyden (Huynh and Heyden, 2001) presented a reconstruction method from affine images with outliers but occlusions are not handled. Recently he extended the method into the perspective case (Heyden, 2002). We present a novel method for outlier detection so that reconstruction from perspective images is solved when occlusions and outliers are present jointly. Our method is independent of image ordering and treats all data uniformly. No six-tuple of points seen in all images is needed.

After problem formulation, philosophy of the new algorithm comes in Section 3, detailed explanation in Sec. 4 and 5. Experiments and summary are in Sec. 6 and 7.

2 PROBLEM FORMULATION

Suppose a set of n 3D points and that some of them are visible in m perspective images. There may be outliers,

*This research was supported by the grants GACR 102/01/0971, MSMT KONTAKT 2001/09, and CTU 0209313. Andrew Zisserman from the University of Oxford kindly provided the Dinosaur data, Marc Pollefeys from K.U.Leuven the Temple and the Castle data, Ondřej Chum from the Czech Technical University in Prague provided the routine for the trifocal tensor estimation, and Tomáš Werner from the University of Oxford provided the routine for the bundle adjustment.

i.e. mismatches, in image measurements. The goal is to reject outliers and to recover 3D structure (point locations) and motion (camera locations) from the remaining image measurements. 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 3×4 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 λ_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*, shortly *MM*. *MM* may have (and in most cases does have) some missing elements and outliers.

3 THE MAIN IDEA OF THE NEW OUTLIER DETECTION ALGORITHM

In the classical RANSAC paradigm, a “good” basis determining the structure of as much data as possible is searched for. Because MM with missing data may not contain any complete column at all, the standard concept of a basis determining the structure of the whole remaining data cannot be used. Hierarchical method (Fitzgibbon and Zisserman, 1998) builds a reconstruction from image triplets using trifocal tensors while image points inconsistent with the tensors are rejected as outliers. Triplets are joined into sub-sequences which can be further hierarchically registered into longer sub-sequences. Compared to this, we suggest a method which does not build on hierarchical approach. All camera matrices are estimated in one step from some image points consistent with some trifocal tensors. Further iterative process ensures a large set of inliers (image measurements consistent with the cameras) to be found.

In this work we assume that the amount of inliers is significantly larger than the amount of outliers. The main idea is that minimal configurations of points in triples of images are sufficient to validate inliers reliably. However, for large scenes, it is computationally infeasible to search for trifocal tensors among many triples of images which would validate all inliers. Therefore, another validation technique was proposed. When sufficiently many inliers are validated using trifocal tensors, it is possible to estimate reconstruction using (Martinec and Pajdla, 2002) and check which image measurements are consistent with the reconstruction. It turned out that a combination of the two above techniques validates inliers reliably and is computationally feasible. Moreover, when the second technique is iterated, a better reconstruction is found and the set of inliers increases.

The advantages conferred by proceeding with exploiting latently all known data at once are the following. There is no dependency on a good estimate from the early frames of the sequence, as opposed to a sequential approach. There is no difference between sequence and wide baseline stereo.

In sequences, a mismatch may cause that a track, i.e. image measurements in a correspondence, consists in fact from two (or more) different sub-tracks: one till the mismatch and the second one from the mismatch on. Consecutively, each of the sub-track will be validated but the whole track is wrong and is to cause large errors in subsequent reconstruction algorithm. The solution inheres in validating sub-tracks of length at least three (which can be done using trifocal tensors) and joining the overlapping sub-tracks (overlap at one image is sufficient), which is very simple compared to computing homographies in (Fitzgibbon and Zisserman, 1998). Since validating an outlier by the trifocal tensor is very unlikely, it is unlikely as well that non-continuous sub-tracks will join in such process.

Figure 1 shows the scheme of the whole algorithm. Measured data in MM are given votes from trifocal tensors

supporting them. Image points with high number of votes are labeled as tentative inliers and joined into sub-tracks, points with low number of votes are labeled as tentative outliers. The reconstruction is computed from tentative inliers by (Martinec and Pajdla, 2002) with tentative outliers regarded as the missing data. The reconstruction $\{P^i\}_{i=1}^m$, $\{X_p\}_{p=1}^n$, as a whole, is good if there is enough correctly reprojected world points X_p by all cameras P^i into the images. If the reconstruction is bad, voting is continued or repeated until a good reconstruction is found.

Tentative inliers may be consistent or inconsistent with the reconstruction. A tentative inlier x_p^i is consistent if its world point X_p projects into all tentative inliers of the j th track precisely enough. Otherwise, since at least one of the tentative inliers in the j th track is inconsistent and it is not known which of them is the outlier, all tentative inliers in the track are tentatively inconsistent and are marked as tentative outliers.¹ Similarly, it is desirable to find which tentative outliers are consistent and which are inconsistent with the reconstruction. Some of them may be the real outliers, others did not get enough votes because they have not been sampled.

Image points of the two cases can be validated using the known camera matrices P . If a track is consistent with P in a triple² of images, it is consistent with the reconstruction in the triple. Overlapping consistent triples of images can be joined into a sub-track. The sub-track, as a whole, may be inconsistent with the reconstruction due to noise in the data (only some of its triples were verified to be consistent). The consistent part of the sub-track can be found by reconstructing the j th world point from the sub-track and reprojecting it into the images. The image measurements consistent with the reconstruction are used as the tentative inliers in the next iteration of reconstructing the whole scene and validating.³ After convergence, tentative inliers are denoted as inliers and tentative outliers as outliers. Algorithm for finding the initial set of tentative inliers is summarized in Algorithm 1 while the whole outlier detection in Algorithm 2. The following two sections explain some steps in more detail.

4 VOTING BY TRIFOCAL TENSORS

For validating the inliers by trifocal tensors, \mathcal{T} , it is crucial to validate only those points which were not used to compute \mathcal{T} . The reason is that the probability that a contaminated \mathcal{T} validates another outlier is very small (with assumption of independent outliers). On the other hand, a \mathcal{T} computed from a contaminated 6-tuple often validates all the six points in the 6-tuple.

¹This is done only to speed up performance of the algorithm. Alternatively, the whole MM could be marked as tentative outliers and validated as described later.

²Pair could be also used but the test is less robust in the presence of noise.

³The whole sub-track may be passed as the tentative inliers into the next iteration but it increases the risk of incorporating an outlier into the set of tentative inliers. On the other hand, (i) it may lead to finding a better local minimum (it avoids stacking in some set of inliers which causes that also very “different” inliers may be validated) and (ii) it speeds up the convergence of the algorithm.

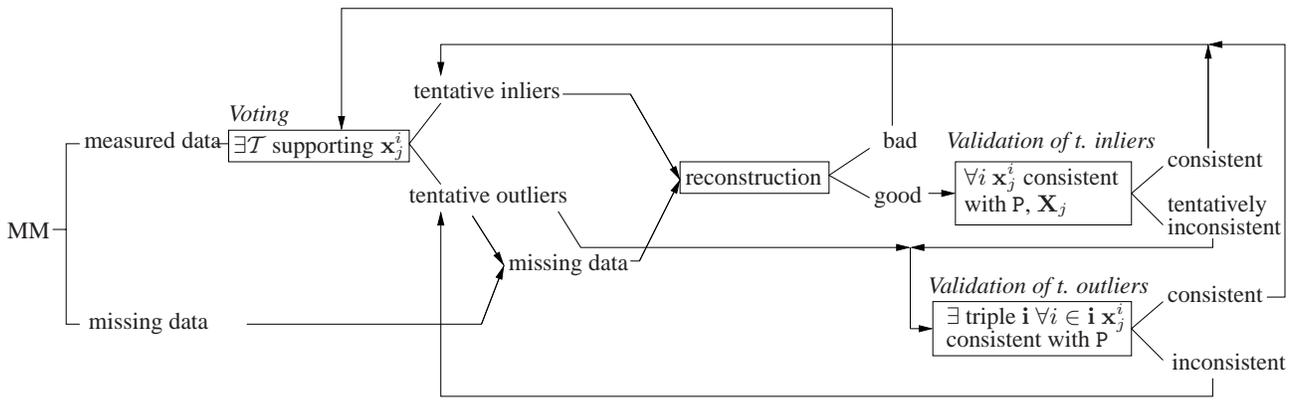


Figure 1: Scheme of outlier detection algorithm

4.1 3D Point Estimation

3D points can be estimated from camera matrices and MM. To make expressions shorter, index sets will be used (in the manner as in Matlab language). The index set in superscript and subscript, resp., will denote the choice of rows and columns, resp. Let index set \mathbf{i} denote a set of images. The following method was used to reconstruct scene points using camera matrices.

1. Estimate depth λ_p^i using camera matrices \mathbf{P}^i , $i \in \mathbf{i}$, by solving a system of linear equations.
2. Find \mathbf{X}_p as the coefficients of the linear combination of columns of \mathbf{P}^i in approximation of \mathbf{R}_p^i by \mathbf{P}^i where

$$\mathbf{P}^i = \begin{bmatrix} \mathbf{P}^{i_1} \\ \vdots \\ \mathbf{P}^{i_l} \end{bmatrix}, \mathbf{R}_p^i = \begin{pmatrix} \lambda_p^{i_1} \mathbf{x}_p^{i_1} \\ \vdots \\ \lambda_p^{i_l} \mathbf{x}_p^{i_l} \end{pmatrix}, \mathbf{i} = \{i_1, \dots, i_l\}$$

3. (optional) Tune point \mathbf{X}_p by a non-linear least-squares bundle adjustment.

4.2 Track Update

It is desirable to check whether some tentative outlier became consistent with \mathbf{P} since \mathbf{P} was changed from the last iteration. If a column p contains some tentative outlier, do the following. For all pairs/triples of images, \mathbf{i} , of the p th correspondence⁴ do the following. Estimate the world point \mathbf{X}_p using \mathbf{P}^i . If repr. errors of the three image points are below a given threshold, add a new sub-track if the triple does not overlap with some formerly validated sub-track or join the overlapping sub-track(s).

5 VALIDATION BY CONSISTENCY WITH RECONSTRUCTION

Image points consistent with \mathbf{P} can be found in the following way. 3D point \mathbf{X}_p is found using inliers in $[\mathbf{x}_p^1 \dots \mathbf{x}_p^m]^\top$ as in Section 4.1. Image points consistent with \mathbf{P} are those whose reprojection error is below a given threshold, i.e. $e_p^i = d(\mathbf{x}_p^i, \mathbf{P}^i \mathbf{X}_p) < t$ (where $d(\mathbf{x}, \mathbf{y})$ is the Euclidean distance between the points \mathbf{x} and \mathbf{y}).

⁴At maximum $\binom{m}{2}$ combinations. Alternatively, choose a pair/triple of images, \mathbf{i} , in random.

Let μ denote the size of the minimal consistent set, $\mu > 6$.

1. Choose randomly a triple of images so that there are at least μ common points in these images. Let $\mathbf{i} = \{i, j, k\}$ denote the index set of the chosen images. Let \mathbf{p} denote the index set of the points visible in images \mathbf{i} .
2. In images \mathbf{i} , choose randomly 6 common points in a non-degenerate configuration and estimate \mathcal{T} and camera matrices \mathbf{P}^i , $i \in \mathbf{i}$ (Hartley and Zisserman, 2000). There will be one or three real solutions.
3. **Finding the consistent set with \mathcal{T} .**
 - (a) Estimate 3D points \mathbf{X}_p , $p \in \mathbf{p}$, using \mathbf{P}^i as in Section 4.1.
 - (b) Calculate the reprojection errors as $e_p = \max_{i \in \mathbf{i}} d(\mathbf{x}_p^i, \mathbf{P}^i \mathbf{X}_p)$
 - (c) Compute the number of inliers consistent with \mathbf{P}^i (\mathcal{T}) by the number of correspondences for which $e_p < t$.
 - (d) If there are three real solutions for \mathcal{T} the number of inliers is computed for each solution, and the solution with most inliers retained.
4. **Voting.** If size of the consistent set is at least μ , then its image points except those used to estimate \mathcal{T} are (i) given a vote and (ii) used for updating the tracks (see Section 4.2).

Repeat steps 1–4 until image points are sufficiently sampled. Image points with high number of votes are tentative inliers, other points are tentative outliers.

Algorithm 1: Algorithm for finding the initial set of tentative inliers in MM using trifocal tensors by voting

- Initial set of inliers.** Find the initial set of tentative inliers in MM using validation by \mathcal{T} s (Alg. 1) that is sufficiently big to find all camera matrices using method (Martinec and Pajdla, 2002). Let T denote the validated sub-tracks of MM given as output.
- Reconstruction.** Set $T' = T$. Create M' from MM by splitting sub-tracks T' (see Section 5.1). Find P , X from tentative inliers in M' using (Martinec and Pajdla, 2002).
- Validation of tentative inliers.** Make tentative inliers from the image points in columns p consistent with P , X_p in all elements.
- Validation of tentative outliers.** In other columns, update T using P (see Sec. 4.2). Make tentative inliers from the validated sub-tracks of T .
- Iteration.** If any new image point consistent with P appeared, go to Step 2.

Image points consistent with P are inliers.

Algorithm 2: Outlier Detection Algorithm

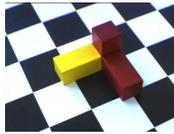
5.1 Splitting Tracks

If some track in MM consists of more sub-tracks, only the first sub-track is left and other sub-tracks are added as single columns to MM.

6 EXPERIMENTS

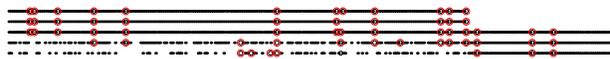
For each experiment, one image, an error table, and the structure of MM are provided. The table includes correspondence detection, accuracy for outlier detection, number of all, contaminated and partially validated tracks, the chosen strategy for depth estimation (see (Martinec and Pajdla, 2002)) and reprojection errors of the reconstruction without and with outliers. In structure of MM, "o" stands for outliers, "•" for scaled image points, "o" for unscaled, and " " for missing.

In Cubes scene, 10 % of image points were randomly chosen and shifted by 40 pixels in random directions simulating outliers. All of them have been correctly detected. Only two points in the third correspondence were outliers, however the remaining two points could not have been validated. In the Temple scene, all bad tracks have been correctly detected and split when possible. In Castle and Dinosaur scene, accuracy for outlier detection was set to one pixel which is quite strong restriction, however many tracks were still used at least partially. In the last two scenes, it has not been verified whether the detected outliers are the real ones due to huge amount of data.

<i>Cubes</i>	22 images [768×576]	
Corresp. / outl. det.	manual / 5	
Depth estimation	central image No. 1	
All / cont. / p. val. tracks	26 / 10 / 9	
Mean error / outl. [pxl]	0.22 / 6.93	



<i>Temple (Leuven)</i>	5 images [867×591]	
Corresp. / outl. det.	Harris' operator / 2	
Depth estimation	sequence	
All / cont. / p. val. tracks	284 / 20 / 6	
Mean error / outl. [pxl]	0.27 / 3.64	



<i>Castle (Leuven)</i>	22 images [768×576]	
Corresp. / outl. det.	Harris' operator / 1	
Depth estimation	sequence	
All / cont. / p. val. tracks	1822 / 716 / 338	
Mean error / outl. [pxl]	0.22 / 11.97	



<i>Dinosaur (Oxford)</i>	36 images [720×576]	
Corresp. / outl. det.	Harris' operator / 1	
Depth estimation	sequence	
All / cont. / p. val. tracks	2683 / 1326 / 587	
Mean error / outl. [pxl]	0.39 / 0.64	



7 SUMMARY AND FUTURE WORK

A new method for outlier detection was developed. Tests on laboratory and outdoor scenes showed its applicability. In the initial inlier detection step, method (Schaffalitzky et al., 2000) could be used. Sequential factorization of matrix R could help to improve convergence.

REFERENCES

- Fitzgibbon, A. W. and Zisserman, A., 1998. Automatic camera recovery for closed or open image sequences. *ECCV(I)*, pp. 311–326.
- Hartley, R. and Zisserman, A., 2000. *Multiple View Geometry in Computer Vision*. Cambridge University Press.
- Heyden, A., 2002. Personal communication.
- Huynh, D. Q., Heyden, A., 2001. Outlier detection in video sequences under affine projection. *CVPR*, pp. 695–701.
- Jacobs, D., 1997. Linear fitting with missing data: Applications to structure from motion and to characterizing intensity images. In: *CVPR*, pp. 206–212.
- Martinec, D., Pajdla, T., 2002. Structure from many perspective images with occlusions. *ECCV(II)*, pp. 355–369.
- Schaffalitzky, F., Zisserman, A., Hartley, R. I. and Torr, P. H. S., 2000. A six point solution for structure and motion. In: *ECCV(I)*, pp. 632–648.
- Sturm, P. and Triggs, B., 1996. A factorization based algorithm for multi-image projective structure and motion. In: *ECCV96(II)*, pp. 709–720.
- Tomasi, C. and Kanade, T., 1992. Shape and motion from image streams under orthography: A factorization method. In: *IJCV(9)*, No. 2, pp. 137–154.