Linear Multi-View Reconstruction for Translating Cameras

Carsten Rother*
Computational Vision and Active Perception Laboratory (CVAP)
Dept. of Numerical Analysis and Computer Science
KTH, SE-10044 Stockholm, Sweden
Email: carstenr@nada.kth.se

Abstract

This paper presents a linear multi view reconstruction algorithm for translating cameras with fixed internal parameters. The main advantages of this method are a) points and camera centers are computed simultaneously from one linear system containing all image data b) the allowance of arbitrary missing data. We show that the key to linearize the SFM problem is the infinite homography which comprises of the cameras' calibration and rotation. This insight unifies reconstruction methods for calibrated cameras, e.g. Oliensis [9], and uncalibrated cameras, e.g. Rother-Carlsson [10]. A further contribution of this paper is the summary and comparison of different approaches to determine the infinite homography.

1 Introduction

Structure from Motion (SFM) is a long studied and fundamental problem in Computer Vision as it can be seen from the number of publications [12, 14, 11, 3, 6, 9, 8, 13, 5, 10] and books [4, 2] devoted to the topic. Ideally an SFM-algorithm should exploit all available image data in one step in order to reconstruct points and cameras simultaneously. For the case of affine cameras a factorization algorithm has been presented [12] which handles all image data uniformly. However, all points have to be visible in all views, i.e. no missing data. The projective counterpart [11] additionally requires known epipolar geometry in order to determine projective depths.

In [10] a linear algorithm for multi view reconstruction has been presented which recovers points and cameras from the null-space of an image data matrix. It relies on a reference plane visible in all views. However, in contrast to factorization ([12, 11]) missing data can be handled. Compared to other reference plane methods, e.g. [13, 5], these advantages makes this approach potentially more powerful. The key idea in [10] is to determine the infinite homography from the reference plane. This paper we will show that there are more multi-view configurations with different constraints on the cameras and the scene for which the infinite homography can be determined. This includes the novel case of translating cameras with fixed internal parameters.

Applying the infinite homography to linearize the SFM-problem has also been exploited by Oliensis, e.g. [9, 8]. It is known (see [4]) that corresponding image points of purely rotating cameras define the infinite homography: $H = K' \; R \; K^{-1}$, where K, K' is the calibration matrix of the first and second camera and R the rotation between them. The basic assumption in Oliensis work is a small movement of the camera. This means that H can be approximately determined and used as initialization for the reconstruction algorithm ([9]). Furthermore, if the calibration is known, i.e. K and K', the rotation can be determined ([8]).

2 Structure, Motion and Infinite Homography

General perspective projection of a 3D point P_i onto the 2D image point p_{ij} can be described in homogeneous coordinates as:

$$p_{ij} \;\sim\; H_j \; (\; I \mid -\bar{Q}_j) \; P_i \;\sim\; H_j \; (\bar{P}_i \;-\; \bar{Q}_j) \qquad (1)$$

where $H_j \; (\; I \mid -\bar{Q}_j)$ represents the 3×4 projection matrix of camera j. Non-homogeneous coordinates are denoted with a bar, e.g. $\bar{Q}_j$, and homogeneous coordinates without a bar, e.g. p_{ij}. The symbol "$\sim$" means equality up to scale. Let us consider the homography H_j in more detail. A point $P = (X, Y, Z, 0)^T$, which lies on the plane at infinity π_∞, is mapped by eqn. (1) on the image plane π_j as:

$$p_{ij} \;\sim\; H_j \; (X, Y, Z)^T. \qquad (2)$$

Therefore, H_j can be considered as the *infinite homography*[1] between the plane at infinity π_∞ and the image plane π_j. Form eqn. (1) we see that if H_j is known, we are left with a linear and symmetric relationship between non-homogeneous points and camera centers:

$$p^*_{ij} \;\sim\; H_j^{-1} p_{ij} \;\sim\; \bar{P}_i \;-\; \bar{Q}_j. \qquad (3)$$

This suggests the following approach for structure and motion recovery:

1. *Determine the infinite homographies H_j*
2. *Reconstruct points and camera centers.*

Section 3 will discuss different ways to determine H_j with different constraints on the cameras or the scene.

*This work was supported by the Swedish Foundation for Strategic Research in the VISIT program.

[1] Note, the infinite homography is slightly differently defined in [4, 2].

If H_j is known, each scene point $\bar{P}_i$ visible in view j provides three linear projection relation which can be obtained from eqn. (3) by eliminating the unknown scale (see [10]). All linear projection can be put into a set of linear equations (SLE) which has for n points and m views the form:

$$\begin{aligned} L\,h &= 0 \text{ with} \qquad (4)\\ h &= (\bar{X}_1, \bar{Y}_1, \bar{Z}_1, \ldots, \bar{X}_n, \bar{Y}_n, \bar{Z}_n, \bar{A}_1, \bar{B}_1, \bar{C}_1, \ldots, \\ &\quad \bar{A}_n, \bar{B}_m, \bar{C}_m,)^T, \end{aligned}$$

with $\bar{P}_i = (\bar{X}_i, \bar{Y}_i, \bar{Z}_i)^T$ and $\bar{Q}_j = (\bar{A}_j, \bar{B}_j, \bar{C}_j)^T$. The nullspace of L, which can be obtained by Singular Value Decomposition, provides the solution for all points and camera centers. Since points on the plane at infinity π_∞ increase the dimensionality of the null-space of L (see [10]), these points have to be excluded from the SLE and reconstructed separately with eqn. (2). Let us summarize the main advantages of this reconstruction method:

- *One linear system containing all image data*
- *Missing data can be handled*
- *Points and cameras are determined simultaneously.*

3 Determine the Infinite Homographies

The key to linearize the problem of structure form motion is the infinite homography. It can be derived with knowledge about the scene, the cameras or the cameras' motion. Various cases are summarized and compared in this section, including the case of purely translating cameras.

3.1 Constant or Known Rotation and Calibration

Eqn. 1 can be written as well as:

$$p_{ij} \sim K_j\, R_j\, (\, I \mid -\bar{Q}_j)\, T\, T^{-1}\, P_i. \qquad (5)$$

We see that the infinite homography H_j depends on the calibration K_j and rotation R_j of camera j, i.e. $H_j = K_j\, R_j$. The 4×4 matrix T represents the free choice of a coordinate system, which has in case of a projective reconstruction 15 degrees of freedom. Let us choose a special T:

$$T_A = \begin{pmatrix} A & \mathbf{0} \\ \mathbf{0}^T & 1 \end{pmatrix} \text{ with } \mathbf{0} = (0,0,0)^T. \qquad (6)$$

This special T_A does not transform the plane at infinity $\pi_\infty = (0,0,0,1)^T$ since $T_A^{-T}\, \pi_\infty = \pi_\infty$. Therefore, depending on the choice of A the final reconstruction is either Euclidean or affine. For this special T_A eqn (6) may be written as:

$$p_{ij} \sim K_j\, R_j\, A\, (\, I \mid -\bar{Q}_j)\, T_A^{-1}\, P_i. \qquad (7)$$

For calibrated cameras with known rotation, i.e. K_j and R_j are known, we may choose $H_j = K_j\, R_j$ and $A = I$. Since T_A is the identity matrix the resulting reconstruction is Euclidean. In case of calibrated, translating cameras, i.e. K_j known and R_j constant, we may choose $H_j = K_j$ and $A = R^{-1}$. The reconstruction is Euclidean since T_A represents a similarity transformation in $\mathcal{P}^3$. Finally, for translating cameras with fixed internal camera parameters, i.e. K_j and R_j constant, we may choose $H_j = I$ and $A = (K\ R)^{-1}$. The reconstruction is affine since T_A represents an affine transformation in $\mathcal{P}^3$. If the camera calibration is approximately known, H_j should be set as $H_j = K$. The fact that purely translating cameras produce affine structure has already been shown by [14].

3.2 Unknown Rotation and Calibration

Methods for computing camera internal calibration K_j, also known as auto- (or self-) calibration techniques, may be divided into two classes: an initial projective reconstruction is *known* or *unknown*. Since multi view reconstruction is our task, only methods of the latter case can be utilized here. These methods exploit known metric properties of the scene, e.g. angles. In [7] a dominant metric property of man-made environments, e.g. architectural scenes, is used: *orthogonal directions*. It has been shown [1, 7] that a camera with known aspect ratio and skew can be calibrated from three vanishing points of orthogonal directions in the scene. If K_j is determined by an arbitrary auto-calibration technique, we are left with the unknown rotation R_j. R_j can be considered as the rotation between image 1 and j, if $A = R_1$ in eqn. 7. Let us assume that a vanishing point v is obtained in image 1 and j, i.e. v_1 and v_j. If K_j is known, each vanishing point represents a direction in the Euclidean camera coordinate frame[2]: $d_1 = \pm K_1^{-1} v_1$ and $d_j = \pm K_j^{-1} v_j$. The directions are related by R_j: $d_j/||d_j|| = R_j\, d_1/||d_1||$. Therefore, two such corresponding directions are sufficient to determine uniquely the three degrees of freedom of R_j.
In summary, the derivation of K_j and R_j from three orthogonal vanishing points as presented in [10] might be the most useful approach in practice.

3.3 Reference Plane

It has been shown in [10] that H_j can be derived from a reference plane visible in all views. It is assumed that four coplanar reference points are visible in all views. The key idea is to choose canonical coordinates for these four points in the scene: $P_1 = (1,0,0,0)^T, P_2 = (0,1,0,0)^T, P_3 = (0,0,1,0)^T, P_4 = (1,1,1,0)^T$ and in each image j: $p_{j1} = (1,0,0)^T, p_{j2} = (0,1,0)^T, p_{j3} = (0,0,1)^T, p_{j4} = (1,1,1)^T$. We see that the reference plane defines the plane at infinity in this particularly chosen projective space. As a consequence, scene points on this particular plane have to be detected and reconstructed separately. The infinite homography is defined as the identity matrix (see eqn. (2)),

[2]Note, the sign of a direction has to be known as well

	Constraints	Reconst.
1	K_j known or auto-calibrated; R_j known or constant	Euclidean
2	K_j known or auto-calibrated; R_j unknown, 2 van.points with direction	Euclidean
3	K_j, R_j constant	affine
4	Plane visible in all views (special case: 4 coplanar points)	projective

Table 1. Different cases to determine H_j

i.e. $H_j = I$. Alternatively, H_j can be derived from the inter-image homographies induced by a reference plane.

3.4 Comparison

The different cases previously discussed are summarized in table 1. The table additionally includes the type of the final reconstruction, which can obviously be upgraded to affine or Euclidean with the given (or further) constraints (see [4]). However, these different cases have two fundamental differences:

- For the first three cases the plane at infinity is at its true position. This has the advantage that the linear system contains *all* finite scene points, i.e. not excluding points on a certain reference plane as in case 4.
- The quality of H_j is a major factor for the quality of the final reconstruction (sec. 4). Those cases which compute H_j from a few reference points, i.e. cases 2 and 4, are potentially inferior to the other cases.

4 Experiments

In the experiments the case of translating cameras with fixed internal parameters is investigated. However, the conclusion drawn from the results apply to all cases (1-3 in table 1), where K_j and R_j are derived explicitly.

4.1 Synthetic Data

In order to demonstrate the performance of the algorithm, it was applied to a wide range of different camera and scene settings. Fig. 1 depicts two of them: lateral movement – LAT (a) and translational movement towards the scene – TOW (b). For the TOW-configuration, points on the baseline, i.e. the line of the camera centers, were removed. The internal calibration matrix was set to diag(1000,1000,1).
In a first experiment the influence of noise on the image data was investigated. Fig. 1 (c) shows the results of our algorithm (Our) and the projective factorization algorithm (Fac.) of Sturm-Triggs [11] for the two synthetic configurations. In order to obtain average performance, the following two steps were conducted 20 times for each noise level: a) randomly determine 50 points b) add Gaussian noise (standard deviation σ) on the reprojected 3D points. The computed reconstructions were evaluated in terms of the Root-Mean-Square (RMS) error between reprojected 3D points and 2D image data (potentially corrupted by noise). The main observation is, that the performance of our algorithm and factorization is close to identical. Furthermore, the performance of both algorithms is close to the theoretical minimum, i.e. Cramera-Rao lower bound. This means that this method is close to optimal with the assumption that the cameras' calibration and rotation is detected correctly.
The second experiment investigated the case if the assumption of pure translating cameras is not satisfied (see fig. 1 (d)). In contrast to the previous experiment Gaussian noise (in degree) was added on the rotation of each camera in an arbitrary direction. The result is as expected: the reconstruction is perfect without noise and is less accurate with increasing noise. This shows that the quality of H_j has a major influence on the final reconstruction. However, for small errors in the rotation, e.g. $\sigma = 1^o$, the resulting RMS error is of the same magnitude as the error for practical noise level on the image data, e.g. $\sigma = 2\,pixel$,

4.2 Real Data

Fig. 2 (a) depicts 3 out of 16 images of a real sequence, where a camera moved translational along a corridor. 46 scene points where selected manually in the sequence. Fig. 2 (b) shows the visibility matrix, where a dot indicates that a certain scene point is visible in a certain view. We see that a point is visible in average in 4 successive frames. The top view of the final reconstruction, which had an RMS error of 6.6, is depicted in fig. 2 (c). The first part of the reconstruction from $0-9m$ is qualitatively correct: the camera (crosses) moved on a fairly straight line along the corridor (dots), which is about $15m$ long and $2m$ wide. However, the second part from $9-15m$ is qualitatively worse: the width of the corridor and the distances between successive camera positions is shrinking. An explanation for this might be that the images of scene points at the end of the corridor are essentially closer to the focus of expansion which negatively influences both the reconstruction of points and cameras.

5 Summary and Conclusions

We have presented a linear multi view reconstruction algorithm for translating cameras with fixed internal parameters. Points and camera centers are computed simultaneously as the nullspace of one linear system constructed from all the image data. In contrast to factorization-based algorithms, e.g. [11, 12], we allow arbitrary missing data,
It has been shown that the key to linearize the SFM problem is the *infinite homography* which comprises of the cameras' calibration and rotation. Several, alternative ways to computing the infinite homography have been presented including the case of a reference plane visible in all views (see [10]). However, a major disadvantage of the reference plane case is that the reference plane is chosen as the plane at infinity. As a consequence, scene points on this particular

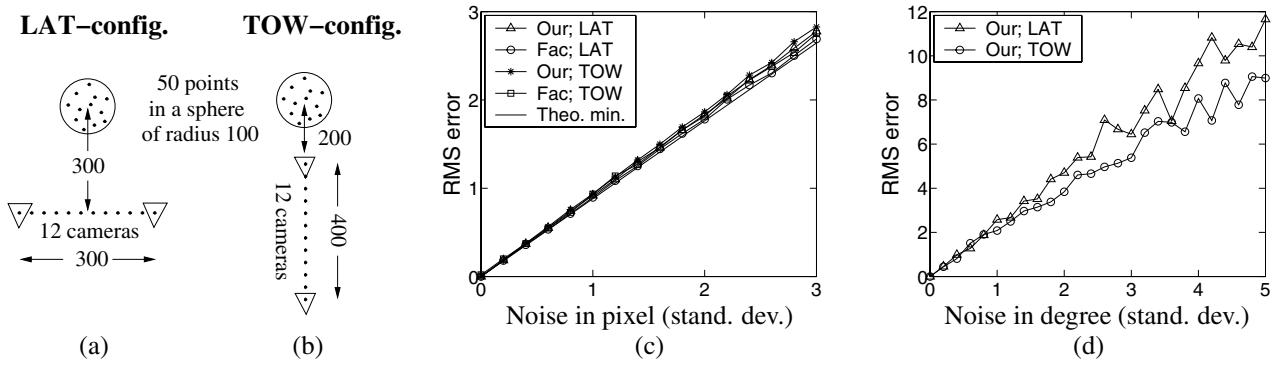

Figure 1. Top view of two synthetic configurations: lateral movement – LAT (a) and translational movement towards the scene – TOW (b). The performance of our algorithm (Our) and factorization (Fac.) in terms of noise on the image data (c) and noise on the camera rotation (d).

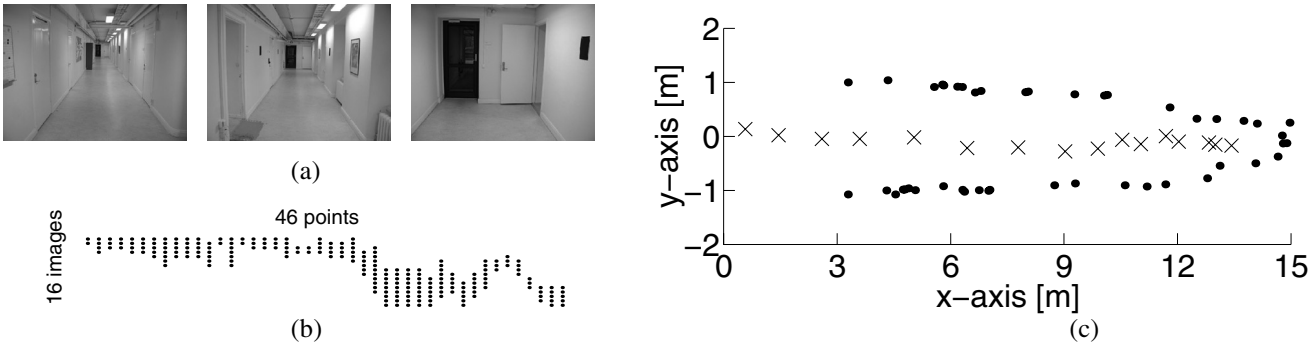

Figure 2. Three images of the corridor sequence (a), the visibility matrix (b) and a top view of the reconstruction (c), where dots represent 3D points and crosses the camera positions.

plane have to be detected and reconstructed separately. This is not necessary in our case, where the infinite homography is derived from the restriction on the cameras' motion, i.e. purely translating cameras.

We demonstrated experimentally that the performance of our method is as good as projective factorization [11] and close to the theoretical minimum. Furthermore, we have seen that the quality of the determined infinite homographies directly influence the quality of the reconstruction.

References

[1] Caprile, B. and Torre, V. 1990 Using vanishing points for camera calibration. In Int. J. Computer Vision, 4:127-140.

[2] Faugeras, O. and Q.-T. Luong 2001. The Geometry of Multiple Images. The MIT Press.

[3] Fitzgibbon, A. W. and Zisserman, A. 1998. Automatic camera recovery for closed or open image sequences. In Europ. Conf. Comp. Vision, Freiburg, Germany, p. 311-326.

[4] Hartley, R. and Zisserman, A. 2000. Multiple View Geometry in Computer Vision. Cambridge University Press.

[5] Hartley, R., Dano, N. and Kaucic, R. 2001. Plane-based Projective Reconstruction. In Int. Conf. Comp. Vision, Vancouver, Canada, p. 420-427.

[6] Koch, R., Pollefeys, M. and VanGool, L. 1998. Multi viewpoint stereo from uncalibrated video sequences. In Europ. Conf. Comp. Vision, Freiburg, Germany, p. 55-65.

[7] Liebowitz, D. and Zisserman, A. 1999. Combining Scene and Auto-Calibration Constraints. In Int. Conf. Comp. Vision, Kerkyra, Greece, p. 293-300.

[8] Oliensis J. and Genc, Y. 1999. Fast algorithms for projective multi-frame structure from motion. In Int. Conf. Comp. Vision, Kerkyra, Greece, p. 536-542.

[9] Oliensis, J. 1999. A Multi-Frame Structure-from-Motion Algorithm under Perspective Projection. In Int. J. Computer Vision, 34(2/3):163-192.

[10] Rother, C. and Carlsson S. 2001. Linear Multi View Reconstruction and Camera Recovery. In Int. Conf. Comp. Vision, Vancouver, Canada, p. 42-51.

[11] Sturm, P. and Triggs, B. 1996. A factorization based algorithm for multi-image projective structure and motion. In Europ. Conf. Comp. Vision, Cambridge, U.K., p. 709-719.

[12] Tomasi, C. and Kanade, T. 1992. Shape and motion from image streams under orthography: a factorization method. In Int. J. Comp. Vision, 9(2):137-54.

[13] Triggs, B. 2000. Plane + Parallax, Tensors and Factorization. In Europ. Conf. Comp. Vision, Dublin, Ireland, p. 522-538.

[14] VanGool, L. and Moons, T. and Proesmans, M. and VanDiest, M. 1994. Affine Reconstruction from Perspective Image Pairs Obtained by a Translating Camera. In Int. Conf. Pattern Recognition, Jerusalem, Israel, p. 290-294.