Slimane LARABI

Computer Vision

From Bidimensional Images to Three Dimensional Scene

January 02, 2025

Springer Nature

Chapter 1

Structure from Motion

1.1 Introduction

Reconstructing 3D structures from 2D images is a fundamental problem in computer vision, known as Structure from motion (SfM) is useful for many applications. We can enumerate Object Recognition, Robotics, Computer Graphics, Image Retrieval, Geo-Localization, Archaeology and Sports.

This process aims to recover the three-dimensional shape of a scene and the motion of the camera from a sequence of 2D projections. The challenge of SfM lies in its inherent ambiguity and the need to infer depth and spatial relationships from limited visual information. This chapter delves into the problem of Structure from Motion, with a specific focus on orthographic projection, a simplified yet powerful model for SfM.

We begin by introducing the problem of Structure from Motion, providing an overview of its significance, challenges, and applications. Unlike perspective SfM, orthographic SfM assumes a simplified imaging model that eliminates perspective effects, making it particularly useful in scenarios where the field of view is narrow or when computational efficiency is critical.

Next, we explore the mathematical expression of orthographic projection, detailing how 3D points in the scene are projected onto a 2D image plane under this model. By formalizing this relationship, we establish the foundation for algorithms that can estimate both the 3D structure of the scene and the motion of the camera from a series of orthographic projections.

Finally, we investigate Orthographic Structure from Motion (SfM) methods, discussing how this approach enables the recovery of 3D structure and motion parameters from 2D observations. By leveraging the mathematical properties of orthographic projection, these methods achieve robust and efficient reconstruction, particularly in controlled environments.

1.2 Problem of Structure from Motion

We suppose that we have a sequence video frames and features points such as corners, SIFT points are detected on each frame (see figure 1.1). Also, we suppose that these features are tracked using known techniques: Template matching, Optical Flow (we know the correspondence between features of acquired images).

In order to reconstruct the 3D scene from a set of acquired images, we will assume, for simplification of the model and time consuming reduction, that images are acquired following the orthographic model of projection.

1.2.1 Structure from Motion Assuming Orthographic Projection

Firstly, we define the orthographic projection as representing three-dimensional objects in two dimensions. The Orthographic projection is a form of parallel projection in which all the projection lines are orthogonal to the projection plane, resulting in every plane of the scene appearing in affine transformation on the viewing surface (see figure 1.1).

Let $(u_{f,p}, v_{f,p}$ a set of corresponding image points (2D) , where f is the frame number and p refers to the image point on that frame.

Advantages of orthographic camera for structure from motion:

- Simplifies the mathematical equations involved in the reconstruction process.

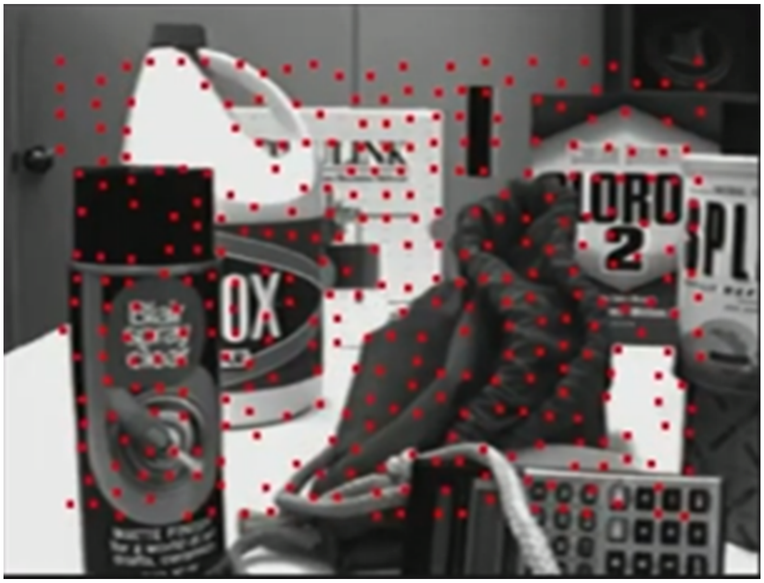

Fig. 1.1 Key points located on one image (from video sequence images).

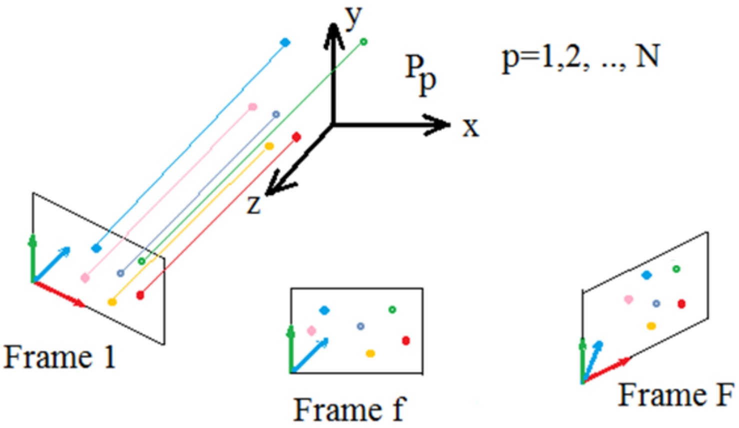

Fig. 1.2 Orthographic projection of 3D points into a video sequence images.

- The ray of projection of a $3D$ point is perpendicular onto the image plane, this simplifies the triangulation process.

- Objects maintain their relative sizes regardless of their distance from the camera.

- It involves simpler mathematical operations, this implies faster computation times.

- It preserves the parallelism of lines and planar structures in the scene.

The limitations:

- It leads to inaccuracies in depth estimation.

- It is difficult to determine the absolute scale of the reconstructed scene without additional constraints or information.

- It leads to distortions and inaccuracies in the reconstructed 3D for scenes with irregular shapes or curved surfaces (non planar).

The main objective is to find the scene points (3D) assuming orthographic camera (see figure 1.2).

1.2.1.1 Mathematical Expression of Orthographic projection

From figure 1.3, we can write using the scalar product and the vector ($x_c = OP$) and the unit vectors i and j of the 2D coordinates frame:

$$u = i.x_c = i^T x_c, v = j.x_c = j^T x_c \tag{1.1}$$

Using the world coordinates, we rewrite in the previous equation x_c using the vectors C_w and x_w (see figure 1.3):

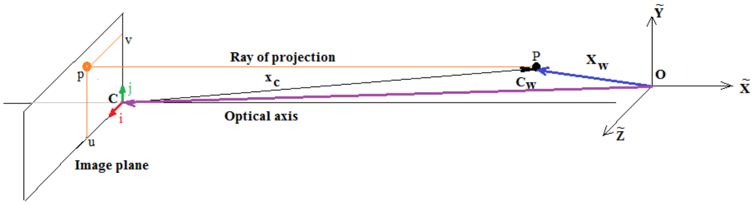

Fig. 1.3 Expression of orthographic projection using the world coordinate frame.

$$u = i.x_c = i^T x_c = i^T(x_w - C_w) = i^T(P - C), v = j.x_c = j^T x_c = j^T(x_w - C_w) = j^T(P - C) \tag{1.2}$$

1.2.1.2 Computation of the Structure

Let $(u_{f,p}, v_{f,p}$ a set of corresponding image points (2D) , where f is the frame number and p refers to the image point on that frame.

The main objective is to find the scene points $(3D)$ assuming orthographic camera.

Camera positions (C_f) and orientations (i_f, j_f) are unknown (see figure 1.4). For each image point P_k in camera frame f we have:

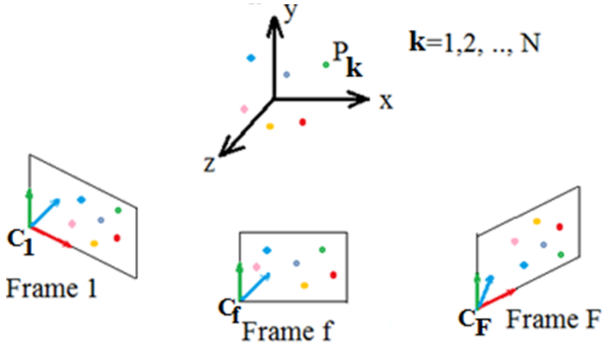

Fig. 1.4 Orthographic projection for SfM.

$$u_k = i_f^T(P_k - C_f), v_k = j_f^T(P_k - C_f) \tag{1.3}$$

We can remove C_f from equations to simplify SFM problem. To do this, we assume that the origin of the world at centroid P of scene points P_k (see figure 1.5).

$$P = \frac{1}{N} \sum_{k=1}^{k=N} P_{f,k} \tag{1.4}$$

The centroid $\overline{p}$ of image points in frame f is given by:

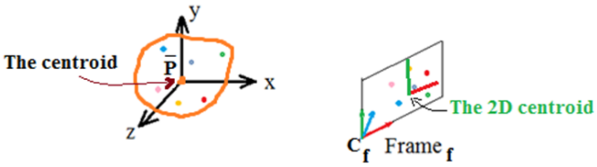

Fig. 1.5 Orthographic projection for SfM.

$$\overline{u}_f = \frac{1}{N}\sum_{k=1}^{k=N} u_{f,k}, \overline{v}_f = \frac{1}{N}\sum_{k=1}^{k=N} v_{f,k} \tag{1.5}$$

$$\overline{u}_f = \frac{1}{N}\sum_{k=1}^{k=N} i_f^T(P_k - C_f), \overline{v}_f = \frac{1}{N}\sum_{k=1}^{k=N} j_f^T(P_k - C_f) \tag{1.6}$$

$$\overline{u}_f = i_f^T\frac{1}{N}\sum_{k=1}^{k=N} P_k - \frac{1}{N}\sum_{k=1}^{k=N} i_f^T C_f, \overline{v}_f = j_f^T\frac{1}{N}\sum_{k=1}^{k=N} P_k - \frac{1}{N}\sum_{k=1}^{k=N} j_f^T C_f \tag{1.7}$$

$$\overline{u}_f = -\frac{1}{N}\sum_{k=1}^{k=N} i_f^T C_f, \overline{v}_f = -\frac{1}{N}\sum_{k=1}^{k=N} j_f^T C_f \tag{1.8}$$

The centroid $(\overline{u}_f, \overline{v}_f)$ of feature points is not a function of the locations of scene points. Then if we shift the origin to the centroid, image points coordinates w.r.t. $(\overline{u}_f, \overline{v}_f)$ will be:

$$\overline{u}_{f,k} = u_{f,k} - \overline{u}_f, \overline{v}_{f,k} = v_{f,k} - \overline{v}_f \tag{1.9}$$

$$\overline{u}_{f,k} = i_{f,k}^T(P_k - C_f) + i_{f,k}^T C_f = i_{f,k}^T P_k, \overline{v}_{f,k} = j_{f,k}^T(P_k - C_f) + j_{f,k}^T C_f = j_{f,k}^T P_k \tag{1.10}$$

We can see that we have the image coordinates without the camera center in the expression. Only the camera orientation (i_f^T, j_f^T) and scene point are present (C_f removed).

We define the observation matrix $W = MS$ where:

$$\begin{bmatrix} \overline{u}_{11} & \overline{u}_{12} & \overline{u}_{13} & ---- & \overline{u}_{1n} \\ \overline{u}_{21} & \overline{u}_{22} & \overline{u}_{23} & ---- & \overline{u}_{2n} \\ ---- & ---- & ---- & --- & --- \\ \overline{u}_{F1} & \overline{u}_{F2} & \overline{u}_{F3} & ---- & \overline{u}_{Fn} \\ \overline{v}_{11} & \overline{v}_{12} & \overline{v}_{13} & ---- & \overline{v}_{1n} \\ \overline{v}_{21} & \overline{v}_{22} & \overline{v}_{23} & ---- & \overline{v}_{2n} \\ ---- & ---- & ---- & --- & --- \\ \overline{v}_{F1} & \overline{v}_{F2} & \overline{v}_{F3} & ---- & \overline{v}_{Fn} \end{bmatrix} = \begin{bmatrix} i_1^T \\ i_2^T \\ --- \\ i_F^T \\ j_1^T \\ j_2^T \\ --- \\ j_F^T \end{bmatrix} \begin{bmatrix} P_1 \ P_2 \ P_3 -- P_n \end{bmatrix} \tag{1.11}$$

The vector structure S is composed by $(3 \times n)$ elements where 3 refers to the coordinates x, y, z. The vector motion M is composed by $(2F \times 3)$ elements, where 2F refers to the 2 unit vectors (i, j) for the F frames. The matrix W is composed by $(2F \times n)$ elements which represent the known n centroid-substracted feature points for the F frames.

1.2.1.3 Computation of *M* and *S* from *W*

Carlo Tomasi and Takeo Kanade, proposed in 1992 in their paper a new method: Shape and motion from image streams under orthography: a factorization method.

The rank of the matrix W is equal to 3.

Rank(W)=Rank(MS) ≤ Rank(M) ≤ min(2F, 3)

Rank(W)=Rank(MS) ≤ Rank(S) ≤ min(n, 3)

Rank(W)=Rank(MS) ≤ min(2F, 3, n)

Rank(W) ≤ 3

For any matrix A there exists a factorization

$$A_{M \times N} = U_{M \times M} \sum V^T_{N \times N} \tag{1.12}$$

where U and V are orthonormal and $\sum$ is orthogonal of dimensions $M \times N$, M here is equal to $2F$.

$$\sum = \begin{bmatrix} \sigma_1 & 0 & 0 & 0 & 0 & 0 & -- & 0 \\ 0 & \sigma_2 & 0 & 0 & 0 & 0 & -- & 0 \\ 0 & 0 & \sigma_3 & 0 & 0 & 0 & -- & 0 \\ 0 & 0 & 0 & \sigma_4 & 0 & 0 & -- & 0 \\ - & - & - & - & - & - & - & - \\ 0 & 0 & 0 & 0 & 0 & 0 & -- & \sigma_n \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{1.13}$$

$\sigma_1 \geq \sigma_2 \geq \sigma_3 .... \geq \sigma_n$ are singular values. As the rank of A is equal to 3, the $\sigma_i, i = 4..n$ are equal to zero. We can write then the previous equation as follow:

$$W = \begin{bmatrix} & & \\ U_1 & U_2 & \\ & & \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & 0 & 0 & 0 & 0 & -- & 0 \\ 0 & \sigma_2 & 0 & 0 & 0 & 0 & -- & 0 \\ 0 & 0 & \sigma_3 & 0 & 0 & 0 & -- & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -- & 0 \\ - & - & - & - & - & - & - & - \\ 0 & 0 & 0 & 0 & 0 & 0 & -- & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} V_1^T \\ \\ V_2^T \end{bmatrix} \tag{1.14}$$

U_1 is with 3 columns and U_2 with $2F - 3$ columns. Since $Rank(W) \leq 3$, $Rank(\sum) \leq 3$, then sub-matrices U_2 and V_2^T do not contribute to W.

We have then: $W = U_1 \sum_1 V_1^T$, where the dimensions of U_1 is $2F \times 3$, of $\sum_1$ is the 3×3 and of V_1^T is $3 \times N$.

For any matrix Q, the following expression is valid:

$$W = MS = (U_1 (\textstyle\sum_1)^{1/2} Q)(Q^{-1} (\textstyle\sum_1)^{1/2} V_1^T)$$

$$\begin{bmatrix} i_1^T \\ i_2^T \\ - \\ i_F^T \\ j_1^T \\ j_2^T \\ - \\ j_F^T \end{bmatrix} = U_1 (\sum_1)^{1/2} Q = \begin{bmatrix} \hat{i}_1^T Q \\ \hat{i}_2^T Q \\ - \\ \hat{i}_F^T Q \\ \hat{j}_1^T Q \\ \hat{j}_2^T Q \\ - \\ \hat{j}_F^T Q \end{bmatrix} \tag{1.15}$$

Knowing the following orthonormality constraints for the frame f:

$\hat{i}_f^T \hat{i}_F^T = 1, \hat{i}_f^T \hat{i}_F^T = 1, \hat{i}_f^T \hat{j}_F^T = 0.$

Then, for one frame we get three equations, where Q is unknown: $\hat{i}_f^T Q Q^T \hat{i}_F^T = 1$, $\hat{i}_f^T Q Q^T \hat{i}_F^T = 1, \hat{i}_f^T Q Q^T \hat{j}_F^T = 0.$

We get then $3F$ quadratic equations with Q is 3×3 matrix, 9 variables.

Q can be solved with 3 or more images ($F >= 3$) using Newton's method.

Figures 1.6, 1.7 show an example of structure from motion (Tomasi 1992, Duke University, USA).

Algorithm

Summary: Orthographic SFM

1- Detect and track feature points

2- Create the centroid subtracted matrix W of corresponding feature points

3- Compute SVD of W end enforce rank constraint. $W = U \sum V^T = U_1 \sum_1 V_1^T$ 4- Set $M = U_1 (\sum_1)^{1/2} Q$ and $S = Q^{-1} (\sum_1)^{1/2} V_1^T$

5- Find Q by enforcing the orthornormality constraint

Figures 1.8, 1.9, 1.10 show the images taken by Marc Pollefeys and Luc Van Gool [12] and the computed motion and the reconstructed scene.

Figures 1.11, 1.12 show the images of outdoor scene and the computed reconstructed scene [11].

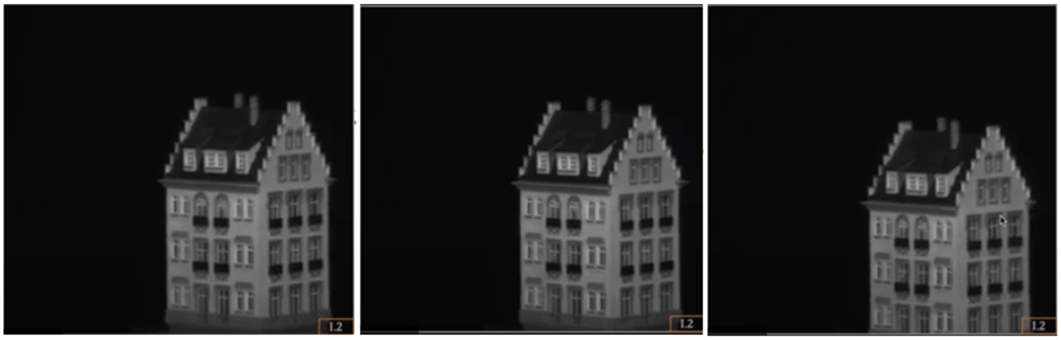

Fig. 1.6 Three images of a scene.

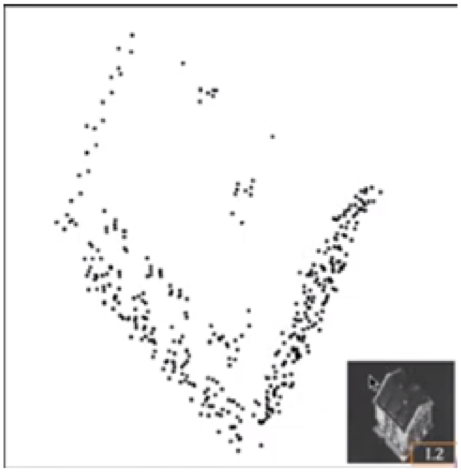

Fig. 1.7 The reconstructed scene.

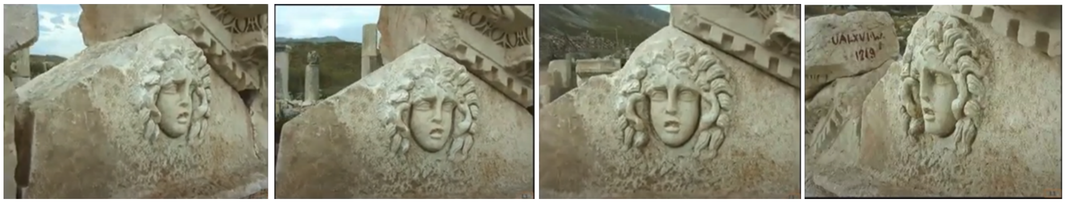

Fig. 1.8 Some images of a scene.

1.3 Other works related to Structure from Motion

In the state of the art, we note that there are two kind of methods: Sequential methods and Factorization methods [36].

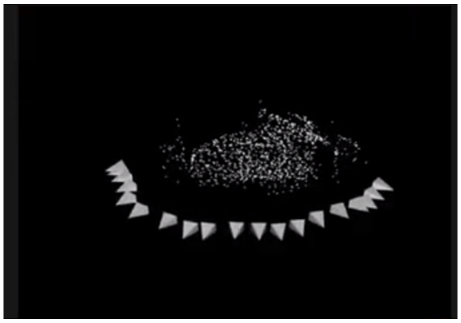

Fig. 1.9 The computed motion of the camera.

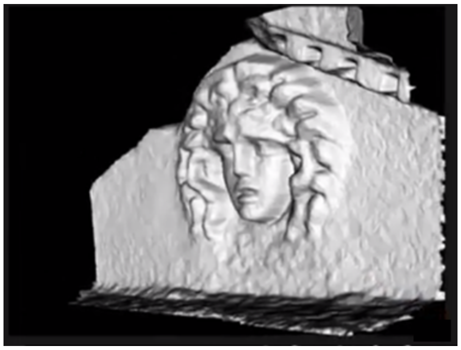

Fig. 1.10 The reconstructed scene.

Fig. 1.11 The used images [11].

1.3.1 Sequential algorithms

Sequential algorithms are the most popular. They work by incorporating successive views one at a time. As each view is registered, a partial reconstruction is extended by computing the positions of all 3D points that are visible in two or more views using triangulation. A suitable initialization is typically obtained by decomposing

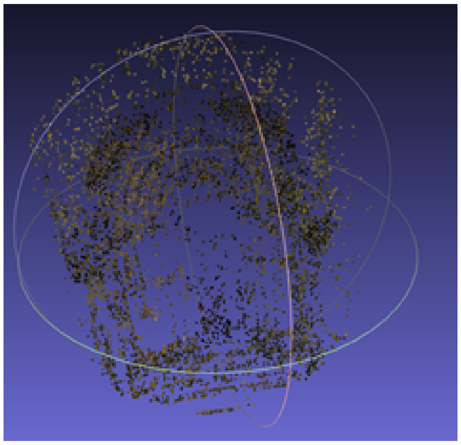

Fig. 1.12 The reconstructed scene.

the fundamental matrix relating the first two views of the sequence. There exist several strategies for registering successive views:
- Epipolar constraints. One possibility is to exploit the two-view epipolar geoemtry that relates each view to its predecessor. For example, where camera intrinsic parameters are known, essential matrices can be used. Essential matrices are estimated linearly using eight or more point correspondences and decomposed to give relative camera orientation and the direction of camera translation. The magnitude of the translation can be fixed using the image in the new view of a single known 3D point, i.e. a point that has already been reconstructed from its image in earlier views.

- Resection. An alternative is to determine the pose of each additional view using already reconstructed 3D points. Six or more 3D to 2D correspondences allow linear solution for the 12 elements of a projection matrix.

- Merging partial reconstructions. Another alternative is to merge partial reconstructions using corresponding 3D points. Typically, two or three view reconstructions are obtained using adjacent image pairs or triplets; then they are merged using corresponding 3D points.

These sequential registration schemes have some important limitations. In the context of interactive modelling systems, one disadvantage is that a large number of corresponding points must be defined in each view. For uncalibrated reconstruction, commercial photogrammetry software (such as ImageModeler5) usually requires a minimum of 7 correspondences per view (and more are recommended for better accuracy). Since corresponding points must usually be visible in three or more views, this means substantial overlap is required. For long sequences of views (e.g. along a city street), this requirement can be prohibitive. Another complication is that there exist various kinds of degenerate structure and motion configuration for which the standard algorithms will fail. For example: (i) camera rotation in the absence of translation, (ii) planar scenes, (iii) a 3D point lying on a line passing through the optical centres of the cameras in which it is visible. In practice, it may be hard to avoid these kinds of degeneracy, especially if views are obtained without careful (or even expert) planning.

1.3.2 Factorization methods

Unlike sequential methods, batch methods work by computing camera pose and scene geometry using all image measurements simultaneously. One advantage is that reconstruction errors can be distributed meaningfully across all measurements; thus, gross errors associated with sequence closure can be avoided.

One family of batch structure from motion algorithms are called factorization methods. Fast and robust linear methods based on direct SVD factorization of the image point measurements have been developed for a variety of simplified linear (affine) camera models, e.g. orthographic (Tomasi and Kanade [37]). These methods generally are not applicable to real-world scenes because real camera lenses are too wide-angle to be approximated as linear. Indeed, the inherent distortion that occurs in images captured by wide-angle lenses, which cannot be accurately modeled using simple linear transformations. Then it's necessary to use more complex, non-linear models (e.g., radial or tangential distortion models) to correct the distortion and accurately map the real-world scene to the image

More recently, a number of researchers have described factorization-like algorithms for perspective cameras too. These methods are iterative and there is no guarantee that they will converge to the optimal solution [38] [39]. that there exist degenerate structure and motion configurations for which they will fail, they are not applicable to sparse modelling problems.

1.4 Bundle adjustment

From image features uij , structure from motion gives an initial estimate of projection matrices Pi and 3D points Xj . Usually it will be necessary to refine this estimate using iterative non-linear optimisation to minimize an appropriate cost function. This is bundle adjustment [41][40]. Bundle adjustment works by minimizing a cost function that is related to a weighted sum of squared reprojection errors. Usually Gauss Newton iteration is used (with an appropriate step control policy) for rapid convergence. This section provides a brief review of established bundle adjustment theory.

13.9.1 Problem definition

The goal of bundle adjustment is to determine an optimal estimate of a set of parameters θ, given a set of noisy observations. Most bundle parameters cannot be observed directly, e.g. projection matrices, 3D point coordinates. Instead, they allow us to make predictions of quantities that can, e.g. the measured pixel coordinates of imaged 3D points. Let the set of predictions be zθ and the set of corresponding observations be z. Then residual prediction error δz is given by: $\delta z = z - z(\theta)$ (13.24) In general, the observation vector z may be partitioned into a set of statistically independent measurements z1 . . . zN with associated predictions $z1(\theta)...zN(\theta)$. Bundle adjustment proceeds by minimizing an appropriate cost function. For a maximum likelihood parameter estimate, the cost function should reflect the likelihood of the residual δz. Under the assumption of Gaussian-distributed measurement noise, the appropriate cost function is a sum of squared errors, which is the negative sum of log likelihoods:

caption FIG 517: Triangulation illustration. Given projection matrices, a 3D point X can be computed from its measured pixel positions (u1, u2, . . .) in two or more views (C1, C2, . . .). Ideally, X should lie at the intersection of the backprojected rays (solid lines). However, because of measurement noise, these rays will not generally intersect. Hence X should be chosen so as to minimize the sum of squared errors between measured and predicted pixel positions (ui and uip).

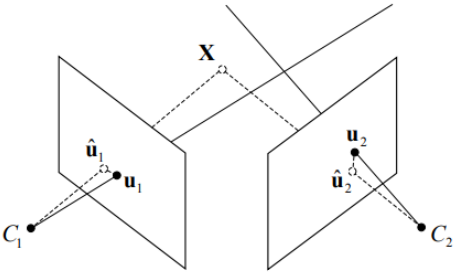

Fig. 1.13 .

Figure 5 18: Sequential registration illustration. Views 1 to 7 are registered one at a time by computing the essential matrices E12, E23, etc. relating each one to its predecessor. The essential matrix can be decomposed to give relative orientation and the direction of translation and 3D to 2D correspondences are used to determine the magnitude of the translation. As each new view is incorporated, the partial reconstruction is extended by reconstructing all 3D points that are visible in two or more views.

From the paper [35]:

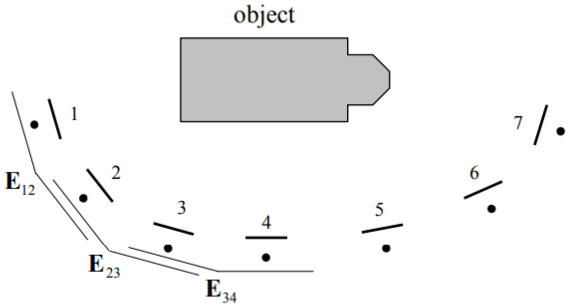

Fig. 1.14 .

This paper shows for the first time that is possible to reconstruct the position of rigid objects and to jointly recover affine camera calibration solely from a set of object detections in a video sequence. In practice, this work can be considered as the extension of Tomasi and Kanade factorization method using objects. Instead of using points to form a rank constrained measurement matrix, we can form a matrix with similar rank properties using 2D object detection proposals. In detail, we first fit an ellipse onto the image plane at each bounding box as given by the object detector. The collection of all the ellipses in the dual space is used to create a measurement matrix that gives a specific rank constraint. This matrix can be factorised and metrically upgraded in order to provide the affine camera matrices and the 3D position of the objects as an ellipsoid. Moreover, we recover the full 3D quadric thus giving additional information about object occupancy and 3D pose. Finally, we also show that 2D points measurements can be seamlessly included in the framework to reduce the number of objects required. This last aspect unifies the classical point-based Tomasi and Kanade approach with objects in a unique framework. Experiments with synthetic and real data show the feasibility of our approach for the affine camera case.

From another paper, CVPR 2012,[34].

Structure from motion (SFM) aims at jointly recovering the structure of a scene as a collection of 3D points and estimating the camera poses from a number of input images. In this paper we generalize this concept: not only do we want to recover

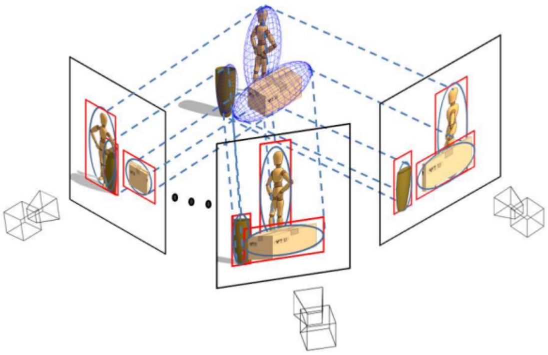

Fig. 1.15 Given multiple views with a set of objects detected in every image, the proposed factorization approach can simultaneously recover the affine camera calibration and the 3D quadrics describing the location and pose of the objects in the scene [35].

3D points, but also recognize and estimate the location of high level semantic scene components such as regions and objects in 3D. As a key ingredient for this joint inference problem, we seek to model various types of interactions between scene components. Such interactions help regularize our solution and obtain more accurate results than solving these problems in isolation. Experiments on public datasets demonstrate that: 1) our framework estimates camera poses more robustly than SFM algorithms that use points only; 2) our framework is capable of accurately estimating pose and location of objects, regions, and points in the 3D scene; 3) our framework recognizes objects and regions more accurately than state-of-the-art single image recognition methods.

.

1.5 Conclusion

In this chapter, we explored the foundational principles of Structure from Motion (SfM) for recovering 3D structure and camera motion from 2D image sequences. First

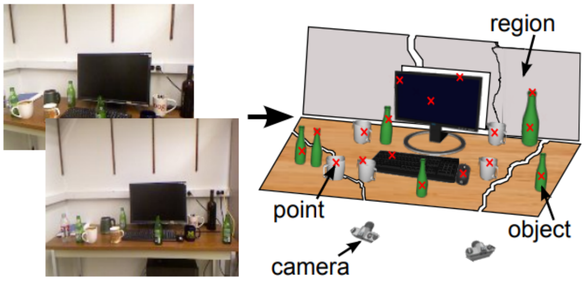

Fig. 1.16 The goal is to recognize semantic elements (e.g. cups, bottles, desk, wall, etc), localize them in 3D, and estimate camera pose from a number of semi-calibrated images. We propose to achieve this goal by modeling interactions among 3D points, regions, and objects [34].

we highlighted the importance of (SfM) in applications such as robotics, augmented reality, and scene reconstruction.

We then examined the problem of Structure from Motion , identifying the key objectives and constraints in reconstructing 3D scenes from 2D projections. We selected the appropriate projection model (orthographic) to simplify computations and reduces computational complexity while retaining sufficient accuracy for many practical applications. This makes orthographic SfM a powerful tool for controlled environments.

References

1. Seymourt Papert. The Summer Vision Project. Artificial Intelligence Group, MIT, 1966. https://dspace.mit.edu/handle/1721.1/11589
2. Bob Sumner. Augmented Creativity, TEDxZurich, 19th of January, 2015. https://www.youtube.com/watch?v=AJJOWemfOYI
3. Sachiko Iwase and Hideo Saito. Tracking soccer players based on homography among multiple views", Proc. SPIE 5150, Visual Communications and Image Processing 2003, (23 June 2003); https://doi.org/10.1117/12.502967
4. Martin A. Fischler, Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM, Volume 24, Issue 6, pp 381–395, https://doi.org/10.1145/358669.358692
5. "The Eye of a Robot: Studies in Machine Vision at MIT" and "TX-O Computer". Courtesy of MIT Museum.
6. Richard Hartley and Andrew Zisserman. Multiple View Geometry in Computer Vision. Cambridge University Press, 2000.
7. D. Scharstein and R. Szeliski. High-accuracy stereo depth maps using structured light. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2003), volume 1, pages 195-202, Madison, WI, June 2003.
8. Zhengyou Zhang. A Flexible New Technique for Camera Calibration. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 22, NO. 11, NOVEMBER 2000
9. A computer algorithm for reconstructing a scene from two projections. Nature volume 293, pages 133–135 (1981)
10. OD Faugeras, QT Luong, SJ Maybank. Camera self-calibration: Theory and experiments. Second European Conference on Computer Vision ECCV'92, 1992
11. Olsson, Carl; Enqvist, Olof, Stable Structure from Motion for Unordered Image Collections Scandinavian Conference on Image Analysis, 2011.
12. Marc Pollefeys and Luc Van Gool. 3-D Modeling from Images. COMMUNICATIONS OF THE ACM July 2002/Vol. 45, No. 7.
13. Peng-Shuai Wang, Yang Liu, Yu-Xiao Guo, Xin ; O-CNN: Octree-based Convolutional Neural Networks for 3D Shape Analysis , July 2017. ACM Transactions on Graphics, 36(4):1-11
14. M. Michalkiewicz, J. K. Pontes, D. Jack, M. Baktashmotlagh, A. Eriksson. Deep Level Sets: Implicit Surface Representations for 3D Shape Inference. 2019. arXiv:1901.06802 [cs.CV]. https://arxiv.org/pdf/1901.06802.pdf

15. A. Podlozhnyuk, S. Pirker, C. Kloss. Efficient implementation of superquadric particles in Discrete Element Method within an open-source framework. Computational Particle Mechanics 4(1), 2016. DOI: 10.1007/s40571-016-0131-6
16. M.A. Turk et al. Face recognition using eigenfaces,CVPR 1991.
17. S.K.Nayar et al., Real-Time 100 Object Recognition System, ICRA, 1996
18. C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese. 3dr2n2: A unified approach for single and multi-view 3d object reconstruction. In European conference on computer vision, pages 628–644. Springer, 2016
19. R. Girdhar, D. F. Fouhey, M. Rodriguez, and A. Gupta. Learning a predictable and generative vector representation for objects. In European Conference on Computer Vision, pages 484–499. Springer, 2016
20. D. J. Rezende, S. A. Eslami, S. Mohamed, P. Battaglia, M. Jaderberg, and N. Heess. Unsupervised learning of 3d structure from images. In Advances in Neural Information Processing Systems, pages 4996–5004, 2016
21. S. R. Richter and S. Roth. Matryoshka networks: Predicting 3d geometry via nested shape layers. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1936–1944, 2018
22. C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition., 1(2):4, 2017.
23. Y. Liao, S. Donne, and A. Geiger. Deep marching cubes: ´ Learning explicit surface representations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2916–2925, 2018
24. H. Fan, H. Su, and L. J. Guibas. A point set generation network for 3d object reconstruction from a single image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition., volume 2, page 6, 2017
25. Mateusz Michalkiewicz et al. Deep Level Sets: Implicit Surface Representations for 3D Shape Inference. arXiv:1901.06802v1 [cs.CV], 21 Jan 2019.
26. Roberts, Lawrence G. 1963. Machine perception of three-dimensional solids. Outstanding PhD dissertations in the computer sciences. Garland Publishing, New York. 1963.
27. Adolfo Guzman-Arénas, Computer Recognition of Three-Dimensional Objects In a Visual Scene. PhD Dissertations in the computer sciences, MIT 1968.
28. M.B. Clowes. On seeing things. Artificial Intelligence, Volume 2, Issue 1, Spring 1971, Pages 79-116.
29. David L. Waltz. GENERATING SEMANTIC DESCRIPTIONS FROM DRAWINGS OF SCENES WITH SHADOWS. MIT Artificial Intelligence Laboratory. Technical Report 271, November 1972.

30. Kemp, M. Julesz's joyfulness. Nature 396, 419 (1998). https://doi.org/10.1038/24753
31. Julesz B. Foundations of Cyclopean Perception. Chicago University Press; Chicago, IL, USA: 1971.
32. David Marr. Vision, A Computational Investigation into the Human Representation and Processing of Visual Information. Originally published: San Francisco : W. H. Freeman, c1982.
33. Arthur Coste. Affine Transformation, Landmarks registration, Non linear Warping. https://www.sci.utah.edu/ãcoste/uou/Image/project3/ArthurCOSTE_Project3.pdf October 2012.
34. Sid Yingze Bao, Mohit Bagra, Yu-Wei Chao, Silvio Savarese. Semantic Structure From Motion with Points, Regions, and Objects. CVPR 2012.
35. Marco Crocco, Cosimo Rubino, Alessio Del Bue. Structure from Motion with Objects, CVPR 2016.
36. D.P. Robertson and R. Cipolla. Structure from Motion. In Varga, M., editors, Practical Image Processing and Computer Vision, John Wiley, 2009.
37. C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: A factorization method. International Journal of Computer Vision, 9(2):137–154, 1992.
38. P. F. Sturm andW. Triggs. A factorization based algorithm for multi-image projective structure and motion. In European Conference on Computer Vision (ECCV'96), pages 709–720, 1996.
39. F. Schaffalitzky, A. Zisserman, R. I. Hartley, and P. H. S. Torr. A six point solution for structure and motion. In European Conference on Computer Vision (ECCV'00), pages 632–648, 2000.
40. W. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon. Bundle adjustment: A modern synthesis. In W. Triggs, A. Zisserman, and R Szeliski, editors, Vision Algorithms: Theory and Practice, LNCS, pages 298–375. Springer Verlag, 2000.
41. D. C. Brown. The bundle adjustment - progress and prospects. International Archives of Photogrammetry, 21(3), 1976.