OVERVIEW OF STRUCTURE FROM MOTION

Svetlana Mijakovska

Faculty of Technical Sciences, University St. Klement of Ohrid, Bitola, Macedonia

ABSTRACT

In this paper is given an overview of tools and techniques for obtain information about the geometry of 3D scenes from 2D images. The generating three-dimensional structure from a series of 2D images or video of scenes is known as Structure from Motion (SfM). A central tenet of structure from motion is that, given the position of a feature in one image, it is possible to find the corresponding position of the same feature in successive images. We described methods for the simultaneous recovery of 3D points and camera projection matrices using corresponding image points in multiple views. Structure from Motion (SfM) is a fascinating field within computer vision that seeks to reconstruct a three-dimensional structure of the environment from a sequence of two-dimensional images. At the heart of SfM lies a set of sophisticated algorithms that enable the extraction of spatial information from a series of images. Two public repositories that may be of use are Open SfM and Colmap.

KEYWORDS

Image processing, image projection, epipolar geometry, camera calibration, projection matrix, triangulation.

1. INTRODUCTION

3D models are used in many fields such as animation, computer games, virtual reality, film industry, computer vision, etc. The generating of 3D models can be done in many ways, from a series of images, manually and also 3D models can be generated from video [1],[2].

The generating three-dimensional structure from a series of 2D images or video of scenes is interesting area for research and this problem is known as Structure from Motion (SfM). This process recovers the 3D structure of a scene and the orientation and position of the camera when each image was captured. SfM is trying to recover, from segments of images the 3D structure of a scene and the position and pose (orientation) of camera at the moment of capturing each image.

Applications for SfM can be broadly split into two categories, those that require geometric accuracy and those that require photorealism:

- Geometric accuracy: These types of application are generally less concerned with the visual appearance of the model but require the scene structure and camera motion to be reconstructed with a high degree of accuracy. Robot navigation, for instance, requires high– accuracy models, but the visual appearance of the model is unimportant. The reverse engineering of existing objects for use in CAD requires the structure of the object to be recovered with a high degree of accuracy. Film special effects that place computer–generated objects into the film and other 'augmented reality' applications require the camera motion to be very accurately reconstructed but the appearance of the structure is irrelevant as it is never seen in the finished product.

DOI:10.5121/acij.2024.15502

- Photorealism: In contrast, there are a growing number of situations where the geometric accuracy of the underlying reconstruction is less important as long as, for the purposes of the application, the model visually resembles the real scene. This is the case for applications such as virtual reality, simulators, computer games and special effects that require a virtual set based on a real scene.

The purpose of structure from motion is getting cloud of 3D points in the scene, which in the process of feature matching and meshing, give the final 3D model. So, given the position of a feature in one image (sequence) we need to find the corresponding position of the same feature in successive image. This is correspondence problem is based on principles of multiple view geometry.

Image formation process is not generally invertible: from its projected position in a camera image plane, a scene point can only be recovered up to a one-parameter ambiguity corresponding to its distance from the camera. The additional information that we need can be received in two ways.

One possibility is to exploit prior knowledge about the scene to reduce the number of degrees of freedom. For example, parallelism and coplanarity constraints can be used to reconstruct simple geometric shapes such as line segments and planar polygons from their projected positions in individual views.

Another possibility is to use corresponding image points in multiple views. Given its image in two or more views, a 3D point can be reconstructed by triangulation [3]. An important prerequisite is the determination of camera calibration and pose, which may be expressed by a projection matrix. The geometrical theory of structure from motion allows projection matrices and 3D points to be computed simultaneously using only corresponding points in each view.

Structure from motion techniques is used in a wide range of applications including photogrammetric survey, the automatic reconstruction of virtual reality models from video sequences, and for the determination of camera motion (e.g. so that computer-generated objects can be inserted into video footage of real-world scenes).

2. The Correspondence Problem

A central tenet of structure from motion is that, given the position of a feature in one image, it is possible to find the corresponding position of the same feature in successive images. This problem, known as the correspondence problem.

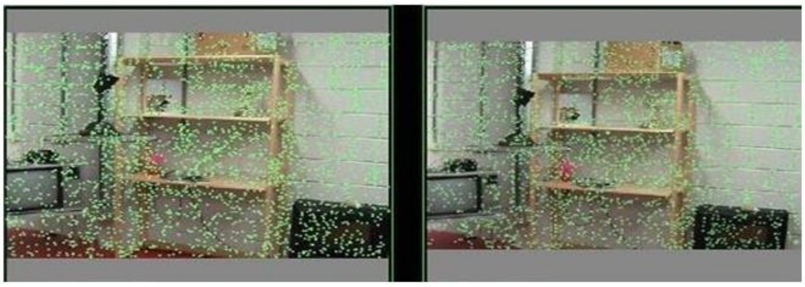

Fig.1 Points of interest detected by SIFT (coloured green)

An underlying assumption of structure from motion is that, given the position of a scene feature in one image, it is possible to find the corresponding position of the same feature in successive images. In any image of a real–world scene, there will be a considerable number of pixels lying in regions of homogeneous texture. This means that each pixel is surrounded by pixels of very similar intensity (and colour), making it virtually impossible to differentiate between the correct, corresponding, pixel and other nearby pixels. Problems are also caused by occlusion, where features lying on part of a scene may be obstructed in subsequent views by other parts of the scene and therefore not actually visible. These factors restrict the number of pixels that it is possible to match to a small percentage of the total pixels in an image sequence, resulting in a sparse set of point correspondences. This set of correspondences is also likely to contain a significant number of mismatches.First, we will need to detect a number of key points in each image (8 minimum). This cold be corners or SIFT points. The scale-invariant feature transform (SIFT) is a computer vision algorithm to detect, describe, and match local features in images. This stage is called feature extraction. Then, we will need to match each key point to its equivalent in each point of view. This is called feature matching and can be performed using template matching, optical flow.

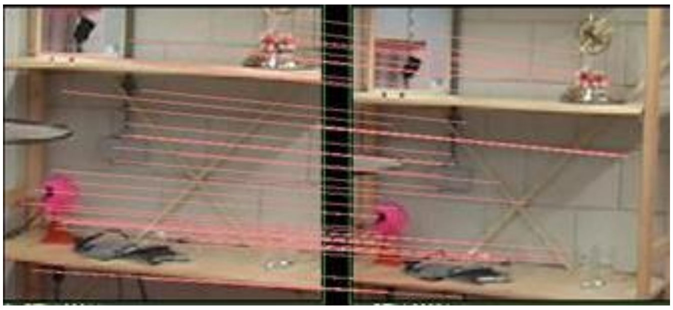

Fig.2 Corresponding position of the same feature in successive images

Another basic assumption made in SfM is that the image sequence represents a rigid scene. In other words, either the entire scene is static and the camera moves through the scene or the entire scene moves past the camera as one object. This is equivalent to multiple cameras each taking one image of a different view of the scene. Along with other assumptions, such as a pinhole camera model, the principles of multiple view geometry can be used to recover the geometry of the cameras and structure of the scene from the known correspondences, up to a projective reconstruction.

The geometrical theory of structure from motion assumes that one is able to solve the correspondence problem, which is to identify points in two or more views that are the projections of the same point in space. One solution is to identify corresponding points interactively in each view. An important advantage is that surfaces can be defined simultaneously with correspondences. A disadvantage is that the interactive approach is time consuming; also, the accuracy of the resulting reconstruction will depend critically on how carefully the user positions the image points.

Structure from Motion algorithms apply the principles of multiple view geometry to features matched across a sequence of images to recover the structure of the scene and the motion of the cameras. These features can include points [4], lines [5] and higher–level primitives such as planes [6]. In general, by far the most common approach is that of point–matching. Points features are easily extracted from images using a corner detector [7], they are widespread in most scene types and they don’t suffer from partial occlusion problems in the same way as lines and curves.

3. STRUCTURE AND CAMERA CALIBRATION

Camera calibration is the first step towards computational computer vision. Although some information concerning the measuring of scenes can be obtained by using uncalibrated cameras, calibration is essential when metric information is required. The use of precisely calibrated cameras makes the measurement of distances in a real world from their projections on the image plane possible. Some applications of this capability include: dense reconstruction, visual inspection, object localization and camera localization. Camera calibration is divided into two phases. First, camera modelling deals with the mathematical approximation of the physical and optical behaviour of the sensor by using a set of parameters. The second phase of camera calibration deals with the use of direct or iterative methods to estimate the values of these parameters. The most used techniques for camera calibration are: Non-linear optimization techniques, Linear techniques which compute the transformation matrix and Two-step techniques.

The basic formulae governing the geometric constraints for the perspective projection of scene features in two views originated in the fields of projective geometry and photogrammetry. A method of computing the epipolar geometry relating two images from 7-point correspondences is produced from Hesse [8] and Sturm [9].

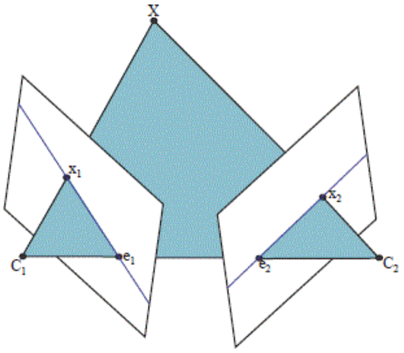

Fig.3 Epipolar geometry of two views

An important prerequisite is the determination of camera calibration and position, which may be expressed by a projection matrix. A matrix which encapsulates the epipolar geometry of two calibrated cameras is call Essential matrix. Initial research in SfM was limited almost exclusively to binocular stereo using cameras that had been previously calibrated with carefully manufactured calibration objects of known geometry. The essential matrix could also be generalised to the case of uncalibrated cameras led to the creation of the Fundamental matrix [10]. This work had important implications as it meant that it was possible to simultaneously recover the projective structure of a scene and camera motion solely from image correspondences, without any knowledge of the parameters of the camera.

Following the discovery of the fundamental matrix, that encapsulates the geometry of two views, the Trifocal tensor [11] for three views and the Quadrifocal tensor [12] for four views were derived. Although the quadrifocal tensor represents the limit for the number of views for which a closed–form solution is available, the need to produce reconstructions from long image sequences has resulted in the creation of methods that merge sets of camera matrices, derived from fundamental matrices or trifocal tensors, into one projective frame.

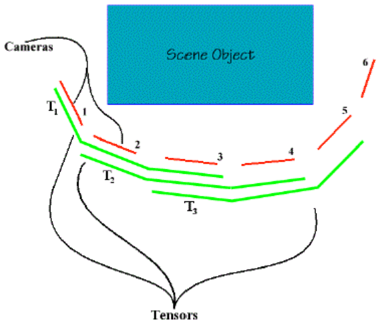

Fig.4 Geometry from N-views

Reconstruction with only knowledge of feature correspondences is only possible up to a projective reconstruction and there are many ways to obtain projection matrices from a geometry constraint, i.e. a fundamental matrix or a focal tensor. Hence projective reconstruction is mainly the recovery of fundamental matrices or focal tensors. Methods, implementation hints, and evaluations are well discussed by Hartley and Zisserman in [13]. If the input, i.e. feature correspondences, includes outliers, robust methods such as RANSAC, LMS can be employed to reject them.

4. AUTO-CALIBRATION

Camera auto-calibration is the process of determining internal camera parameters directly from multiple uncalibrated images of unstructured scenes. The intrinsic parameters and hence upgrade the reconstruction from projective to metric. Most auto–calibration methods are based on the concept of the absolute conic; a conic that is invariant to Euclidean transformations. One of the most important concepts for self-calibration is the Absolute Conic (AC) and its projection in the images. Since it is invariant under Euclidean transformations, its relative position to a moving camera is constant. For constant intrinsic camera parameters its image will therefore also be constant. This means that its relative position to a moving camera is constant and therefore its image in any view depends only on the intrinsic parameters of the camera. Recovery of the image of the absolute conic in a given view allows the intrinsic parameters of the camera for that view to be determined. A particularly convenient representation, the absolute dual quadric, projects to the dual image of the absolute conic in any view, and calculating this enables all cameras and 3D structure to be upgraded to a metric reconstruction. Accurate calibration ensures precise measurements and reliable analysis by correcting distortions and estimating intrinsic and extrinsic camera parameters.

5. 3D MODEL FROM SERIES OF IMAGES

We made experiment to create 3D model from several images. For resolve the correspondence problem we used fundament matrix and using several algorithms for SfM we create 3D model of the object which was capture with hand-held camera (Figure 5). This step is actually the main step in 3D modelling from several images, because in this step we must choose which algorithm to be used for find corresponding points of more images with moving cameras at different points in time. We used RANSAC, MSAC and MLESAC. Random sample consensus (RANSAC) is algorithm which continuously generate hypothetical solutions estimated from randomly selected, minimal data sets and testing each solution. That means the solution can be computed from the smallest sample and the likelihood of a sample containing distance is minimized. A M–estimator known as MSAC (M–estimator SAmple Consensus) solves this problem with giving outliers a fixed penalty related to the threshold and scoring inliers according to their error. The Maximum Likelihood Sample Consensus (MLESAC) algorithm is a version of RANSAC that is probabilistic. Using mixture model, it maximises a likelihood. The comparasion of these algorithms are given in Table I. The standard deviation of the prediction errors is Root Mean Square Error (RMSE).

Table I. Comparison of the speed and RMSE parameter of the algorithms for elimination of unnecessary points (outliers)

Algorithm	Speed	RMSE
MSAC	35s	0,817
RANSAC	75s	0,729
MLESAC	30s	0,699

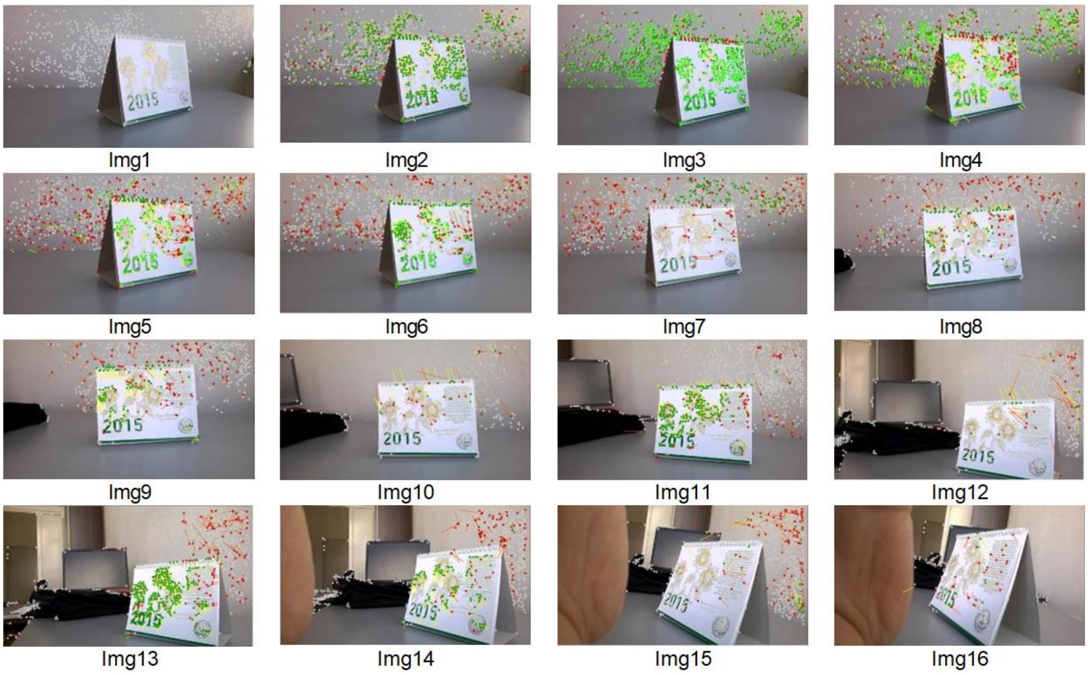

Fig.5 Cloud of 3D points in several images

After getting cloud of 3D points, we connect and match them in order to create 3D model of the object from several images. The result is shown in Figure 6.

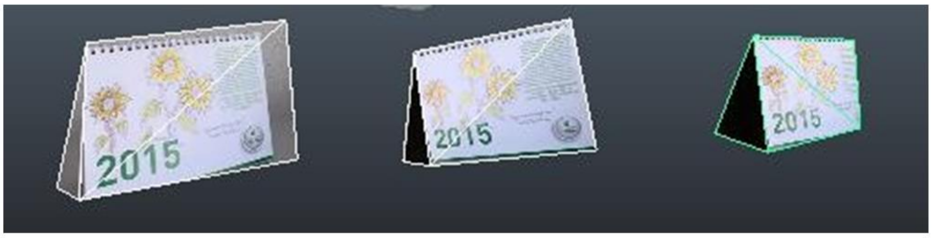

Fig.6 Final 3D model

6. CONCLUSIONS

Structure and motion recovery recovers the structure of the scene and the motion information of the camera. The motion information is the position, orientation, and intrinsic parameters of the camera at the captured views. The structure information is captured by the 3D coordinates of features. Because we want to get 3D model from several 2D images, for this step we must research 3D reconstruction from multiple views i.e. multiple view geometry. The first step is getting cloud of 3D points (using algorithms for the elimination of outliers and their comparison), and with their connecting and matching, create 3D model of object that is capture in those images. Structure from Motion (SfM) stands as a captivating realm within computer vision, bridging the gap between two-dimensional images and the intricate three-dimensional world.

The future of Structure from Motion (SfM) is poised for transformative advancements through the integration of deep neural networks. Recent developments, such as Gaussian splatting techniques or the recent paper from Oxford university and Meta AI, showcase a paradigm shift in SfM methodologies.

REFERENCES

[1] S.Mijakovska, I.Nedelkovski, F.Popovski, 2014, Generating 3D model from Video, Advanced Computing: An International Journal (ACIJ), Vol.5, No.5/6, November 2014.

[2] S.Mijakovska, I.Nedelkovski, F.Popovski, 2014, Overview of the process of 3D modelling from video, International Journal of Engineering Sciences & Emerging Technologies, Dec. 2014, ISSN: 2231 – 6604 Volume 7, Issue 3, pp: 680-686 ©IJESET.

[3] “Triangulation Method in Process of 3D Modelling from Video”, Svetlana Mijakovska, Filip Popovski, Roberto Pasic and Ivo Kuzmanov; Asian Journal of Social Science and Management Technology, 3 (6). pp. 16-21. ISSN 2313-7410, November-December 2021.

[4] H. C. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. IEEE Transactions on Robotics and Automation, 293:133 135, September 1981.

[5] R. Hartley. Projective reconstruction from line correspondences. In Proceed ings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 903–907, Seattle, USA, 1994.

[6] X. Gang, T. Jun-Ichi, and S. Heung-Yeung. A linear algorithm for camera self-calibration, motion and structure recovery for multi-planar scenes from two perspective images. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, volume 2, pages 474–479, 2000.

[7] ["Influence of Corner Detectors in the Process of 3D Modeling from Video "; S. Mijakovska, R. Pasic, I. Kuzmanov; International Journal of Scientific and Engineering Research, Volume 7, Issue 9, Sep 2016, pages 228-232, ISSN 2229-5518.

[8] O. Hesse. Die cubische gleichung, von welcher die lsung des problems der homograpie von m. chasles abh ngt. J. Reine Angew. Math., 26:188–192, 1863.

[9] R. Sturm. Das problem der projectivit t und seine anwendung auf die ae chenzweiten grades. Math. Ann., 1:533–574, 1869.

[10] O. Faugeras, Q. Luong, and S. Maybank. Camera self-calibration: Theory and experiments. In European Conference on Computer Vision, pages 321 334, 1992.

[11] L. Quan. Invariants of 6 points from 3 uncalibrated images. In Proceedings of the third European conference on Computer Vision (Vol. II), pages 459–470. Springer-Verlag New York, Inc., 1994

[12] R. Hartley. Computation of the quadrifocal tensor. In Proceedings of the 5th European Conference on Computer Vision, volume 1, pages 20–35, 1998.

[13] M. Armstrong, A. Zisserman, and R. Hartley. Self-Calibration from Image Triplets. In ECCV, volume 1, pages 3–16, 1996.

AUTHOR

Prof. Svetlana Mijakovska is a Doctor of Technical Sciences in Graphic Engineering at Faculty of Technical Sciences in Bitola, Macedonia. She is interested in computer graphics, visualization, 3d Virtual reality.