Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2012, Article ID 856523, 17 pages
doi:10.1155/2012/856523

Research Article

Key Issues in Modeling of Complex 3D Structures from Video Sequences

Shengyong Chen,[1] Yuehui Wang,[2] and Carlo Cattani[3]

[1] *College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China*
[2] *College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China*
[3] *Department of Mathematics, University of Salerno, Via Ponte Don Melillo, 84084 Fisciano, Italy*

Correspondence should be addressed to Shengyong Chen, sy@ieee.org

Received 1 July 2011; Accepted 22 August 2011

Academic Editor: Gani Aldashev

Construction of three-dimensional structures from video sequences has wide applications for intelligent video analysis. This paper summarizes the key issues of the theory and surveys the recent advances in the state of the art. Reconstruction of a scene object from video sequences often takes the basic principle of structure from motion with an uncalibrated camera. This paper lists the typical strategies and summarizes the typical solutions or algorithms for modeling of complex three-dimensional structures. Open difficult problems are also suggested for further study.

1. Introduction

Over the past two decades, many researchers seek to reconstruct the model of a three-dimensional (3D) scene structure and camera motion from video sequences taken with an uncalibrated camera or unordered photo collections from the Internet. Most traditionally, depth measurement and 3D metric reconstruction can be done from two uncalibrated stereo images [1]. Nowadays, reconstructing a 3D scene from a moving camera is one of the most important issues in the field of computer vision. This is a very challenging task because of its computational efficiency, generality, complexity, and exactitude. In this paper, we aim to show the development and current status of the 3D reconstruction algorithms on this topic.

The basic concept and knowledge of the problem can be found from the fundamentals of the multiview geometry through the books and thesis such as *Multiple View Geometry in Computer Vision* [2], *The Geometry of Multiple Images* [3], *Triangulation* [4], and some typical publications [5–8], which are independent for implementing an entire system. Multiple-view geometry is most fundamental in computer vision, and the algorithms of structure from motion are based on the perspective geometry, affine geometry, and the Euclidean geometry. For simultaneous computation of 3D points and camera positions, this is a linear algorithm framework for the Euclidean structure recovery utilizing a scaled orthographic view and perspective views based on having a reference plane visible in all views [9]. There is an affine framework for perspective views that are captured by a single extremely simple equation based on a viewer-centered invariant, called relative affine structure [10]. A comprehensive method is used for estimating scene structure and camera motion from an image sequence taken by affine cameras which can incorporate all point, line, and conic features in a unified manner [11]. The other approach tries to calculate the cameras along with the 3D points, only relying on established correspondences between the observed images. These systems and improvements are covered in many publications [2, 6, 12–15]. The literature gives a compact yet accessible overview covering a complete reconstruction system.

For multiview modeling of a rigid scene, an approach is presented in [16], which merges traditional approaches to reconstructing image-extractable features, and modeling via user-provided geometry includes steps to obtain features for a first guess of the structure and motion, fit geometric primitives, correct the structure so that reconstructed features would lie exactly on geometric primitives, and optimize both structure and motion in a bundle adjustment manner. A nonlinear least square algorithm is presented in [17] for recovering 3D shape and motion from image streams.

Sparse 3D measurements of real scenes are readily estimated from N-view image sequences using structure-from-motion techniques. There is a fast algorithm for rigid structure from image sequences in [18]. Hilton presents a geometric theory for reconstruction of surface models from sparse 3D data captured from N camera views [19] for 3D shape reconstruction by using vanishing points [20]. Relative affine structure is given for canonical model for 3D from 2D geometry and applications [10].

The paper describes the progress in automatic recovering 3D scene structures together with 3D camera positions from a sequence of images acquired by an unknown camera undergoing unknown movement [12]. The main departure from previous structure from motion strategies is that the processing is not sequential. Instead, a hierarchical approach is employed for building from image triplets and associated trifocal tensors. A method is presented for dealing with hundreds of images without precise calibration knowledge [21]. Optimizing just over the motion unknowns is fast, and given the recovered motion, one can recover the optimal structure algebraically for two images [4].

In fact, reconstruction of nonrigid scenes is very important in structure from motion. The recovery of 3D structure and camera motion for nonrigid scenes from single-camera video footages is a key problem in computer vision. For an implicit imaging model of nonrigid scenes, there is an approach that gives a nonrigid structure-from-motion algorithm based on computing matching tensors over subsequences, and each nonrigid matching tensor is computed, along with the rank of the subsequence, using a robust estimator incorporating a model selection criterion that detects erroneous image points [22]. Uncalibrated motion captures exploiting articulated structure constraints [23] such as humans. The technique shows promise as a means of creating 3D animations of dynamic activities such as sports events. For the problem of 3D reconstruction of nonrigid objects from uncalibrated image sequences, under the assumption of an affine camera and that the nonrigid object is composed of a rigid part and a deformation part, a stratification approach can be used to recover the structure of nonrigid objects by first reconstructing the structure in affine space and then upgrading it to the Euclidean space [24]. In addition, a general framework of locally rigid motion for solving the M-point and N-view structure-from-motion problem for unknown bodies deforming under orthography is presented in [25]. An incremental approach is presented in [26], where a new framework for nonrigid structure from motion simultaneously addresses three significant challenges: severe occlusion, perspective camera projection, and large non-linear deformation.

With the development of structure-from-motion algorithms, geometry constraint and optimization are necessary for reconstructing a good 3D model of the object or scene. Many researchers give us some useful approaches. For example, a technique is proposed in [27] for estimating piecewise planar models of objects from their images and geometric constraints and 3D structure from a single calibrated view using distance constraints [28]. Marques and Costeira present an approach to estimating 3D shape from degenerated sequences with missing data [29]. Beyond the epipolar constraint, it improves the effect of structure from motion [30].

3D affine measurements may be computed from a single perspective view of a scene given only minimal geometric information determined from the image. This minimal information is typically the vanishing line of a reference plane and a vanishing point for a direction not parallel to the plane. Without camera parameters, Criminisi et al. [31] show how to (i) compute the distance between planes parallel to the reference plane; (ii) compute area and length ratios on any plane parallel to the reference plane; (iii) determine the camera' location. Direct estimation is the fundamental estimation of scene structure and camera motion from a sequence of images. No computation of optical flow or feature correspondences is required [32]. A good critique on structure-from-motion algorithms can be found in [33] by Oliensis.

The remainder of this paper is organized as follows. Section 2 briefly gives some typical applications of structure from video sequences. Section 3 introduces the general reconstruction principle of structure from video sequences and unstructured photo collections. Section 4 outlines the methods for structure and motion estimation. Section 5 discusses the relevant available algorithms for every step to obtain a better result. We offer our impressions of current and future trends in the topic and conclude the development in Sections 6 and 7.

2. Typical Applications

2.1. Modeling and Reconstruction of 3D Buildings or Landmarks

For 3D reconstruction of an object or building, Pollefeys et al. typically present a complete system to build visual model with a hand-held camera [6]. There is a system for photorealistic 3D reconstruction from hand-held cameras [34]. Sinha et al. [35] present an algorithm for interactive 3D architectural models from unordered photo collections. There is a fully automated 3D reconstruction and visualization system for architectural scenes including its interiors and exteriors [36]. The system utilizes structure-from-motion, multiview stereo and a stereo algorithm.

The 3D models of historical relics and buildings, for example, the Emperor Qin's Terracotta Warriors and Piazza San Marco, have very significant meanings for archeologists. A system that can match and reconstruct 3D scenes from extremely large collections of photographs has been developed by Agarwal et al. [37]. A method for enabling existing multiview stereo algorithms to operate on extremely large unstructured photograph collections has been contrived by Furukawa et al. [38]. This approach is to decompose the collection into a set of overlapping sets of photos that can be processed in parallel and to merge the resulting reconstructions [38]. People want to sightsee the famous buildings or landscapes from the Internet; they could tour the world via building a web-scale landmark recognition engine [39].

Modeling and recognizing landmarks at world scale is a useful yet challenging task. There exists no readily available list of worldwide landmarks. Obtaining reliable visual models for each landmark can also pose problems, and efficiency is another challenge for such a large-scale system. Zheng et al. leverage the vast amount of multimedia data on the web, the availability of an Internet image search engine, and advances in object recognition and clustering techniques, to address these issues [39].

2.2. Urban Reconstruction

Modeling the world and reconstructing a city present many challenges for a visualization system in computer vision. It can use some products such as Google Earth Google Map. For instance, Pollefeys et al. [40] present a system for automatic, georegistered, real-time multiview stereo 3D reconstruction form long image sequences of urban scenes. The system collects video streams, as well as GPS and inertia measurements in order to obtain the georegistered coordinates of the 3D models [40]. Faugeras et al. [41] address the problem of recovery of a realistic textured model of a scene from a sequence of images, without any prior knowledge either about the parameters of the cameras or about their motion.

2.3. Navigation

If the world's model or the city's reconstruction is exhaustively completed, we can obtain relative location of the buildings and find related views for navigation for robots or other vision systems. Photo Tourism can enable full 3D navigation and exploration of the set of images and world geometry, along with auxiliary information such as overhead maps [14]. It gives several modes for navigation, including free-fight navigation, moving between related views, object-based navigation, and creating stabilized slideshows. The system by Pollefeys et al. also contains the navigation function [40]. Supplying realistically textured 3D city models at ground level promises to be useful for previsualizing upcoming traffic situations in car navigation systems [42].

2.4. Visual Servoing

In the literature, there are applications that can employ SfM algorithms successfully in practical engineering. For instance, based on structure from controlled motion or on robust statistics, a visual servoing system is presented in [43]. A general-purpose image understanding system via a control structure is designed by Marengoni et al. [44] and 3D video compression via topology matching [45]. More applications are being developed by researchers and engineers in the community.

2.5. Scene Recognition and Understanding

3D reconstruction is an important application to face recognition, facial expression analysis, and so on. Fidaleo and Medioni [46] design a model-assisted system for reconstruction of 3D faces from a single-consumer quality camera using a structure-from-motion approach. Park and Jain [47] present an algorithm for 3D-model-based face recognition in video.

Reconstruction of 3D scene geometry is an important element for scene understanding, autonomous vehicle and robot navigation, image retrieval, and 3D television [48]. Nedovic et al. propose accounting for the inherent structure of the visual world when trying to solve the scene reconstruction problem [48].

3. Information Organization

The goal of structure-form-motion is automatic recovery of camera motion and scene structure from two or more images. The problem of using pixel correspondences or track points to determine camera and point geometry in this manner is known as structure from motion. It is a self-calibration technique and called automatic camera tracking or match moving. We must consider several questions like

(1) Correspondence (feature extracting and tracking or matching): given a point in one image, how does it constrain the position of the corresponding point in other images?

(2) Scene geometry (structure): given point matches in two or more images, where are the corresponding points in 3D?

(3) Camera geometry (motion): given a set of corresponding points in two or more images, what are the camera matrices for these views?

Based on these questions, we can give the 3D reconstruction pipeline as in Figure 1. The goal of correspondence is to build a set of matching 2D coordinates of pixels across the video sequences. It is a significant step in the flow of the structure from motion. Correspondence is always a challenging task in computer vision. So far, many researchers have developed some practical and robust algorithms. Given a video sequence of scene, how can we find matching points?

Firstly, there are some well-known algorithms for image sequences or videos; one popular is the KLT tracker [49–51]. It gives us an integrated system that can automatically detect the KLT feature points and track them. However, it cannot apply to the situations with wide baseline, illustration changing, variant scale, duplicate and similar structure, occlusion, noise, image distortion, and so on. Generally speaking, for video sequences, the KLT tracker can perform a good effect. Figures 2 and 3 show examples of the feature points of the KLT detector output with example images from http://www.ces.clemson.edu/~stb/klt/.

In the KLT tracker [49–51], if the time interval between two frames of video is sufficiently short, we can suppose that the positions of feature points move, but their intensities do not change; that is,

$$I(\mathbf{x}, t) = I(\delta(\mathbf{x}), t + \Delta t), \tag{3.1}$$

where $\mathbf{x}$ is the position of a feature point and $\delta(\mathbf{x})$ is a transformation function.

In the papers of Lucas and Kanade [49], Tomasi and Kanade [50], and Shi and Tomasi [51], the authors made an important hypothesis that for high enough frame rates, $\delta(\mathbf{x})$ can be approximated with a displacement vector $\mathbf{d}$:

$$I(\mathbf{x}, t) = I(\mathbf{x} + \mathbf{d}, t + \Delta t). \tag{3.2}$$

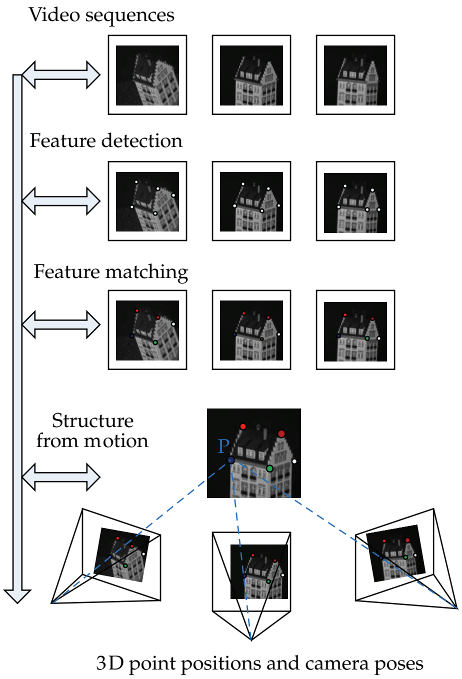

Figure 1: 3D reconstruction pipeline.

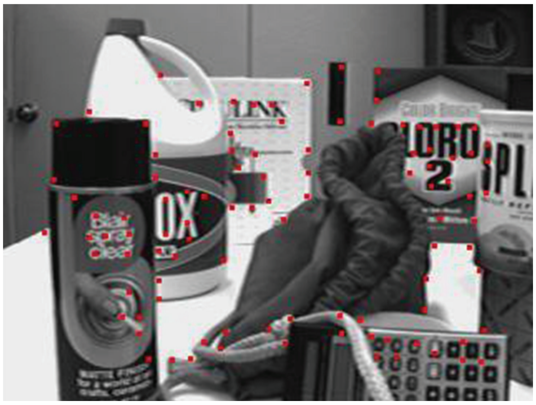

Figure 2: Example set of detected KLT features.

Then symmetric definition for the dissimilarity between two windows, one in image $I(\mathbf{x}, t)$ and one in image $I(\mathbf{x} + \mathbf{d}, t + \Delta t)$, is as follows:

$$\varepsilon = \iint_W [I(\mathbf{x} + \mathbf{d}, t + \Delta t) - I(\mathbf{x}, t)]^2 \omega(\mathbf{x}) d\mathbf{x}, \tag{3.3}$$

where $\omega(\mathbf{x})$ is the weighting function, usually set to the constant 1. The algorithm is calculating the vector $\mathbf{d}$ which minimizes. Now, utilizing the first-order Taylor expansion of

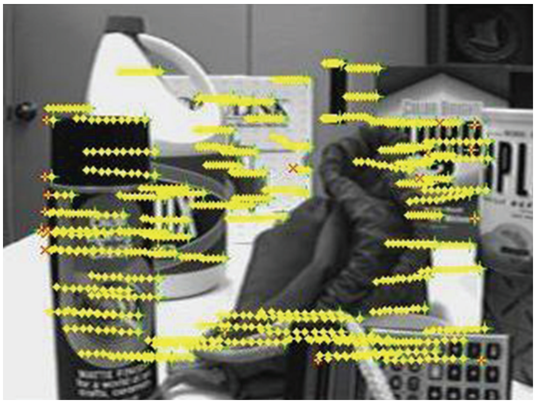

Figure 3: Tracking trajectory of KLT tracker through a video sequence.

$I(\mathbf{x}+\mathbf{d}, t+\Delta t)$ to truncate to the linear term and setting the derivative of $\boldsymbol{\varepsilon}$ with respect to $\mathbf{d}$ to $\mathbf{0}$, obtaining the linear equation:

$$Z\mathbf{d} = \mathbf{e}, \tag{3.4}$$

where Z is the following 2×2 matrix:

$$Z = \iint_W g(\mathbf{x}) g^T(\mathbf{x}) \omega(\mathbf{x}) d\mathbf{x} \tag{3.5}$$

and $\mathbf{e}$ is the following 2×1 vector:

$$\mathbf{e} = \iint_W [I(\mathbf{x}+\mathbf{d}, t+\Delta t) - I(\mathbf{x}, t)] g(\mathbf{x}) \omega(\mathbf{x}) d\mathbf{x}, \tag{3.6}$$

where $g(\mathbf{x}) = \partial I / \partial \mathbf{x}$.

On the other hand, for a completely unorganized set of images, the tracker becomes invalid. There is another popular algorithm in computer vision area, named scale-invariant feature transform (SIFT) [52]. It is effective to feature detection and matching in a wide class of image transformation, including rotations, scales, and changes in brightness or contrast, and to recognize panoramas [53]. Figures 4 and 5 show examples of the feature points of the SIFT output with example images from http://www.cs.ubc.ca/~lowe/keypoints/.

4. Structure and Motion Estimation

Assume that we have obtained a set of correspondences between images or video sequence, and then we use the set to reconstruct the 3D structure of each point in the set of correspondences and recover the motion of a camera. This task is called structure from motion. The problem has been an active research topic in computer vision since the development of the Longuet-Higgins eight-point algorithm [54] that focused on reconstructing geometry

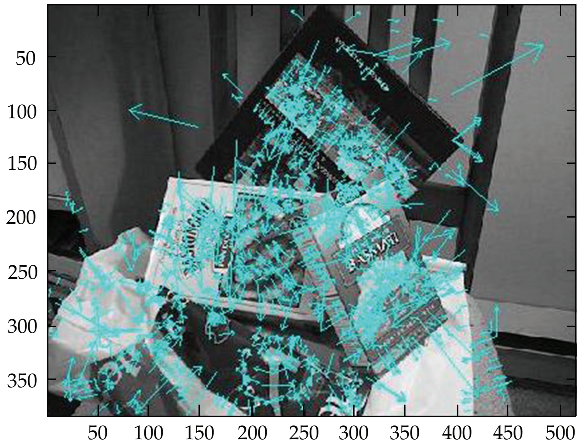

Figure 4: Example set of detected SIFT features.

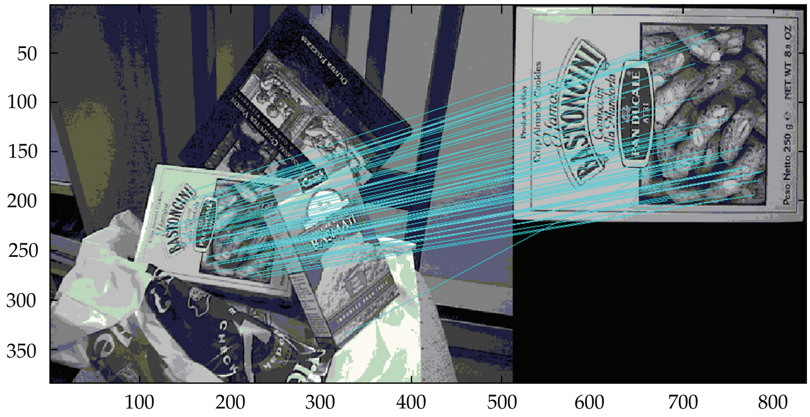

Figure 5: SIFT feature matches between images.

from two views. In the literature [2], several different approaches to solve the structure-from-motion problem are given.

4.1. Factorization

There is a popular factorization algorithm for image streams under orthography, using many images and tracking many feature points to obtain highly redundant feature position information, which was firstly developed by Tomasi and Kanade [55] in the 1990s. The main idea of this algorithm is to factorize the tracking matrix into structure and motion matrices simultaneously via singular value decomposition (SVD) method with low-rank approximation, taking advantage of the linear algebraic properties of orthographic projection.

However, an orthographic formulation limits the range of motions the method can accommodate. Perspective projection is a projection model that closely approximates perspective projection by modeling several effects not modeled under orthographic projection, while retaining linear algebraic properties [56, 57]. Poelman and Kanade [56] have developed a paraperspective factorization method that can be applied to a much wider range of motion scenarios, including image sequences containing motion toward the camera and aerial image sequences of terrain taken from a low-altitude airplane.

With the development of factorization method, a factorization- based algorithm for multi-image projective structure and motion is developed by Sturm and Triggs [57]. This technique is a practical approach for recovery of scaled feature points, using fundamental matrix and epipoles estimated from the image sequences.

Because matrix factorization is a key component for solving several computer vision problems, Tardif et al. have proposed batch algorithms for matrix factorization [58] that are based on closure and basis constraints, which handle the presence of missing or erroneous data, which often arise in structure from motion.

In mathematical expression of the factorization algorithm, assume that the tracked points are $\{(x_i^j, y_i^j) \mid i = 1, \ldots, n;\ j = 1, \ldots, m\}$. The algorithm defines the measurement matrix W : $\mathbf{W} = \left|\begin{smallmatrix}\mathbf{U}\\ \mathbf{V}\end{smallmatrix}\right|$. The rows of $\mathbf{U}$ and $\mathbf{V}$ are then registered by subtracting from each entry the mean of the entries in that row:

$$\begin{aligned} \overline{x}_i^j &= x_i^j - \frac{1}{n}\sum x_i^j, \\ \overline{y}_i^j &= y_i^j - \frac{1}{n}\sum y_i^j. \end{aligned} \tag{4.1}$$

The goal of the Tomasi-Kanade algorithm [55] is to factorize $\overline{\mathbf{W}}$ into two matrices as follows:

$$\overline{\mathbf{W}} = \mathbf{MX}, \tag{4.2}$$

where $\mathbf{M}$, named motion matrix, is a $2m \times 3$ matrix which represents the camera rotation in each frame and $\mathbf{X}$, named structure matrix, is a $3 \times n$ matrix which denotes the positions of the feature points in object space. So in the absence of the Gauss noise, rank $(\overline{\mathbf{W}}) \leq 3$.

Then we can compute SVD decomposition of $\overline{\mathbf{W}}$ to obtain $\mathbf{UDV^T}$:

$$\overline{\mathbf{W}} = \mathbf{UDV^T}, \tag{4.3}$$

where if the singular value of $\overline{\mathbf{W}}$ is $[\sigma_1, \sigma_2, \sigma_3]$, we can get the matrix $\mathbf{M} = [\sigma_1\mathbf{u_1}, \sigma_2\mathbf{u_2}, \sigma_3\mathbf{u_3}]$ and $\mathbf{X} = [\mathbf{v_1}, \mathbf{v_2}, \mathbf{v_3}]$.

The method can also handle and obtain a full solution from a partially filled-in measurement matrix, which occurs when features appear and disappear in the video due to occlusions or tracking failures [55]. This method gives accurate results and does not introduce smoothing in structure and motion. Using the above method, the problem can be solved for the video of general scene such as building and sculpture (Figure 6).

4.2. Bundle Adjustment

Bundle adjustment is a significant component of most structure from motion systems. It is the joint nonlinear refinement of camera and point parameters, so it can consume a large amount of time for large problems. Unfortunately, the optimization underlying structure from motion

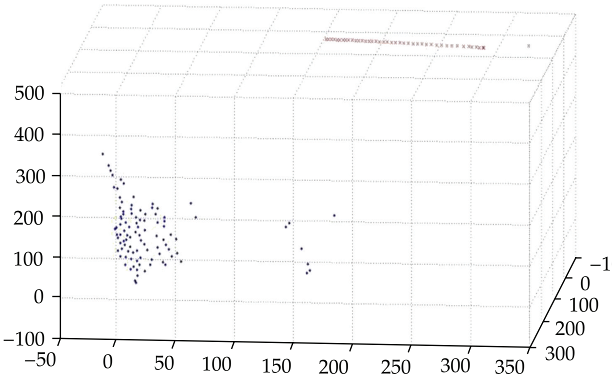

Figure 6: Example of recovering structure and motion.

involves a complex, nonlinear objective function with no closed-form solution, due to nonlinearities in perspective geometry. Most modern approaches use nonlinear least squares algorithms [17] to minimize this objective function, a process known as bundle adjustment; [53] that is, basic mathematics of the bundle adjustment problem is well understood [59]. Generally speaking, bundle adjustment is a global algorithm, but it consumes much time and cannot achieve real time to solve the minimize restriction. Mouragnon et al. [60] propose an approach for generic and real-time structure from motion using local bundle adjustment. It allows 3D points and camera poses to be refined simultaneously through the image sequence. Zhang et al. [61] apply bundle optimization to further improve the results of consistent depth maps from a video sequence.

4.3. Self-Calibration

To upgrade the projective and affine reconstruction to a metric reconstruction (i.e., determined up to an arbitrary Euclidean transformation and a scale factor), calibration techniques, to which we follow the approach described in [2, 6, 9, 15, 62], can deal with this problem. It can be done by imposing some constraints on the intrinsic camera parameters. This approach that is called self-calibration has received a lot of attention in recent years. The ambiguity on the reconstruction is restricted from projective to metric through self-calibration [6]. Mostly self-calibration algorithms are concerned with unknown but constant intrinsic camera parameters [2, 4, 12]. The paper presented the problem of 3D Euclidean reconstruction of structured scenes from uncalibrated images based on the property of vanishing points [63]. They propose a multistage linear approach, with structure from motion technique based on point and vanishing point matches in images [64].

4.4. Correlative Improvement

Traditional SFM algorithms using just two images often produce inaccurate 3D reconstructions, mainly due to incorrect estimation of the camera’ motion. Thomas and Oliensis [65] present a practical algorithm that can deal with noise in multiframe structure from motion. It describes a new incremental algorithm for reconstructing structure from multi-image sequences which estimates and corrects for the error in computing the camera motion.

The research of structure from motion has shown great progress throughout several decades, but the algorithms on structure from motion still exhibit some faults and shortages. The result of Structure from Motion cannot satisfy people in many situations. However, many researchers present a lot of improving approaches, such as dual computation of projective shape and camera positions from multiple images [66].

For incremental algorithms that solve progressively larger bundle adjustment problems, Crandall et al. present an alternative formulation for structure from motion based on finding a coarse initial solution using a hybrid discrete-continuous optimization and then improve the solution using bundle adjustment. The initial optimization step uses a discrete Markov random field (MRF) formulation, coupled with a continuous Levenberg-Marquardt refinement [67].

For time efficiency, Havlena et al. present a method of efficient structure from motion by graph optimization [68]. Gherardi et al. improve the algorithm of efficiency with hierarchical structure and motion [69].

For duplicate or similar structure, Roberts et al. couple an expectation maximization (EM) algorithm for structure from motion for scenes with large duplicate structures [70]. A hierarchical framework that resamples 3D reconstructed points to reduce computation cost on time and memory for very-large-scale structure from motion [71]. Savarese and Bao propose a formulation called semantic structure from motion (SSFM), where SSFM takes advantages of both semantic and geometrical properties associated with objects in the scene [72].

5. Relevant Algorithms

5.1. Features

(1) Line

For the problem of camera motion and 3D structure reconstruction from line correspondences across multiple views, there is a triangulation algorithm that outperforms standard linear and bias-corrected quasi-linear algorithms, and that bundle adjustment using our orthonormal representation yields results similar to the standard maximum likelihood trifocal tensor algorithm, while being usable for any number of views [73]. Spetsakis and Aloimonos [74] present a system for structure from motion using line correspondences. The recovery algorithm is formulated in terms of an objective function which measures the total squared distance in the image plane between the observed edge segments and the projections of the reconstructed lines [75]. A linear method is developed for reconstruction using lines and points simultaneously [76].

(2) Curve

Tubic et al. [77] present an approach for reconstructing a surface from a set of arbitrary, unorganized, and intersecting curves. There is an approach for reconstructing open surfaces from image data [78]. Kaminski and Shashua [79] introduce a number of new results in the context of multiview geometry from general algebraic curves, which start with the recovery of camera geometry from matching curves. Berthilsson et al. present a method for reconstruction of general curves, using factorization and bundle adjustment [80].

(3) Silhouette

Liang and Wong [81] develop an approach that produces relatively complete 3D models similar to volumetric approaches, with the topology conforming to what is observed from the silhouettes. In addition, the method neither assumes nor depends on the spatial order of viewpoints. Hartley and Kahl give us critical configurations for projective reconstruction from multiple views in [82]. Joshi et al. design an algorithm for structure and motion estimation from dynamic silhouettes under perspective projection [83]. Liu et al. present a method that is shaped from silhouette outlines using an adaptive dandelion model [84]. Yemez and Wetherilt develop a volumetric fusion technique for surface reconstruction from silhouettes and range data [85].

5.2. Other Aspects

(1) Multiview Stereo

Multiview stereo (MVS) techniques take as input a set of images with known camera parameters (i.e., position and orientation of the camera, focal length, image distortion parameters) [38, 53, 86]. We can refer to [87] for a classification and evaluation of recent MVS techniques.

(2) Clustering

There are clustering techniques to partition the image set into groups of related images, based on the visual structure represented in the image connectivity graph for the collection [88, 89].

6. Existing Problems and Future Trends

While algorithms of structure from motion have been developed for 3D reconstruction in many applications, some problems of reconstructing geometry from video sequences still exist in computer vision and photography. Until recently, however, there have been no good computer vision techniques for recovering this kind of structure from motion. Many researchers are still making efforts to improve the methods mainly in the following aspects.

6.1. Feature Tracking and Matching

Zhang et al. give a robust and efficient algorithm on efficient nonconsecutive feature tracking for structure from motion via two main steps, that is, consecutive point tracking and nonconsecutive track matching [90]. They improve the KTL tracker by the invariant feature points and a two-pass matching strategy to significantly extend the track lifetime and reduce the sensitivity of feature points to variant scale, duplicate and similar structure, and noise and image distortion. The results can be found at http://www.cad.zju.edu.cn/home/gfzhang/.

6.2. Active Vision

The method is based on the structure from controlled motion that constrains camera motions to obtain an optimal estimation of the 3D structure of a geometrical primitive [91]. Stereo geometry is acquired from 3D egomotion streams [92]. Wide-area egomotion estimation is acquired from known 3D structure [93]. A work on estimating surface reflectance properties of a complex scene under captured natural illumination can be found in [94]. Other algorithms are also attempted on selective attention of human eyes.

6.3. Unorganized Images

To solve the resulting large-scale nonlinear optimization, we reconstruct the scene incrementally, starting from a single pair of images, then adding new images and points in rounds, and running a global nonlinear optimization after each round [53]. Structure from motion could be applied to photos found in the wild, reconstructing scenes from several large Internet photo collections [14]. The large redundancy in online photo collections means that a small fraction of images may be sufficient to produce high-quality reconstructions. An investigation has begun to explore by extracting image "skeletons" from large collections [95]. Perhaps the most important challenge is to find ways to effectively parallelize all the steps of the reconstruction pipeline to take advantage of multicore architectures and cloud computing [37, 38, 53, 89].

7. Conclusion

This paper has summarized the recent development of structure from motion algorithm that is able to metrically reconstruct complex scenes and objects. The wide applications have been addressed in computer vision area. Typical contributions are introduced for feature point detection, tracking, matching, factorization, bundle adjustment, multiview stereo, self-calibration, line detection and matching, modeling, and so forth. Representative works are listed for readers to have a general overview of the state of the art. Finally, a summary of existing problems and future trends of structure modeling is addressed.

Acknowledgments

This work was supported by the National Natural Science Foundation of China and Microsoft Research Asia (NSFC nos. 61173096, 60870002, and 60802087), Zhejiang Provincial S&T Department (2010R10006, 2010C33095), and Zhejiang Provincial Natural Science Foundation (R1110679).

References

[1] H. Kuffar and K. Takaya, "Depth measurement and 3D metric reconstruction from two uncalibrated stereo images," in *Proceedings of the Canadian Conference on Electrical and Computer Engineering (CCECD '07)*, pp. 1460–1463, April 2007.

[2] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2nd edition, 2004.

[3] O. Faugeras and Q. T. Luong, *The Geometry of Multiple Images*, MIT Press, Cambridge, Mass, USA, 2001.

[4] R. I. Hartley and P. Sturm, "Triangulation," in *Proceedings of the American Image Understanding Workshop*, pp. 957–966, 1994.

[5] P. F. McLauchlan and D. W. Murray, "Unifying framework for structure and motion recovery from image sequences," in *Proceedings of the 5th International Conference on Computer Vision (ICCV '95)*, pp. 314–320, June 1995.

[6] M. Pollefeys, L. Van Gool, M. Vergauwen et al., “Visual modeling with a hand-held camera,” *International Journal of Computer Vision*, vol. 59, no. 3, pp. 207–232, 2004.

[7] S. Avidan and A. Shashua, “Trajectory triangulation: 3D reconstruction of moving points from a monocular image sequence,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 4, pp. 348–357, 2000.

[8] N. Molton and M. Brady, “Practical structure and motion from stereo when motion is unconstrained,” *International Journal of Computer Vision*, vol. 39, no. 1, pp. 5–23, 2000.

[9] A. Marugame, J. Katto, and M. Ohta, “Structure recovery with multiple cameras from scaled orthographic and perspective views,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 7, pp. 628–633, 1999.

[10] A. Shashua and N. Navab, “Relative affine structure: canonical model for 3D from 2D geometry and applications,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 9, pp. 873–883, 1996.

[11] F. Kahl and A. Heyden, “Affine structure and motion from points, lines and conics,” *International Journal of Computer Vision*, vol. 33, no. 3, pp. 163–180, 1999.

[12] A. W. Fitzgibbon and A. Zisserman, “Automatic camera recovery for closed or open image sequences,” in *Proceedings of the ECCV*, pp. 311–326, 1998.

[13] F. Schaffalitzky and A. Zisserman, “Multi-view matching for unordered image sets, or “How do I organize my holiday snaps?”,” in *Proceedings of the IEEE Conference on Computer Vision*, vol. 1, pp. 414–431, 2002.

[14] N. Snavely, S. M. Seitz, and R. Szeliski, “Photo tourism: exploring photo collections in 3D,” in *Proceedings of the International Conference on Computer Graphics and Interactive Technologies*, pp. 835–846, 2006.

[15] R. Hartley, “Euclidean reconstruction from uncalibrated views,” in *Applications of Invariance in Computer Vision*, J. L. Mundy, A. Zisserman, and D. Forsyth, Eds., Lecture Notes in Computer, 1994.

[16] A. Bartoli and P. Sturm, “Constrained structure and motion from multiple uncalibrated views of a piecewise planar scene,” *International Journal of Computer Vision*, vol. 52, no. 1, pp. 45–64, 2003.

[17] R. Szeliski and S. B. Kang, “Recovering 3D Shape and Motion from Image Streams Using Nonlinear Least Squares,” *Journal of Visual Communication and Image Representation*, vol. 5, no. 1, pp. 10–28, 1994.

[18] P. M. Q. Aguiar and J. M. F. Moura, “A fast algorithm for rigid structure from image sequences,” in *Proceedings of the IEEE International Conference on Image Processing (ICIP ’99)*, vol. 3, pp. 125–129, Kobe, Japan, 1999.

[19] A. Hilton, “Scene modelling from sparse 3D data,” *Image and Vision Computing*, vol. 23, no. 10, pp. 900–920, 2005.

[20] P. Parodi and G. Piccioli, “3D shape reconstruction by using vanishing points,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 2, pp. 211–217, 1996.

[21] M. Lhuillier, “Toward flexible 3D modeling using a catadioptric camera,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR ’07)*, pp. 1560–1567, June 2007.

[22] A. Bartoli and S. I. Olsen, “A batch algorithm for implicit non-rigid shape and motion recovery,” in *Proceedings of the International Conference on Dynamical Vision*, 2006.

[23] D. Liebowitz and S. Carlsson, “Uncalibrated motion capture exploiting articulated structure constraints,” *International Journal of Computer Vision*, vol. 51, no. 3, pp. 171–187, 2003.

[24] G. Wang and Q. M. J. Wu, “Stratification approach for 3-D euclidean reconstruction of nonrigid objects from uncalibrated image sequences,” *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 38, no. 1, pp. 90–101, 2008.

[25] J. Taylor, A. D. Jepson, and K. N. Kutulakos, “Non-rigid structure from locally-rigid motion,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR ’10)*, pp. 2761–2768, June 2010.

[26] S. Zhu, L. Zhang, and B. M. Smith, “Model evolution: an incremental approach to non-rigid structure from motion,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR ’10)*, pp. 1165–1172, June 2010.

[27] M. Farenzena and A. Fusiello, “Stabilizing 3D modeling with geometric constraints propagation,” *Computer Vision and Image Understanding*, vol. 113, no. 11, pp. 1147–1157, 2009.

[28] R. Gong and G. Xu, “3D structure from a single calibrated view using distance constraints,” *IEICE Transactions on Information and Systems*, vol. 87, no. 6, pp. 1527–1536, 2004.

[29] M. Marques and J. Costeira, “Estimating 3D shape from degenerate sequences with missing data,” *Computer Vision and Image Understanding*, vol. 113, no. 2, pp. 261–272, 2009.

[30] T. Brodsky, C. Fermuller, and Y. Aloimonos, "Structure from motion: beyond the epipolar constraint," *International Journal of Computer Vision*, vol. 37, no. 3, pp. 231–258, 2000.

[31] A. Criminisi, I. Reid, and A. Zisserman, "Single view metrology," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 123–148, 2000.

[32] H. Joachim, "Direct estimation of structure and motion from multiple zframes," MIT AI Lab. Memo 1190, Massachusetts Institute of Technology, Mass, USA, 190.

[33] J. Oliensis, "Critique of structure-from-motion algorithms," *Computer Vision and Image Understanding*, vol. 80, no. 2, pp. 172–214, 2000.

[34] T. Rodriguez, P. Sturm, P. Gargallo et al., "Photorealistic 3D reconstruction from handheld cameras," *Machine Vision and Applications*, vol. 16, no. 4, pp. 246–257, 2005.

[35] S. N. Sinha, D. Steedly, R. Szeliski, M. Agrawala, and M. Pollefeys, "Interactive 3D architectural modeling from unordered photo collections," *ACM Transactions on Graphics*, vol. 27, no. 5, article 159, 2008.

[36] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski, "Reconstructing building interiors from images," in *Proceedings of the International Conference on Computer Vision*, pp. 80–87, 2009.

[37] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski, "Building Rome in a day," in *Proceedings of the 12th International Conference on Computer Vision (ICCV '09)*, pp. 72–79, Kyoto, Japan, October 2009.

[38] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski, "Towards internet-scale multi-view stereo," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '10)*, pp. 1434–1441, San Francisco, Calif, USA, June 2010.

[39] Y. T. Zheng, M. Zhao, Y. Song et al., "Tour the World: building a web-scale landmark recognition engine," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR '09)*, pp. 1085–1092, June 2009.

[40] M. Pollefeys et al., "Detailed real-time urban 3D reconstruction from video," *International Journal of Computer Vision*, vol. 78, no. 2-3, pp. 143–167, 2008.

[41] O. Faugeras, L. Robert, S. Laveau et al., "3-D reconstruction of urban scenes from image sequences," *Computer Vision and Image Understanding*, vol. 69, no. 3, pp. 292–309, 1998.

[42] N. Cornelis, B. Leibe, K. Cornelis, and L. Van Gool, "3D urban scene modeling integrating recognition and reconstruction," *International Journal of Computer Vision*, vol. 78, no. 2-3, pp. 121–141, 2008.

[43] C. Collewet and F. Chaumette, "Visual servoing based on structure from controlled motion or on robust statistics," *IEEE Transactions on Robotics*, vol. 24, no. 2, pp. 318–330, 2008.

[44] M. Marengoni, A. Hanson, S. Zilberstein, and E. Riseman, "Decision making and uncertainty management in a 3D reconstruction system," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 7, pp. 852–858, 2003.

[45] T. Tung, F. Schmitt, and T. Matsuyama, "Topology matching for 3D video compression," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '07)*, pp. 2719–2726, June 2007.

[46] D. Fidaleo and G. Medioni, "Model-assisted 3D face reconstruction from video," in *Proceedings of the 3rd International Workshop on Analysis and Modeling of Faces and Gestures (AMFG '07)*, vol. 4778 of *Lecture Notes in Computer Science*, pp. 124–138, 2007.

[47] U. Park and A. Jain, "3D model-based face recognition in video," in *Proceedings of the Proceedings International Conference on Advances in Biometrics (ICB '07)*, vol. 4642 of *Lecture Notes in Computer Science*, pp. 1085–1094, 2007.

[48] V. Nedovic, A. W. M. Smeulders, A. Redert, and J. M. Geusebroek, "Stages as models of scene geometry," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1673–1687, 2010.

[49] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 674–679, 1981.

[50] C. Tomasi and T. Kanade, "Detection and tracking of point features," Tech. Rep. CMU-91-132, CMU, 1991.

[51] J. Shi and C. Tomasi, "Good features to track," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 593–600, June 1994.

[52] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[53] N. Snavely, I. Simon, M. Goesele, R. Szeliski, and S. M. Seitz, "Scene reconstruction and visualization from community photo collections," *Proceedings of the IEEE*, vol. 98, no. 8, Article ID 5483186, pp. 1370–1390, 2010.

[54] H. C. Longuet-higgins, “A computer algorithm for reconstructing a scene from two projections,” *Nature*, vol. 293, no. 5828, pp. 133–135, 1981.

[55] C. Tomasi and T. Kanade, “Shape and motion from image streams under orthography: a factorization method,” *International Journal of Computer Vision*, vol. 9, no. 2, pp. 137–154, 1992.

[56] C. J. Poelman and T. Kanade, “A paraperspective factorization method for shape and motion recovery,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 3, pp. 206–218, 1997.

[57] P. Sturm and B. Triggs, “A factorization based algorithm for multi-image projective structure and motion,” in *Proceedings of the 4th European Conference on Computer Vision*, 1996.

[58] J. P. Tardif, A. Bartoli, M. Trudeau, N. Guilbert, and S. Roy, “Algorithms for batch matrix factorization with application to structure-from-motion,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR ’07)*, June 2007.

[59] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, “Bundle adjustment—a modern synthesis,” in *Proceedings of the International Workshop on Vision Algorithms*, pp. 298–372, 1999.

[60] E. Mouragnon, M. Lhuillier, M. Dhome, F. Dekeyser, and P. Sayd, “Generic and real-time structure from motion using local bundle adjustment,” *Image and Vision Computing*, vol. 27, no. 8, pp. 1178–1193, 2009.

[61] G. Zhang, J. Jia, T. T. Wong, and H. Bao, “Consistent depth maps recovery from a video sequence,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 6, pp. 974–988, 2009.

[62] T. Jebara, A. Azarbayejani, and A. Pentland, “3D structure from 2D motion,” *IEEE Signal Processing Magazine*, vol. 16, no. 3, pp. 66–84, 1999.

[63] G. Wang, H. T. Tsui, and Q. M. Jonathan Wu, “What can we learn about the scene structure from three orthogonal vanishing points in images,” *Pattern Recognition Letters*, vol. 30, no. 3, pp. 192–202, 2009.

[64] S. N. Sinha, D. Steedly, and R. Szeliski, “A multi-stage linear approach to structure from motion,” in *Proceedings of the European Conference on Computer Vision (ECCV ’10)*, 2010.

[65] J. I. Thomas and J. Oliensis, “Dealing with noise in multiframe structure from motion,” *Computer Vision and Image Understanding*, vol. 76, no. 2, pp. 109–124, 1999.

[66] S. Carlsson and D. Weinshall, “Dual computation of projective shape and camera positions from multiple images,” *International Journal of Computer Vision*, vol. 27, no. 3, pp. 227–241, 1998.

[67] D. Crandall, A. Owens, N. Snavely, and D. Huttenlocher, “Discrete-continuous optimization for large-scale structure from motion,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR ’11)*, 2011.

[68] M. Havlena, A. Torii, and T. Pajdla, “Efficient structure from motion by graph optimization,” in *Proceedings of the European Conference on Computer Vision (ECCV ’10)*, 2010.

[69] R. Gherardi, M. Farenzena, and A. Fusello, “Improving the efficiency of hierarchical structure-and-moton,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR ’10)*, pp. 1594–1600, June 2010.

[70] R. Roberts, S. Sinha, R. Szeliski, and D. Steedly, “Structure from motion for scenes with large duplicate structures,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR ’11)*, 2011.

[71] T. Fang and L. Quan, “Resampling structure from motion,” in *Proceedings of the European Conference on Computer Vision (ECCV ’10)*, 2010.

[72] S. Savarese and S. Y. Z. Bao, “Semantic structure from motion,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR ’11)*, 2011.

[73] A. Bartoli and P. Sturm, “Structure-from-motion using lines: representation, triangulation, and bundle adjustment,” *Computer Vision and Image Understanding*, vol. 100, no. 3, pp. 416–441, 2005.

[74] M. E. Spetsakis and J. Aloimonos, “Structure from motion using line correspondences,” *International Journal of Computer Vision*, vol. 4, no. 3, pp. 171–183, 1990.

[75] C. J. Taylor and D. J. Kriegman, “Structure and motion from line segments in multiple images,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 11, pp. 1021–1032, 1995.

[76] R. I. Hartley, “Linear method for reconstruction from lines and points,” in *Proceedings of the 5th International Conference on Computer Vision (ICCV ’95)*, pp. 882–887, June 1995.

[77] D. Tubic, P. Hebert, and D. Laurendeau, “3D surface modeling from curves,” *Image and Vision Computing*, vol. 22, no. 9, pp. 719–734, 2004.

[78] J. E. Solem and A. Heyden, “Reconstructing open surfaces from image data,” *International Journal of Computer Vision*, vol. 69, no. 3, pp. 267–275, 2006.

[79] J. Y. Kaminski and A. Shashua, “Multiple view geometry of general algebraic curves,” *International Journal of Computer Vision*, vol. 56, no. 3, pp. 195–219, 2004.

[80] R. Berthilsson, K. Astrom, and A. Heyden, “Reconstruction of general curves, using factorization and bundle adjustment,” *International Journal of Computer Vision*, vol. 41, no. 3, pp. 171–182, 2001.

[81] C. Liang and K. Y. K. Wong, “3D reconstruction using silhouettes from unordered viewpoints,” *Image and Vision Computing*, vol. 28, no. 4, pp. 579–589, 2010.

[82] R. Hartley and F. Kahl, “Critical configurations for projective reconstruction from multiple views,” *International Journal of Computer Vision*, vol. 71, no. 1, pp. 5–47, 2007.

[83] T. Joshi, N. Ahuja, and J. Ponce, “Structure and motion estimation from dynamic silhouettes under perspective projection,” *International Journal of Computer Vision*, vol. 31, no. 1, pp. 31–50, 1999.

[84] X. Liu, H. Yao, and W. Gao, “Shape from silhouette outlines using an adaptive dandelion model,” *Computer Vision and Image Understanding*, vol. 105, no. 2, pp. 121–130, 2007.

[85] Y. Yemez and C. J. Wetherilt, “A volumetric fusion technique for surface reconstruction from silhouettes and range data,” *Computer Vision and Image Understanding*, vol. 105, no. 1, pp. 30–41, 2007.

[86] N. Snavely, S. M. Seitz, and R. Szeliski, “Modeling the world from Internet photo collections,” *International Journal of Computer Vision*, vol. 80, no. 2, pp. 189–210, 2008.

[87] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, “A comparison and evaluation of multi-view stereo reconstruction algorithms,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, pp. 519–526, June 2006.

[88] I. Simon, N. Snavely, and S. M. Seitz, “Scene summarization for online image collections,” in *Proceedings of the IEEE 11th International Conference on Computer Vision (ICCV '07)*, October 2007.

[89] S. Agarwal, Y. Furukawa, N. Snavely, B. Curless, S. M. Seitz, and R. Szeliski, “Reconstructing Rome,” *Computer*, vol. 43, no. 6, pp. 40–47, 2010.

[90] G. Zhang, Z. Dong, J. Jia, T. T. Wong, and H. Bao, “Efficient non-consecutive feature tracking for structure-from-motion,” in *Proceedings of the European Conference on Computer Vision (ECCV '10)*, 2010.

[91] E. Marchand and F. Chaumette, “Active vision for complete scene reconstruction and exploration,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 1, pp. 65–72, 1999.

[92] F. Dornaika and C. K. R. Chung, “Stereo geometry from 3-D ego-motion streams,” *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 33, no. 2, pp. 308–323, 2003.

[93] O. Koch and S. Teller, “Wide-area egomotion estimation from known 3D structure,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '07)*, pp. 437–444, June 2007.

[94] P. Debevec, C. Tchou et al., “Estimating surface reflectance properties of a complex scene under captured natural illumination,” Tech. Rep. ICT-TR-06.2004, University of Southern California Institute for Creative Technologies, Marina del Rey, Calif, USA, 2004.

[95] N. Snavely, S. M. Seitz, and R. Szeliski, “Skeletal graphs for efficient structure from motion,” in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08)*, June 2008.

Advances in
Operations Research
Advances in
Decision Sciences
Mathematical Problems
in Engineering
Algebra
Journal of
Probability and Statistics
The Scientific
World Journal
International Journal of
Differential Equations
International Journal of
Combinatorics
Advances in
Mathematical Physics
Journal of
Complex Analysis
Journal of
Mathematics
International Journal of
Stochastic Analysis
Abstract and
Applied Analysis
Discrete Dynamics in
Nature and Society
International
Journal of
Mathematics and
Mathematical
Sciences
Journal of
Discrete Mathematics
Journal of
Function Spaces
Journal of
Applied Mathematics
Journal of
Optimization