Electronic Letters on Computer Vision and Image Analysis 13(2): 42-44; 2014

3D SCENE MODELING AND UNDERSTANDING FROM IMAGE SEQUENCES

Hao Tang

Computer Information Systems Department, BMCC, CUNY, 245 Greenwich Street, New York, NY, USA
Advisor/s: Dr. Zhigang Zhu, Dr. Ioannis Stamos, Dr. Jizhong Xiao and Dr. Rakesh Kumar
6 December 2012, The City University of New York

Received 9th Feb 2014; accepted 26th May 2014

1 Abstract

Reconstructing and representing large-scale 3D scenes from image sequence have many important applications, including airborne or ground video surveillance for moving target extraction, automated 3D urban scene construction, airborne/ground traffic survey, and image-based modelling and rendering. To deal with two major challenges in large-scale 3D modelling - huge amount of data and intensive computation, we propose a dynamic mosaic-based approach that can model a large scale 3D scene with dynamic objects (Fig. 1a). The work in this thesis has the following four original contributions:

- First, we extend the previous work on stereo mosaics from static scenes to dynamic scenes, thus allowing the handling of independent moving objects. This is significant in low-altitude aerial video surveillance for detecting moving targets (such as vehicles on roads) from an aerial platform.
- Second, an effective and efficient patch-based stereo matching algorithm has been proposed to extract both 3D and motion information from stereo mosaics of urban scenes, which allows 3D reconstruction of scenes with a lot of texture-less regions with sharp depth boundaries (such as buildings).
- Third, we perform thorough experimental analysis of the robustness and accuracy of 3D reconstruction using parallel-perspective stereo mosaics, showing the advantage of the parallel-perspective representation.
- Fourth, a content-based 3D mosaic (CB3M) representation is proposed, and a graph-based higher-level scene understanding algorithm is proposed. The method shows the advantage of CB3M data representation in grouping planar patches into higher-level object entities, such as roads, buildings, etc.

A Three-Phase Approach

We have developed a three-phase procedure for this goal, as shown in Fig. 1(a). In the first phase, a set of parallel-perspective (pushbroom) mosaics with varying viewing directions is generated to capture both the 3D and dynamic aspects of the scene under the camera coverage. The set of the multi-view dynamic pushbroom mosaics, with a pair of stereo mosaics as the minimum sub-set, is a compact visual representation for a long video sequence of a 3D scene with independent moving targets.

Correspondence to: haotang@gmail.com

Recommended for acceptance by <Alicia Fornés and Volkmar Frinken>
ELCVIA ISSN: 1577-5097
Published by Computer Vision Center / Universitat Autonoma de Barcelona, Barcelona, Spain

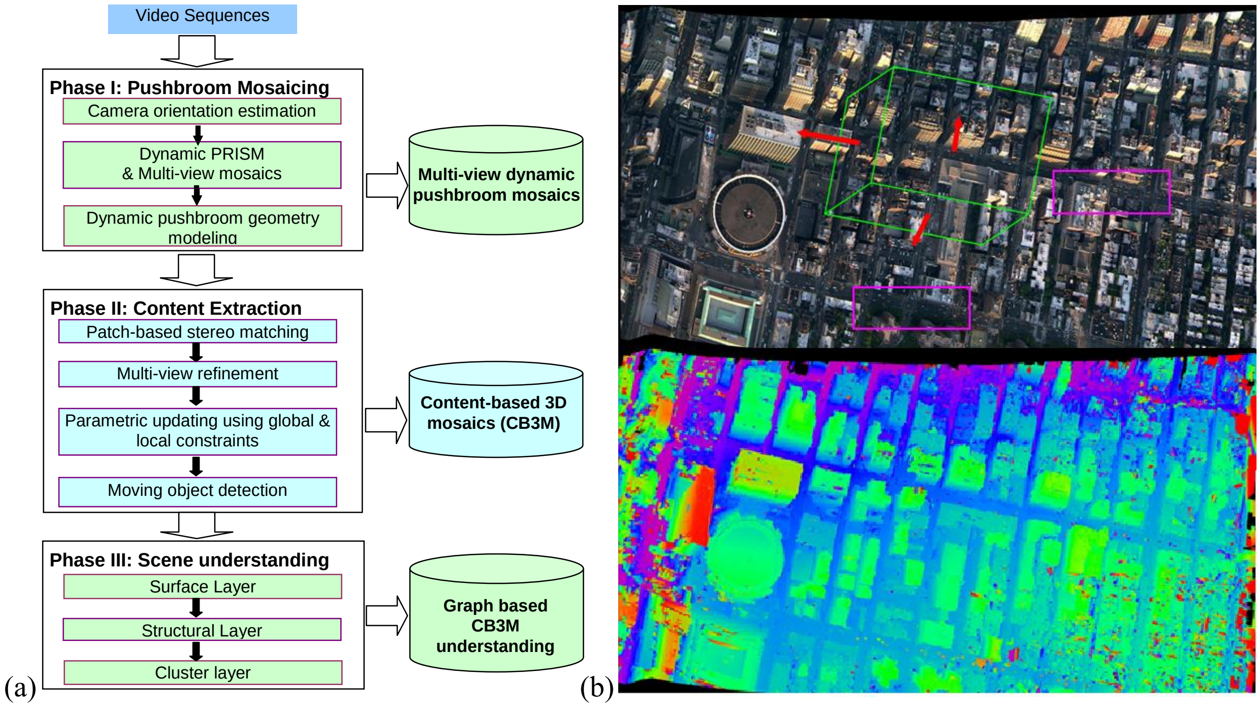
Video Sequences
Phase I: Pushbroom Mosaicing
Camera orientation estimation
Dynamic PRISM
& Multi-view mosaics
Dynamic pushbroom geometry modeling
Multi-view dynamic pushbroom mosaics
Phase II: Content Extraction
Patch-based stereo matching
Multi-view refinement
Parametric updating using global & local constraints
Moving object detection
Content-based 3D mosaics (CB3M)
Phase III: Scene understanding
Surface Layer
Structural Layer
Cluster layer
Graph based CB3M understanding
(a)
(b)

Figure 1. (a) System diagram. (b) Mosaic and color-coded depth image generated from one minute aerial image sequence

However, the 2D mosaic representation is still an image-based one without object content representations. Therefore, in the second phase, a segmentation-based stereo matching approach is proposed to extract parametric representation of the color, structure and motion of the dynamic and/or 3D objects in urban scenes, where a lot of planar surfaces exist. In our approach, we use the fact that all the static objects obey the epipolar geometry, i.e. along the epipolar lines of pushbroom stereo. An independent moving object, on the other hand, either violates the epipolar geometry if the motion is not in the direction of sensor motion, or exhibits unusual 3D structure otherwise, e.g., obviously hanging above the road or hiding below the road. Furthermore, multiple pairs of stereo mosaics and local/global spatial constraints are used for facilitating reliable stereo matching, occlusion handling, accurate 3D reconstruction and robust moving target detection.

Based on the above two phases, in the third phase, a content-based 3D mosaic (CB3M) representation is created for a long video sequence. This is a highly compressed visual representation for the video sequence of a dynamic 3D scene. More importantly, the CB3M representation has high-level object contents. A scene is represented in parametric forms of planar regions with their 3D, their boundaries, their motion, and their relations. In the third phase, the CB3M representation is used for a higher-level scene understanding. The proposed graph-based method parses the CB3M in three steps that is the generation of three layers, including a surface layer, a structural layer and a cluster layer.

Experimental Results

The proposed approach for the content-based 3D mosaic representations was applied to multi-view pushbroom mosaics generated from some real world video sequences (Fig. 1b). In addition, we also performed evaluations on the accuracy of 3D and motion estimation on a simulated video sequence generated with the ground truth data (Fig. 2). Fig. 2a is one of the mosaics generated in Phase 1; the depth map in (b) is computed based on total 9 mosaics in Phase 2; the CB3M representation in (c) is built in Phase 3, after 3D model is constructed and moving targets are identified. Each region (generated in the segmentation step) is rendered by its average color, plane parameters (a,b,c,d) (in blue) and boundaries for several representative surfaces, and motion displacements (sx, sy) (in red) of the detected moving targets are labelled. Fig. 1b shows the results of a real-world scene - NYC mid-town, including a mosaic image generated from one minute aerial video and the color-coded depth map.

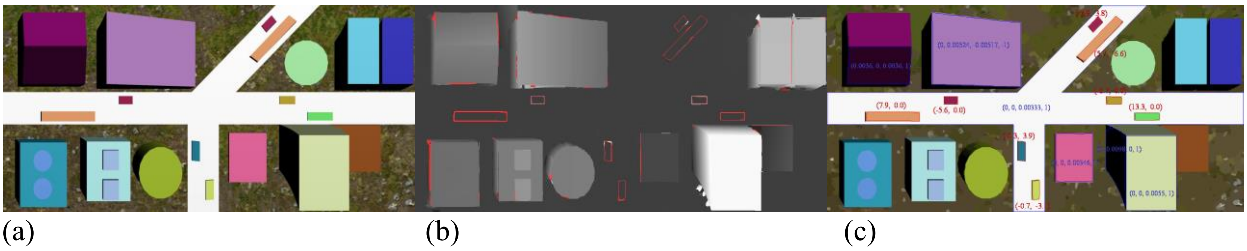

Figure 2. (a) The leftmost mosaic, (b) depth map and (c) CB3M data representation.

References

[1] H. Tang, 3D Scene Modeling and Understanding from Image Sequences, Doctoral Dissertation, Department of Computer Science, City University of New York New York, NY, USA, December 2013. ISBN: 978-1-267-92574-9. http://dl.acm.org/citation.cfm?id=2520689

[2] H. Tang and Z. Zhu, Content-Based 3D Mosaics for Representing Videos of Dynamic Urban Scenes, IEEE Transactions on Circuits and Systems for Video Technology, 22(2), 295-308, 2012. DOI: 10.1109/TCSVT.2011.2178729