Computer Vision and Image Understanding 139 (2015) 104–121

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Block world reconstruction from spherical stereo image pairs

Hansung Kim*, Adrian Hilton

Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, Surrey, GU2 7XH, UK

ARTICLE INFO

Article history:
Received 22 May 2014
Accepted 6 April 2015

Keywords:
3D reconstruction
Scene modelling
Spherical imaging
Block world interpretation

ABSTRACT

We propose a block-based scene reconstruction method using multiple stereo pairs of spherical images. We assume that the urban scene consists of axis-aligned planar structures (Manhattan world). Captured spherical stereo images are converted into six central-point perspective images by cubic projection and façade alignment. Depth information is recovered by stereo matching between images. Semantic regions are segmented based on colour, edge and normal information. Independent 3D rectangular planes are constructed by fitting planes aligned with the principal axes of the segmented 3D points. Finally cuboid-based scene structure is recovered from multiple viewpoints by merging and refining planes based on connectivity and visibility. The reconstructed model efficiently shows the structure of the scene with a small amount of data.

1. Introduction

3D scene reconstruction from photographic images has been an important research topic for various domains. Applications include visual sets in film and game production, 3D map generation, virtual tourism and urban planning. There have been many studies into outdoor scene reconstruction from multi-view images [1–3]. Strecha et al. created a benchmarking site for the quantitative evaluation of algorithms against ground-truth by LIDAR scanning [4]. However, the quality of pure image-based reconstruction largely depends on the capture environment.

Firstly, real environments include complex appearance causing errors in reconstruction from images. Textureless and non-Lambertian surfaces often result in errors in matching and reconstruction. Scenes reflected on glass or water induce false depth. Moving pedestrians and cars in the scene can be occluders in urban scene modelling.

Another problem is that normal cameras with a limited field-of-view (FOV) capture only a partial observation of the surrounding environment. Reconstruction of a complete model of the 3D environment requires additional views to capture the scene and occluded regions. Reconstruction of scene models from multiple images or video acquired with a standard camera has been the focus of considerable research. However, the limited FOV presents a challenging problem to ensure complete scene coverage for reconstruction. Agarwal et al. [5] reconstructed full 3D street models from 150,000 photos from the internet using grid computing. Pollefeys et al. [6] used 3000 video frames to reconstruct one building and 170,000 frames for a small town. The relatively narrow FOV and low resolution of normal cameras require acquisition and processing of large image sets for scene model reconstruction.

Finally, conventional dense reconstruction methods such as LIDAR scans or image-based reconstruction result in millions of points with a high-level redundancy which do not efficiently represent the scene structure. The task of extracting a structured representation for subsequent visualisation is typically performed manually. When we applied our previous dense reconstruction algorithm [7] for datasets covering areas of 30 m diameter surrounded by buildings, it produced more than 100 million faces with 60 million vertices. This occupies huge amount of system memory and may require out of core techniques [8] to visualise and render. Applications such as 3D structure representation and pre-visualisation require scene models in a structured form for efficient storage, transmission and rendering.

Piecewise-planar, plane-based and block-based scene modelling methods provide a good solution for the above problems. These approaches start from the assumption that man-made environments such as urban areas or building interiors are composed of piecewise planar surfaces. Furukawa et al. [9] and Gupta et al. [10] used the strong assumption of a piecewise-axis-aligned-planar world (Manhattan world).

We previously presented a dense environment model reconstruction [7] and a plane-based reconstruction [11] using a line-scan camera and manual segmentation. In this paper, we propose an automatic block-based environment model reconstruction method based on the same input data. This produces a more complete scene model with a compact representation for storage and transmission. The geometry can be refined for higher resolution mesh models if dense depth information is available. The approach provides a compact scene

* Corresponding author.
E-mail address: h.kim@surrey.ac.uk, coolhs99@hotmail.com (H. Kim).

http://dx.doi.org/10.1016/j.cviu.2015.04.001
1077-3142/

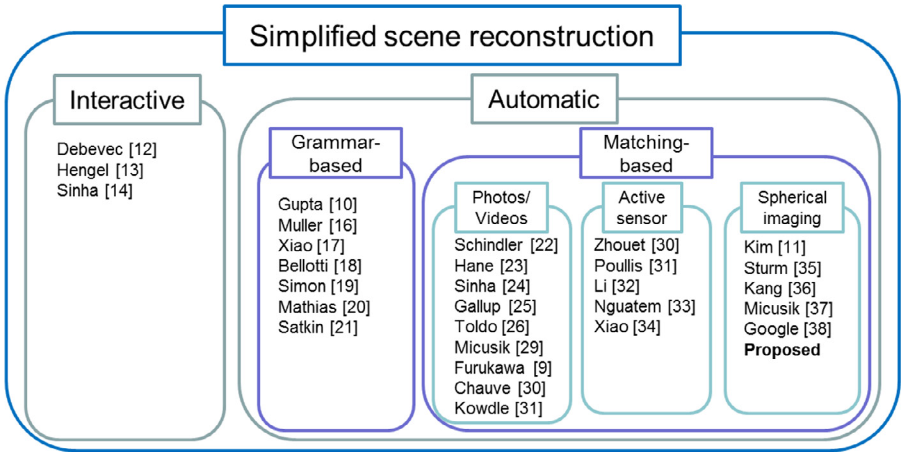

Fig. 1. Categories of simplified scene modelling methods.

model for hierarchical geometry representation of the detailed scene structure.

The main contributions of this paper are:

- We propose a 3D block-based scene reconstruction system. This is a simple and efficient way to represent the structure of a scene with high completeness for transmission and interactive visualisation.
- Spherical stereo imaging enables full scene reconstruction with a small number of input images. This saves considerable time in scene capture and reconstruction.
- We propose a façade alignment algorithm to find regions in the scene for optimal alignment and cubic projection. Cubic projection decomposes the spherical image into six central-point perspective images. The central-point perspective image is advantageous in feature matching and 3D plane reconstruction because it is distortion-free and has a vanishing point at the centre of the image aligned with the principal axes for a Manhattan world.
- We propose an automatic extraction of plane and cuboid structure from colour and depth images. Optimal block-based representation of the scene is recovered based on visibility, occupancy, point density and physical stability.
- We provide an optional user interaction to constrain primitive reconstruction to keep specific geometrical details or refine erroneous regions.
- High resolution texture mapping from the original images to the block-based representation gives a quick rendering of the scene.

The rest of this paper is organised as follows: Section 2 introduces related previous works and Section 3 outlines overview of the proposed method. Section 4 presents capture method and cubic projection with façade alignment. Depth reconstruction and region segmentation methods are proposed in Section 5. In Section 6, we introduce plane primitives reconstruction and structured block reconstruction methods. Experimental results and discussion are given in Section 7, and Section 8 makes conclusions of this work. Supplemental video is also available at: http://www.cvssp.org/hkim/BlockWorld/BlockRecon-CVIU.mov showing results of reconstruction for various scenes.

2. Related work

Simplified scene modelling has been a long-standing area of research. Previous approaches can be separated into two categories: interactive and fully automatic methods. The automatic methods are divided into grammar-based and matching-based approaches according to the registration strategy and the matching-based approach uses various input modalities as illustrated in Fig. 1

The FAÇADE system introduced by Debevec et al. [12] pioneered interactive environment modelling from images. In this approach, a simplified geometric model of the architecture is recovered interactively with manual correspondence using multiple view geometry. Novel views are rendered using view-dependent texture mapping, and additional geometric detail is recovered automatically through stereo correspondence. Their research was commercialised as ImageModeler.[1]

Hengel et al. [13] proposed an interactive 3D modelling method from video frames by tracing the shape of objects in the scene. They used structure from motion (SfM), feature point tracking and superpixel segmentation to get 3D information from 2D video frames. If users draw 2D primitives such as lines and circles on frames, then the system automatically builds 3D primitives from the user's input and reconstructed 3D information. This concept was extended by Sinha et al. [14] using feature-matching and SfM methods with line detection and vanishing point detection algorithms for interactive 3D architectural modelling from photo collections. SketchUp[2] provides a simple 3D reconstruction tool from multiple photos. This is similar to the Sinha's method but it does not use any matching method, just manual vanishing point alignment for photo registration to 3D coordinates. This tool is useful to build very simple scenes but has limitations in building complex scenes because it requires manual matchings for each primitive.

Automatic scene reconstruction can be divided into two categories: grammar-based and matching-based reconstruction. Grammar-based reconstruction uses semantic region detection and recognition to compose the world according to pre-defined rules. Gupta et al. [10] proposed block world reconstruction from a single outdoor image, inspired by the "Blocks World" work in the 1960s and Hoiem et al.'s "pop up 3D" [15]. They assume that the world is composed of blocks and match 2D image regions into 3D block view classes. They also estimate the density of each block using visual cues and use it to generate 3D parse graphs which describe geometric and mechanical relationships between objects within an image. Muller et al. [16] proposed a rule-base city modelling method using shape grammar rules from façade images, now commercialised as CityEngine.[3] Xiao et al. [17] proposed an automatic approach to generate street-side 3D photo-realistic models from images captured along streets at ground level with an assumption that building façades have two principal directions. They use a SfM method for the initial

[1] ImageModeler, http://usa.autodesk.com/adsk/servlet/pc/index?id=11390028&siteID=123112.

[2] SketchUp, http://www.sketchup.com/.

[3] City Engine, http://www.esri.com/software/cityengine/.