Computer Vision and Image Understanding 000 (2016) 1–12

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Efficient 3D scene abstraction using line segments

Manuel Hofer*, Michael Maurer, Horst Bischof

Institute for Computer Graphics and Vision, Graz University of Technology, Graz, Austria

ARTICLE INFO

Article history:
Received 10 November 2015
Revised 17 March 2016
Accepted 24 March 2016
Available online xxx

Keywords:
Structure-from-Motion
3D reconstruction
Line segments
Scene abstraction
Multi-view Stereo

ABSTRACT

Extracting 3D information from a moving camera is traditionally based on interest point detection and matching. This is especially challenging in urban indoor- and outdoor environments, where the number of distinctive interest points is naturally limited. While common Structure-from-Motion (SfM) approaches usually manage to obtain the correct camera poses, the number of accurate 3D points is very small due to the low number of matchable features. Subsequent Multi-view Stereo approaches may help to overcome this problem, but suffer from a high computational complexity. We propose a novel approach for the task of 3D scene abstraction, which uses straight line segments as underlying features. We use purely geometric constraints to match 2D line segments from different images, and formulate the reconstruction procedure as a graph-clustering problem. We show that our method generates accurate 3D models with low computational costs, which makes it especially useful for large-scale urban datasets.

1. Introduction

Recovering 3D information from an image sequence used to be a very challenging and time consuming task. Today, thanks to freely available software (Moulon et al.; Schoenberger et al., 2014; Snavely et al., 2006; Sweeney; Wu, 2013) or sophisticated professional tools (Agisoft photoscan; Capturingreality; Strecha et al.), even non-expert users are able to generate accurate 3D models from arbitrary scenes within hours. Since these so-called Structure-from-Motion (SfM) approaches operate on a sparse set of distinctive feature points (e.g., SIFT Lowe (2004) features), the resulting 3D point cloud is usually quite sparse as well. The important part of the SfM result are the camera poses for each input image, which enable subsequent Multi-View Stereo (MVS) pipelines (e.g., PMVS Furukawa et al. (2010) or SURE (Rothermel et al., 2012)) to create a (semi-) dense point cloud.

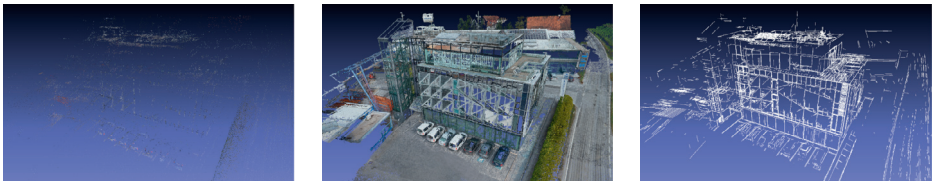
While the first part of this two-step procedure (pose estimation via SfM) can be computed very efficiently even for large crowd-sourced datasets (Frahm et al., 2010; Havlena and Schindler, 2014), the second part (dense reconstruction via MVS) is still computationally expensive and can take up to several days on modern desktop computers. Moreover, the resulting 3D point cloud easily consists of millions of points and just viewing it in a point-cloud viewer becomes a very tedious task. The same holds for any kind of automatic data analysis or post processing (e.g., meshing Labatut et al. (2007) and texturing (Waechter et al., 2014)). This is due to using point clouds as 3D representation. On the one hand, shapes of arbitrary complexity can be described by a set of 3D points, but on the other hand, the number of points needed to do so can explode very quickly.

Desirable is an efficient way of abstracting the 3D model, to decrease the amount of data without reducing the embedded 3D information. A natural choice is to use more complex geometric primitives as data representation, such as planes (e.g., Kim and Manduchi (2014); Raposo et al. (2014); Sinha et al. (2009)) or lines (e.g., Hofer et al. (2015); Micusik and Wildenauer (2014); Zhang and Koch (2014)). While this might not be sufficient for natural scenes (e.g., forests, etc.), it is especially useful for urban indoor- and outdoor environments, where most of the structures are piecewise planar/linear.

In this paper, we propose a system, denoted as *Line3D++*, to generate a semantically meaningful 3D model of urban environments, by using 2D line segments as image features. Our method uses an oriented image sequence as input, whose camera poses can be obtained by any conventional SfM pipeline. We use weak epipolar constraints to establish a large set of potential line correspondences between images, and use a scoring formulation based on mutual support to separate correct from incorrect matches for each segment individually. We obtain a final line-based 3D model by clustering 2D segments from different views, using an efficient graph-clustering formulation. In addition, we show how the SfM result as well as the 3D line model can be further improved by employing a combined bundle adjustment over the reconstructed points and lines. Our method scales almost linearly with the

* Corresponding author. Tel.: +433168735090.
E-mail addresses: hofer@icg.tugraz.at (M. Hofer), maurer@icg.tugraz.at (M. Maurer), bischof@icg.tugraz.at (H. Bischof).
URL: http://www.icg.tugraz.at (M. Hofer)

http://dx.doi.org/10.1016/j.cviu.2016.03.017
1077-3142/



SfM [23]
97,699 points
(a) runtime: 0.75 hours

PMVS [15]
12,156,664 points
(b) runtime: > 11 hours

Line3D++ (proposed)
7,877 lines
(c) runtime: **210.65 sec**

Fig. 1. Three different 3D representations of the *Building* sequence (344 images). (a) Sparse 3D model (Irschara et al., 2007) (SIFT Lowe (2004) features). (b) Semi-dense point-cloud (PMVS Furukawa et al. (2010)). (c) 3D line model using *Line3D++*. It is hardly possible to recognize the building in the sparse 3D model, while it is clearly recognizable in both the semi-dense- and the line-based 3D model. Compared to PMVS, our method has much lower runtime- and memory requirements.

number of input images and has low memory requirements, to easily reconstruct scenes from hundreds of images. We demonstrate our approach on several challenging real-world datasets, as well as a synthetic sequence with ground truth. The implementation of this work is freely available,[1] and works off-the-shelf with the output of several state-of-the-art SfM pipelines (Moulon et al.; Schoenberger et al., 2014; Snavely et al., 2006; Strecha et al.; Wu, 2013).

Fig. 1 shows a comparison between a sparse-, a dense-, and a line-based 3D model for an urban scene. As we can see, our reconstruction provides a high degree of semantically meaningful 3D information, despite its sparsity compared to the dense model. Moreover, our method is much more efficient than computing a dense 3D model, with a runtime of just a few minutes even for this large-scale dataset.

2. Related work

While line segments have been used for tasks such as image registration or 3D reconstruction for a long time (e.g., Ayache and Faverjon (1987); Baillard et al. (1999); Schmid and Zisserman (1997)), in recent years image-based 3D reconstruction has been dominated by the use of image feature-points and their invariant descriptors (e.g., SIFT Lowe (2004)). Only quite recently, the principles of feature-point descriptors have been successfully adapted to the task of line segment matching (e.g., Wang et al. (2009); Zhang et al. (2012, 2011)), but line-based 3D reconstruction for real-world scenarios is still rarely used.

Bartoli and Sturm (2005) proposed a full SfM pipeline based on line segments, including bundle adjustment. They make use of Plücker coordinates as 3D line representations and a trifocal tensor, since two views do not put enough constraints on the camera motion when using lines alone. In an earlier work, they successfully used line segments for SfM with only two images at a time, by establishing point-wise correspondences between lines from different images (Bartoli et al., 2004). Bay et al. (2006) use line segments from two un-calibrated images to determine relative camera poses, and to compute a piecewise planar 3D model. They establish correspondences between the images by using a color histogram based line descriptor (Bay et al., 2005). Further, Schindler et al. (2006) demonstrated how the Manhattan-world assumption can be incorporated into the reconstruction procedure, to decrease the computational complexity and to reconstruct buildings from two views.

Recently, Zhang and Koch (2014) proposed a very sophisticated line-based SfM pipeline. They initialize the reconstruction from matched lines across three images in closed from, and incrementally add new images by establishing 3D-2D line correspondences for absolute pose estimation. Additionally, they introduce the *Cayley* 3D line representation, which can be efficiently derived from Plücker coordinates, and only requires four parameters to encode a 3D line. This representation allows a more efficient optimization of 3D lines during bundle adjustment. In the same year, Micusik and Wildenauer (2014) presented a SLAM-like line-based SfM system, which is able to reconstruct large-scale urban scenes with line segments alone. They use relaxed endpoint constraints for line matching, which requires narrow baselines. They showed impressive results, especially for indoor scenes.

The aforementioned methods try to solve both camera pose estimation and 3D reconstruction using only line segments. This is especially challenging for pose estimation, since line correspondences between two images do not put enough constraints on the epipolar geometry for the general case. Hence, additional constraints are needed to solve for the unknown camera orientations (e.g., explicit point-wise correspondences between lines (Bartoli et al., 2004), a trifocal tensor (Bartoli and Sturm, 2005), narrow baselines (Micusik and Wildenauer, 2014), or the Manhattan-world assumption (Kim and Manduchi, 2014)). Related to these methods, several approaches for the task of line-based pose estimation have been presented over the years, e.g., by utilizing special line triplets for relative- (Elqursh and Elgammal, 2011), and 2D-3D line correspondences for the absolute pose problem (Zhang and Koch, 2012).

In contrast to pure line-based SfM, several MVS algorithms that focus on the reconstruction procedure with given camera poses have been presented. Jain et al. (2010) proposed a method that does not require explicit correspondences between line segments from different images, which enables 3D reconstruction under difficult lighting conditions or around highly non-planar objects (such as power pylons), where patch-based line descriptors would fail. They formulate the reconstruction procedure as an optimization problem, where the unknown depth of the endpoints of 2D line segments in the images is modelled as a random variable. They compute the most probable 3D locations for the segment endpoints by minimizing the re-projection error among several neighboring views, and compute a final 3D model by merging individual 3D hypotheses that are sufficiently close together. While their approach generates very promising and visually pleasant results, the continuous optimization of the endpoint depths (in a potentially large range) renders the method inefficient for large-scale datasets.

Since the approach by Jain et al. (2010) is very versatile, we decided to go a similar way in our research. To overcome the high runtime requirements, we switch from using no explicit 2D line correspondences at all to using weak epipolar matching constraints, without any kind of appearance information. We compute a set of potential matches for each 2D line segment in its neighboring images, and limit its potential 3D locations to a discrete set coinciding with these matches. We have shown how this simple modification significantly boosts the performance without negatively affecting the accuracy (Hofer et al., 2013a; 2013b). We further replace the greedy line-merging from Jain et al. (2010) with a scale invariant graph clustering formulation (Hofer et al., 2014b),

[1] https://github.com/manhofer/Line3Dpp.

Please cite this article as: M. Hofer et al., Efficient 3D scene abstraction using line segments, Computer Vision and Image Understanding (2016), http://dx.doi.org/10.1016/j.cviu.2016.03.017