Computers & Graphics 39 (2014) 89–100

Contents lists available at ScienceDirect

Computers & Graphics

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

Technical Section

High-resolution depth for binocular image-based modeling☆

David C. Blumenthal-Barby [a,b,*], Peter Eisert [a,b,1]

[a] Fraunhofer Heinrich Hertz Institute, Berlin, Germany
[b] Humboldt Universität zu Berlin, Berlin, Germany

ARTICLE INFO

Article history:
Received 23 August 2013
Received in revised form
2 December 2013
Accepted 2 December 2013
Available online 21 December 2013

Keywords:
3D reconstruction
Image-based modeling
Optimization
Interaction in 3D reconstruction

ABSTRACT

We propose a binocular stereo method which is optimized for reconstructing surface detail and exploits the high image resolutions of current digital cameras. Our method occupies a middle ground between stereo algorithms focused on depth layering of cluttered scenes and multi-view "object reconstruction" approaches which require a higher view count. It is based on global non-linear optimization of continuous scene depth rather than discrete pixel disparities. We propose a mesh-based data-term for large images, and a smoothness term using robust error norms to allow detailed surface geometry. We show that the continuous optimization approach enables interesting extensions beyond the core algorithm: Firstly, with small changes to the data-term camera parameters instead of depth can be optimized in the same framework. Secondly, we argue that our approach is well suited for a semi-interactive reconstruction work-flow, for which we propose several tools.

1. Introduction

Binocular stereo algorithms compute a depth map from a pair of photographs or video frames. Driven by benchmarks used in the Computer Vision community [1] and applications like driver assistance, the leading algorithms are optimized for recovering depth-layers of cluttered scenes and precise object boundaries. Often, they use a finite set of disparities, and therefore the resulting depth maps show little surface detail. Operating on the pixel grid, many algorithms are limited to low image resolutions, especially if high quality global optimization methods are used. However, graphics applications like image-based modeling, photo relighting, or the fabrication of physical models with 3D printing require detailed surface geometry rather than depth layering of cluttered scenes. Surface meshes with impressive details can be computed with state of the art multi-view algorithms. Unsurprisingly, however, those rely heavily on the availability of a large number of views.

In this paper, we propose an algorithm between these poles—a binocular stereo method optimized for computing detailed surface geometry. It exploits image resolutions in the 10–20 megapixel range which is typical for today's digital cameras. Combined with self-calibration, our approach enables high quality "walk-along" stereo on pairs of casual images shot free-hand, a few footsteps apart.

Like many binocular stereo algorithms, we use global energy minimization in a "data-term/smoothness-term" framework. However, we deviate from typical binocular stereo schemes in several respects, using strategies more common in multi-view methods: We use a triangle mesh in the image plane to decouple the number of variables from the number of pixels, allowing us to exploit high image resolutions *and* global optimization. Similar to patch-based reconstruction and surface evolution methods, continuous depth parametrization is used instead of discrete disparities, which has several advantages: Most importantly, it does neither restrict surface detail nor impose a tradeoff between detail and computational complexity due to a larger label set. It also allows us to omit image rectification, which is beneficial for casual free-hand stereo where camera rotation between the shots induces strong distortions in the rectification. Finally, we use a versatile Gauss–Newton-type optimizer which is straightforward to implement on top of widely available numerical software libraries. These topics are addressed in the first part of the paper (Section 3).

In the second part we outline two extensions of the core method, demonstrating the versatility of the proposed approach. We show that with small modifications to the energy function we can optimize camera parameters instead of depth in the same framework, which is useful to correct errors in calibration (Section 4.1). Secondly, we describe a set of interactive tools which allow the user to influence or correct the reconstruction. User interaction in 3D reconstruction is a little-discussed topic, despite the great practical success of intelligent semi-interactive tools for

☆This article was recommended for publication by Bedrich Benes.

* Corresponding author.

E-mail addresses: dc.schneid@gmail.com (D.C. Blumenthal-Barby), peter.eisert@hhi.fraunhofer.de (P. Eisert).

[1] Fraunhofer Heinrich Hertz Institute, Image Processing Department, Einsteinufer 37, 10827 Berlin, Germany.

0097-8493/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cag.2013.12.001

many tasks in graphics, such as segmentation, image retargeting or camera tracking. We argue that our approach is well suited for an interactive "optimize–adjust–re-optimize" work-flow, similar to semi-automatic segmentation (Section 4.2).

Finally, we address initialization and convergence, and present results which we compare to other methods (Section 5).

2. Related work

2.1. Binocular stereo

A taxonomy and review of classic stereo algorithms is given by Scharstein and Szeliski [1]. In the following we outline the differences between the focus of typical binocular stereo methods and the scope of our algorithm. Most research in binocular stereo is aimed at the reconstruction of scenes with multiple objects and a complex cluttered depth structure, prototypically represented by the Middlebury stereo benchmarks [1,2]. This prevalent scene type influences the strategies employed by the algorithms: For example, color similarity is often used as an indicator for depth continuity [3–5]. Combining scene segmentation with stereo has proven to be highly effective, such that algorithms like [6] perform these tasks in a joint optimization.

For detailed reconstruction of objects, which is the focus of our work, these scene priors often do not apply. Color information or segmentation, for example, cannot be used for a detailed reconstruction of artifacts like the one shown in Fig. 1.

Most stereo algorithms assume image pairs to be rectified such that correspondences can be searched along scan lines. This enables them to represent depth by disparity, measured in pixels, rather than by distance in a world coordinate system. As the pixel grid is discrete, many algorithms operate on a finite set of disparity values and use discrete global optimization methods such as Graphcuts and its descendants [6], belief propagation [4,5] or other message passing methods [7] . While this yields impressive results for the scenes these algorithms aim for, it limits the amount of surface detail for object reconstruction. This can be relaxed to some degree by using subpixel disparities (e.g. [8]) at the cost of a larger label set, or by additional refinement steps after the reconstruction process (e.g. [9,10]). Our method, like many multi-view approaches, is based on continuous depth parametrization and continuous optimization, which has the advantage of not limiting detail.

2.2. Multi-view reconstruction

Detailed surface reconstruction is the domain of multi-view methods which compute a mesh or volume representation of the scene rather than a depth map. As a comprehensive review of this area is beyond scope, we focus this section mostly on the relation of our work to multi-view approaches which use depth-maps computed from a small number of images from similar viewpoints at an intermediate stage, merging them later to obtain the final mesh. Examples of these methods are [7,9–11].

The most significant conceptual difference is that multi-view methods, even if based on intermediate depth maps, rely on the availability of more than two views. The larger number of views can be exploited in several ways. Most importantly, it provides more image data to verify correspondence hypotheses based on photometric consistency. This makes the estimation of depth (and, sometimes, normals as in [12]) significantly more reliable than in the binocular case. Due to this increased matching robustness, many algorithms omit costly global smoothness terms which are used in binocular stereo and compute stereo matches locally [9–12]. Smoothness is enforced later in outlier filtering, meshing or refinement stages. On the opposite side, surface evolution approaches like [13,14] use a single continuous global optimization on a highly sophisticated error functional that models surface visibility over multiple views in a mathematically precise manner. These methods, however, are quite involved with respect to numerics.

Some methods (e.g. [9]) restrict depth computation to binocular view pairs. While these approaches do not exploit the larger view count in the matching stage of the algorithm, the redundant scene coverage provided by multiple views allows them to be very strict in filtering out potential mismatches without risking holes in the overall reconstruction. In contrast, optimization-based binocular algorithms such as ours employ smoothness terms to propagate information into areas of low matching confidence. Many multi-view methods, including, for example, [9,12], also use the visual hull of the object as a constraint or to filter outliers. This is not possible in the binocular case.

We conclude this section by pointing out three specific related works. Firstly, an interesting but specialized stereo approach is used by Beeler et al. [10] for face reconstruction. They describe a complex iterative process of strictly local matching, filtering, and refinement steps on an image pyramid which produces detailed depth maps of the face in the binocular stereo stage. The impressive results of their overall system are based on multi-view data and on a face-specific approach to detail enhancement.

Secondly, there is an interesting connection of our work to Patch-Bases Multiview Stereo (PMVS) [12], which is still one of the most successful multi-view approaches. PMVS uses a three stage approach. First, corner points are matched across images to obtain an initial sparse point cloud. Second, depth together with surface normals is computed using nonlinear optimization on isolated patches, using the detected points as seed locations. Then, a closed surface is computed, which is refined by repeatedly optimizing patches located at the surface′s vertices, smoothness, and consistency with the scene silhouettes. Our method shares the initialization stage, where we use SIFT [15] instead of Harris corners. Then we directly optimize the reprojection of a connected triangle mesh, treating the mesh triangles as patches, where the connectivity induced by the mesh enables us to use small triangles. We also employ smoothness terms, which are critical to our approach due to the low number of views.

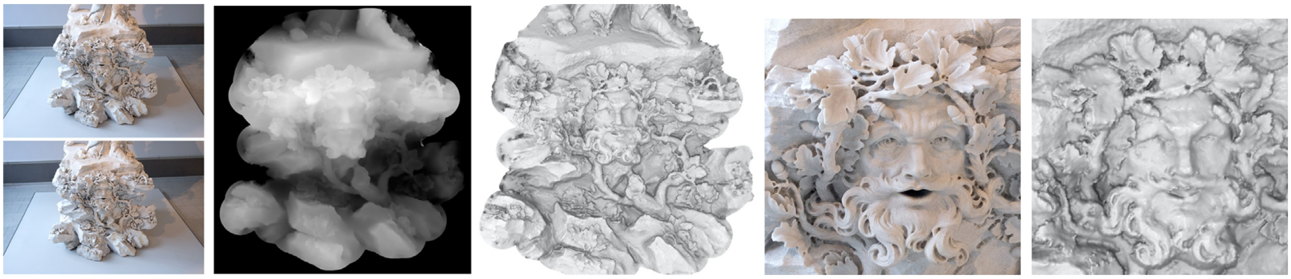

Fig. 1. Left to right: (1) Complete set of input images, 12 megapixels, shot freehand in a museum under available light; (2) depth map representation of our reconstruction; (3) reconstruction rendered as shaded mesh; (4,5) detail crop of the left view and the shaded reconstruction.