Computer Vision and Image Understanding 140 (2015) 127–143

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Hierarchical structure-and-motion recovery from uncalibrated images☆

Roberto Toldo[a,1], Riccardo Gherardi[a,2], Michela Farenzena[a,3], Andrea Fusiello[b,*]

[a] Dipartimento di Informatica, Università di Verona, Strada Le Grazie, 15 - 37134 Verona, Italy
[b] Dipartimento di Ingegneria Elettrica, Gestionale e Meccanica, University of Udine, Via Delle Scienze, Udine 208 - 33100, Italy

ARTICLE INFO

Article history:
Received 11 September 2014
Accepted 30 May 2015
Available online 6 June 2015

Keywords:
Structure and motion
Image orientation
Bundle adjustment
Autocalibration
3D

ABSTRACT

This paper addresses the structure-and-motion problem, that requires to find camera motion and 3D structure from point matches. A new pipeline, dubbed SAMANTHA, is presented, that departs from the prevailing sequential paradigm and embraces instead a hierarchical approach. This method has several advantages, like a provably lower computational complexity, which is necessary to achieve true scalability, and better error containment, leading to more stability and less drift. Moreover, a practical autocalibration procedure allows to process images without ancillary information. Experiments with real data assess the accuracy and the computational efficiency of the method.

1. Introduction

The progress in three-dimensional (3D) modeling research has been rapid and hectic, fueled by recent breakthroughs in keypoint detection and matching, the advances in computational power of desktop and mobile devices, the advent of digital photography and the subsequent availability of large datasets of public images. Today, the goal of definitively bridging the gap between physical reality and the digital world seems within reach given the magnitude, breadth and scope of current 3D modeling systems.

Three dimensional modeling is the process of recovering the properties of the environment and optionally of the sensing instrument from a series of measures. This generic definition is wide enough to accommodate very diverse methodologies, such as time-of-flight laser scanning, photometric stereo or satellite triangulation. The structure-and-motion (a.k.a. structure-from-motion) field of research is concerned with the recovery of the three dimensional geometry of the scene (the structure) when observed through a moving camera (the motion). Sensor data is either a video or a set of exposures; additional information, such as the calibration parameters, can be used if available. This paper describes our contributions to the problem of structure-and-motion recovery from unordered, uncalibrated images i.e, the problem of building a three dimensional model of a scene given a set of exposures. The sought result (the "model") is generally a 3D point-cloud consisting of the points whose projection was identified and matched in the images and a set of camera matrices, identifying position and attitude of each image with respect to an arbitrary reference frame.

The main challenges to be solved are computational efficiency (in order to be able to deal with more and more images) and generality, i.e., the amount of ancillary information that is required.

To address the efficiency issue we propose to describe the entire structure-and-motion process as a binary tree (called *dendrogram*) constructed by agglomerative clustering over the set of images. Each leaf corresponds to a single image, while internal nodes represent partial models obtained by merging the left and right sub-nodes. Computation proceeds from bottom to top, starting from several seed couples and eventually reaching the root node, corresponding to the complete model. This scheme provably cuts the computational complexity by one order of magnitude (provided that the dendrogram is well balanced), and it is less sensitive to typical problems of sequential approaches, namely sensitivity to initialization [1] and drift [2]. It is also scalable and efficient, since it partitions the problem into smaller instances and combines them hierarchically, making it inherently parallelizable.

On the side of generality, our aim is to push the "pure" structure-and-motion technique as far as possible, to investigate what can be achieved without including any auxiliary information. Current structure-and-motion research has partly sidestepped this issue using ancillary data such as EXIF tags embedded in conventional image formats. Their presence or consistency, however, is not guaranteed. We describe our approach to autocalibration, which is the process of automatic estimation from images of the internal parameters of

☆ This paper has been recommended for acceptance by J.-O. Eklundh.

* Corresponding author.
E-mail addresses: roberto.toldo@3dflow.net (R. Toldo), riccardo.gherardi@crl.toshiba.co.uk (R. Gherardi), michela.farenzena@evsys.net (M. Farenzena), andrea.fusiello@uniud.it, andrea.fusiello@gmail.com (A. Fusiello).

[1] R.T. is now with 3Dflow s.r.l. Verona, Italy.
[2] R.G. is now with Toshiba Cambridge Research Laboratory.
[3] M.F. is now at is now with eVS s.r.l.

http://dx.doi.org/10.1016/j.cviu.2015.05.011
1077-3142/

the cameras that captured them, and we therefore demonstrate the first structure-and-motion pipeline capable of using unordered, uncalibrated images.

1.1. Structure-and-motion: related work

The main issue to address in structure-and-motion is the computational complexity, which is dominated by the bundle adjustment phase, followed by feature extraction and matching.

A class of solutions that have been proposed are the so-called partitioning methods [3]. They reduce the structure-and-motion problem into smaller and better conditioned subproblems which can be effectively optimized. Within this approach, two main strategies can be distinguished.

The first one is to tackle directly the bundle adjustment algorithm, exploiting its properties and regularities. The idea is to split the optimization problem into smaller, more tractable components. The subproblems can be selected analytically as in [4], where spectral partitioning has been applied to structure-and-motion, or they can emerge from the underlying 3D structure of the problem, as described in [5]. The computational gain of such methods is obtained by limiting the combinatorial explosion of the algorithm complexity as the number of images and points increases.

The second strategy is to select a subset of the input images and points that subsumes the entire solution. Hierarchical sub-sampling was pioneered by [3], using a balanced tree of trifocal tensors over a video sequence. The approach was subsequently refined by [6], adding heuristics for redundant frames suppression and tensor triplet selection. In [7] the sequence is divided into segments, which are resolved locally. They are subsequently merged hierarchically, eventually using a representative subset of the segment frames. A similar approach is followed in [8], focusing on obtaining a well behaved segment subdivision and on the robustness of the following merging step. The advantage of these methods over their sequential counterparts lies in the fact that they improve error distribution on the entire dataset and bridge over degenerate configurations. In any case, they work for video sequences, so they cannot be applied to unordered, sparse images. The approach of [9] works with sparse datasets and is based on selecting a subset of images whose model provably approximates the one obtained using the entire set. This considerably lowers the computational requirements by controllably removing redundancy from the dataset. Even in this case, however, the images selected are processed incrementally. Moreover, this method does not avoid computing the epipolar geometry between all pairs of images.

Within the solutions aimed at reducing the impact of the bundle adjustment phase, hierarchical approaches include [10–12] and this paper. The first can be considered as the first paper where the idea has been set forth: a spanning tree is built to establish in which order the images must be processed. After that, however, the images are processed in a standard incremental way. The approach described in [11] is based on recursive partitioning of the problem into fully-constrained sub-problems, exploiting the bipartite structure of the visibility graph. The partitioning operates on the problem variables, whereas our approach works on the input images.

Orthogonally to the aforementioned approaches, a solution to the computational complexity of structure-and-motion is to throw additional computational power at the problem [13]. Within such a framework, the former algorithmic challenges are substituted by load balancing and subdivision of tasks. Such a direction of research strongly suggest that the current monolithic pipelines should be modified to accommodate ways to parallelize and optimally split the workflow of structure-and-motion tasks. In [14] image selection (via clustering) is combined with a highly parallel implementation that exploits graphic processors and multi-core architectures.

The impact of the bundle adjustment phase can also be reduced by adopting a different paradigm in which *first* the motion is recovered and *then* the structure is computed. All these methods start from the relative orientation of a subset of camera pairs (or triplets), computed from point correspondences, then solve for the absolute orientation of all the cameras (*globally*), reconstruct 3D points by intersection, and finally run a single bundle adjustment to refine the reconstruction. Camera internal parameters are required.

The method described in [15] solves a homogeneous linear system based on a novel decomposition of the essential matrix that involves the absolute parameters only. In [16] nonlinear optimization is performed to recover camera translations given a network of both noisy relative translation directions and 3D point observations. This step is preceded by outlier removal among relative translations by solving simpler low-dimensional subproblems. The authors of [17] propose a discrete Markov random field formulation in combination with Levenberg–Marquardt minimization. This technique requires additional information as input, such as geotag locations and vanishing points. Other approaches (e.g. [18–20]) compute translations together with the structure, involving a significant number of unknowns. The method presented in [21] proposes a fast spectral solution by casting translation recovery in a graph embedding problem. Govindu in [22] derives a homogeneous linear system of equations in which the unknown epipolar scaling factors are eliminated by using cross products, and this solution is refined through iterative reweighted least squares. The authors of [23] propose a linear algorithm based on an approximate geometric error in camera triplets. Moulon et al. [24] extract accurate relative translations by using an a-contrario trifocal tensor estimation method, and then recover simultaneously camera positions and scaling factors by using an ℓ_∞-norm approach. Similar to [23], this method requires a graph covered by contiguous camera triplets. The authors of [25] propose a two-stage method in which relative translation directions are extracted from point correspondences by using robust subspace optimization, and then absolute translations are recovered through semidefinite programming.

Another relevant issue in structure-and-motion is the level of generality, i.e., the number of assumption that are made concerning the input images, or, equivalently the amount of extra information that is required in addition to pixel values. Existing pipelines either assume known internal parameters [15,19,24,26–29], or constant internal parameters [30], or rely on EXIF data plus external information (camera CCD dimensions) [31,32]. Methods working in large scale environments usually rely on a lot of additional information, such as camera calibration and GPS/INS navigation systems [2,33] or geotags [17].

1.2. Autocalibration: related work

Autocalibration (a.k.a. self-calibration) has generated a lot of theoretical interest since its introduction in the seminal paper by Maybank and Faugeras [34]. The attention created by the problem however is inherently practical, since it eliminates the need for off-line calibration and enables the use of content acquired in an uncontrolled setting. Modern computer vision has partly sidestepped the issue by using ancillary information, such as EXIF tags embedded in some image formats. Unfortunately it is not always guaranteed that such data will be present or consistent with its medium, and do not eliminate the need for reliable autocalibration procedures.

A great deal of published methods rely on equations involving the dual image of the absolute quadric (DIAQ), introduced by Triggs in [35]. Earlier approaches for variable focal lengths were based on linear, weighted systems [36,37], solved directly or iteratively [38]. Their reliability has been improved by more recent algorithms, such as [39], solving super-linear systems while directly forcing the positive definiteness of the DIAQ. Such enhancements were necessary because of the structural non-linearity of the task: for this reason the problem has also been approached using branch and bound schemes, based