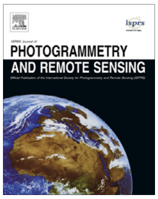
ISPRS Journal of Photogrammetry and Remote Sensing xxx (2018) xxx–xxx

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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



Prioritized multi-view stereo depth map generation using confidence prediction

Christian Mostegel [*], Friedrich Fraundorfer, Horst Bischof

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

ARTICLE INFO

Article history:
Received 17 September 2017
Received in revised form 22 December 2017
Accepted 21 March 2018
Available online xxxx

Keywords:
Multi-view stereo
Machine learning
Confidence measures
View prioritization
Image clustering
View cluster ranking

ABSTRACT

In this work, we propose a novel approach to prioritize the depth map computation of multi-view stereo (MVS) to obtain compact 3D point clouds of high quality and completeness at low computational cost. Our prioritization approach operates before the MVS algorithm is executed and consists of two steps. In the first step, we aim to find a good set of matching partners for each view. In the second step, we rank the resulting view clusters (i.e. key views with matching partners) according to their impact on the fulfillment of desired quality parameters such as completeness, ground resolution and accuracy. Additional to geometric analysis, we use a novel machine learning technique for training a confidence predictor. The purpose of this confidence predictor is to estimate the chances of a successful depth reconstruction for each pixel in each image for one specific MVS algorithm based on the RGB images and the image constellation. The underlying machine learning technique does not require any ground truth or manually labeled data for training, but instead adapts ideas from depth map fusion for providing a supervision signal. The trained confidence predictor allows us to evaluate the quality of image constellations and their potential impact to the resulting 3D reconstruction and thus builds a solid foundation for our prioritization approach. In our experiments, we are thus able to reach more than 70% of the maximal reachable quality fulfillment using only 5% of the available images as key views. For evaluating our approach within and across different domains, we use two completely different scenarios, i.e. cultural heritage preservation and reconstruction of single family houses.

1. Introduction

In this work, we aim to improve the efficiency of multi-view stereo (MVS) approaches based on depth maps. This type of approach is very popular (e.g. Goesele et al., 2007; Rothermel et al., 2012; Zheng et al., 2014; Galliani et al., 2015; Schönberger et al., 2016) as it is inherently parallelizable and delivers state-of-the-art results. One drawback of such approaches is that they typically generate one depth map per image in the dataset. For modern cameras, this means that 3D points in the order of 10^7 are created per image. With a few hundred images this leads to billions of points, that have to be stored, visualized and/or handled by subsequent processing steps such as depth map fusion or surface reconstruction. In this work, we propose a way to significantly reduce the number of generated points, while at the same time preserving the reconstruction accuracy and completeness to a large extent (see Fig. 1). This does not only speed up the depth map computation process, but also significantly reduces the load for all subsequent steps.

The key to our improved efficiency is a combination of geometric reasoning based on a preliminary scene reconstruction and machine learning to represent important properties of MVS that cannot be geometrically modeled. These unmodeled properties stem from the fact that MVS is an ill-posed task and to solve this task all MVS algorithms have to rely on a combination of some similarity measure (e.g. Census Transform or Normalized Cross Correlation) with a set of reasonable assumptions about the scene. The most popular assumptions include visual saliency, local planarity and a static scene. While these assumptions work well in many environments, there are also many environments where parts of the assumptions do not hold. Thus, MVS reconstructions very often contain outliers and/or fail to reconstruct certain objects completely.

[*] Corresponding author.
E-mail addresses: mostegel@icg.tugraz.at (C. Mostegel), fraundorfer@icg.tugraz.at (F. Fraundorfer), bischof@icg.tugraz.at (H. Bischof).

https://doi.org/10.1016/j.isprsjprs.2018.03.022
0924-2716/© 2018 International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS). Published by Elsevier B.V. All rights reserved.

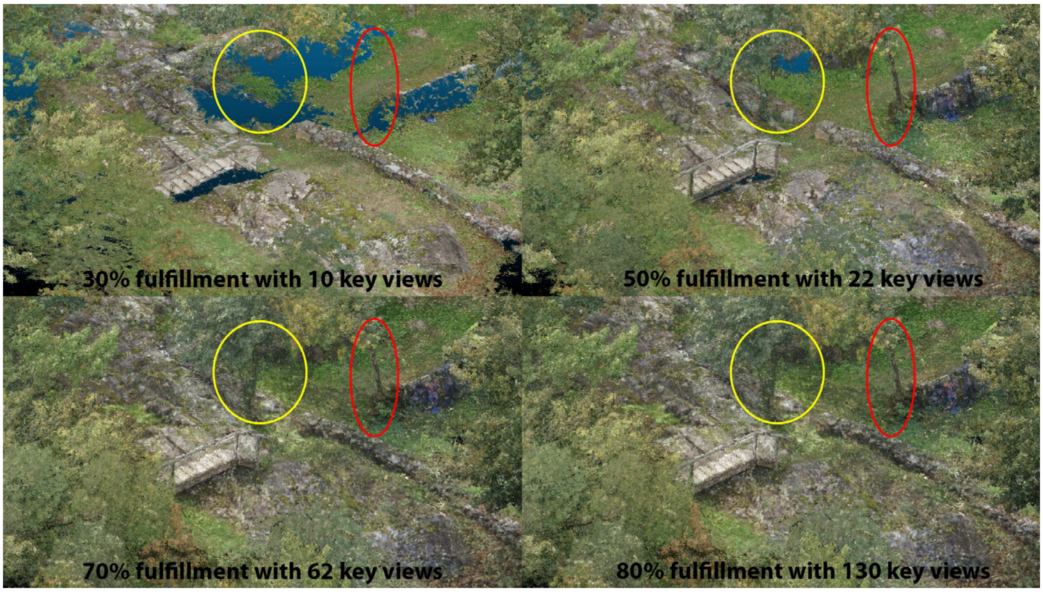

30% fulfillment with 10 key views
50% fulfillment with 22 key views
70% fulfillment with 62 key views
80% fulfillment with 130 key views

Fig. 1. View cluster prioritization. Our approach allows us to prioritize/rank view clusters (i.e. key views with matching partners) such that a highly complete and accurate point cloud can be obtain with a very small fraction of the available images. Here, we show the point clouds from the raw depth maps of the view clusters (with 11 matching partners) ranked with our approach after reaching 30%, 50%, 70% and 80% of the maximal achievable quality fulfillment (i.e. completeness with respect to a desired ground sampling distance and accuracy of 1 cm). Already with 50% fulfillment and only 22 key views (i.e. 1.8% of the available images), most parts of this complex scene are already contained in the reconstruction (red ellipse) and only a small part is missing (yellow ellipse). With 70% fulfillment, even strongly occluded parts such as tree trunks (see ellipses) are contained in the point cloud, although this point cloud is computed from only 62 key views (i.e. 5% of the available images). For going from 70–80% fulfillment, the number of necessary key views already has to be more than doubled, however, the visual difference between those two point clouds is nearly imperceptible (see also the supplemented video). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

In this work, we use our unsupervised machine learning framework (Mostegel et al., 2016a,b) to predict these failures and help us to reduce the number of key views (i.e. the image for computing a depth map) and necessary matching partners. Our method consists of two main steps. First, we select suitable matching partners for each view. Second, we prioritize/rank the resulting view clusters (i.e. views with matching partners) depending on their impact on a quality fulfillment function. This quality fulfillment function respects important photogrammetric parameters, such a ground resolution and 3D uncertainty, together with the scene coverage. The confidence prediction supports this whole process and allows us to obtain this ranking without having to execute the actual MVS algorithm within the ranking procedure. We formulate this quality fulfillment function as monotone submodular function and optimize this function with our ranking procedure in a greedy fashion. Although the overall problem is still NP-hard (as it includes the NP-hard maximum coverage problem), this formulation gives us strong optimality guarantees in the function space (Nemhauser et al., 1978).

This formulation has many advantages. First of all, the computed quality fulfillment function provides the opportunity to decide how many view clusters are necessary to obtain a certain quality fulfillment level. Thus, the operator can either choose to reconstruct the *n* best view clusters and has a estimation of the expect level of fulfillment or can simply query how large *n* should be to reach a certain level. The second advantage is that the inherent parallelism of MVS based on depth maps is maintained as our ranking procedure happens before executing the MVS reconstruction step. Third, the overall efficiency of the MVS reconstruction step can be significantly improved without changing the MVS algorithm itself. Thus, we were able to obtain a quality fulfillment (i.e. completeness with respect to a desired resolution and accuracy) of 70% with only 5% of the available view clusters. This leads to a speed up factor of approximately 10 and a complexity/memory reduction factor of approximately 20 for the resulting point cloud without losing much information.

2. Related work

Our prioritization approach is related to two different research areas, which are namely: Matching partner selection and Next-Best View (NBV) planning. In the following, we discuss the relation to these two interwoven areas.

2.1. Matching partner selection

Most MVS approaches based on depth maps formulate some kind of heuristic to select the *k* best matching partners for each key view to increase the efficiency of MVS. The heuristics for matching partner selection strongly depend on how the images are acquired (structured versus unstructured) and the requirements of the MVS algorithm. If the images are acquired in a regular grid, the *k* closest images are a natural choice to maximize the completeness. For more unstructured settings, the connectivity in the sparse reconstruction (i.e. how many sparse 3D points are shared between two cameras) is typically a more reliable cue to determine if the dense MVS matching step will work or not. To avoid that images with insufficient parallax are chosen as matching partners, Goesele et al. (2007) combine the connectivity with geometric constraints in a greedy fashion. Their formulation down-weights connections (shared features) with a triangulation angle below 10° and dissimilar scale. Additionally to these two terms, Bailer et al. (2012) also add a coverage term, which favors connections that have not been covered by other selected images. Shen (2013) use a formulation without connectivity only based on the geometric constraints on the triangulation angle and the distance between images. For very small datasets where all images nearly see the same part of the scene (as in the DTU dataset Aanæs et al., 2016), also random selection of matching partners can lead to good results (Galliani et al., 2015). Of all formulations mentioned above, Bailer et al. (2012) seems to be the closest related formulation to our approach. Similar to their approach,

Please cite this article in press as: Mostegel, C., et al. Prioritized multi-view stereo depth map generation using confidence prediction. ISPRS J. Photogram. Remote Sensing (2018), https://doi.org/10.1016/j.isprsjprs.2018.03.022