Neurocomputing 178 (2016) 46–61

Contents lists available at ScienceDirect

Neurocomputing

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

Multi-view stereo via depth map fusion: A coordinate decent optimization method

Zhaoxin Li [a], Kuanquan Wang [a,*], Deyu Meng [b], Chao Xu [c]

[a] School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China
[b] Institute for Information and System Sciences, Xi'an Jiaotong University, Xi'an, China
[c] School of Computer Software, Tianjin University, Tianjin, China

ARTICLE INFO

Article history:
Received 12 February 2015
Received in revised form
26 July 2015
Accepted 8 September 2015
Available online 10 November 2015

Keywords:
Coordinate decent
Depth map fusion
Multi-view stereo
Weighted median filtering

ABSTRACT

Multi-view stereo (MVS) plays a critical role in many practically important vision applications. Among the existing MVS methods, one typical approach is to fuse the depth maps from different views via minimization of the energy functional. However, these methods usually have expensive computational cost and are inflexible for extending to large neighborhood, leading to long run time and reconstruction artifacts. In this work, we propose a simple, efficient and flexible depth-map-fusion-based MVS reconstruction method: CoD-Fusion. The core idea of the method is to minimize the anisotropic or isotropic TV+L_1 energy functional using the coordinate decent (CoD) algorithm. CoD performs TV+L_1 minimization via solving a serial of voxel-wise L_1 minimization sub-problems which can be efficiently solved using fast weighted median filtering (WMF). We then extend WMF to larger neighborhood to suppress reconstruction artifacts. The results of quantitative and qualitative evaluation validate the flexibility and efficiency of CoD-Fusion as a promising method for large scale MVS reconstruction.

1. Introduction

Multi-view stereo (MVS) is a classic while challenging topic in computer vision [1–6], and plays a critical role in many practically important vision applications, e.g., augmented reality [63], urban reconstruction [7], and object detection, tracking and recognition [64–66]. The main task of MVS is to reconstruct a 3D scene surface from multiple calibrated 2D images. Based on the cues adopted in MVS, the existing approaches can be classified into four categories: reprojection error-based [7–12], photo-consistency-based [13–20], depth-map-based [21–33], and feature-based methods [34–36]. Among them, depth-map-based methods can be conveniently integrated with the existing stereo matching methods, and thus are more scalable and efficient. When accurate stereo correspondence is obtained, the reconstruction will be perfect. However, due to the occlusion, inaccurate camera calibration and lack of texture, false correspondence usually is inevitable, leading to incomplete and inaccurate reconstruction. For better reconstruction, regularization terms usually should be included to represent the prior of the scene surfaces.

A number of volumetric-based methods have been proposed for depth map fusion by integrating stereo image cues with regularization term [21,24,29,30,32,33], which share several difficulties. First, the computational cost is high for reconstructing scene with deep concave regions. Although coarse-to-fine scheme can be used for acceleration, the protrusion of surface removed in the low-resolution optimization is hard to be recovered in high-resolution. Second, although large neighborhood carries more context information for suppressing noise and artifacts, most existing methods are not scalable in taking large neighborhood in optimization.

In this paper, we propose a fast and flexible depth-map-fusion-based MVS method, i.e., CoD-Fusion. Compared with existing volumetric-based depth map fusion methods, CoD-Fusion is more computationally efficient, and can be extended from small 6-neighborhood in classic anisotropic TV to a larger neighborhood (e.g., 26 or 124) for better denoising and removal of irrelevant background regions. The major contributions of the proposed method are two folds:

- The coordinate decent (CoD) method [62] is used to decompose the anisotropic TV+L_1 energy functional[1] into a serial of L_1

* Corresponding author. Tel.: +86 451 8641 2871.
E-mail address: wangkq@hit.edu.cn (K. Wang).

http://dx.doi.org/10.1016/j.neucom.2015.09.109
0925-2312/

[1] In many literatures on variational methods for MVS [8–10,12,29,30,32], "energy functional" have been used. The meaning of "functional" indicates that the input variable of the function is also a function.

minimization sub-problems for each voxel. Fast weighted median filtering can then be used to solve these sub-problems.
- An approximation for isotropic 3D TV is given and analyzed, which allows us to employ CoD-Fusion for isotropic TV+L_1 minimization.

Multiple public datasets and self-captured datasets are used to evaluate the effectiveness of the proposed method. In the Middlebury MVS benchmark, compared with all published methods, the proposed method achieves the tied for the first and tied for the third on the *dino ring* and *temple ring* datasets in terms of completeness, respectively.

The rest of the paper is organized as follows: the related work is discussed in Section 2, the proposed CoD-Fusion method is presented in Section 3, including the energy functional and optimization algorithm. Experimental results are analyzed in Section 4 and concluding remarks are provided in Section 5.

2. Related work

In volumetric-based depth map fusion, various methods have been developed to minimize the energy functional, such as Graph cut [37] and continuous convex optimization [38]. Both of these two kinds of methods are able to reach the global minimum of the energy functional. However, graph cut based methods are difficult in parallel computation and may introduce significant metrication errors due to discretization [39]. Comparably, the continuous convex optimization methods are more promising on these issues.

Zach et al. [29] proposed a truncated-signed-distance-field (TSDF)-based fusion method. They transfer the depth maps to a set of TSDFs and then use an isotropic total variation (TV)+L_1 energy functional to integrate them. The optimization of isotropic TV+L_1 can be implemented by alternatively solving two sub-problems: a Rudin–Osher–Fatemi (ROF) sub-problem and a point-wise scalar sub-problem. The ROF sub-problem is solved using Chambolle projection method [40] and the scalar sub-problem is solved using soft-thresholding. Gottfried et al. [30] extended [29] for more efficient GPU implementation, where improved primal-dual method [41] was adopted for the ROF sub-problem and a generalized soft-thresholding [42] was employed for the scalar sub-problem. To avoid the over-smoothing behavior caused by isotropic TV, Schroers et al. [32] adopted anisotropic smoothing to better control the smoothness with respect to local structures. Although much of improvement has been made, these continuous methods still have expensive computational cost and are inflexible for extending to large neighborhood, leading to long run time and reconstruction artifacts.

3. Methods

The pipeline of the proposed CoD-Fusion method is summarized in Fig. 1. Given input images with corresponding camera parameters, the adjacent image pairs are stereo-rectified and the stereo matching algorithm is performed to generate depth maps. Then, all the depth maps from multiple image views are integrated on the volume to compute the truncated signed distance function (TSDF). Finally, the TV+L_1 energy functional is solved using CoD.

3.1. Estimation of depth maps via stereo matching

Given input images together with cameras parameters (intrinsic parameters and camera poses), depth maps can be inferred by searching the correspondence between image pairs. For MVS, the observed images are usually captured by a movable camera with arbitrary camera poses. To accurately and efficiently estimate the depth, stereo rectification is essential which usually involves two steps: (i) select suitable neighboring images for each reference image to build stereo pairs; (ii) perform stereo rectification to simplify the correspondence search.

After stereo rectification, stereo matching is then utilized to infer depth information for each stereo pair. Therefore we can employ the well-studied stereo matching methods in binocular stereo vision [43–47] to improve the 3D reconstruction performance. For a rectified stereo pair, the pixels in the left image (reference image) are matched in the right images with some criteria. Under the assumption of Lambertian reflection model, the correspondence of pixels between left and right images should have maximum photo-consistency, and a prior is usually included to enforce the local smoothness. The result of correspondence search assigns a disparity for each pixel of reference image. To remove the occlusion part, a left-right check is performed to filter the inconsistent pixels. The result disparity maps are then mapped to real depth metrics according to camera intrinsic parameters. In the paper, we generate depth maps using an existed stereo matching algorithm, libelas [59]. For libelas, the correspondence is constructed based on the features after Sobel operation in a small window for each pixel. For pixel correspondence, the L_1 distance between two features is used as the measure criteria. The libelas constructs a generative model based on a set of robust seed points on high-textured regions, and uses the Delaunay triangulation to constrain the search range of correspondences for each pixel in an admissible range. As a result, the running time of libelas is irrelevant to the disparity range, making it more suitable to large-scale scenes.

The depth map from a reference image records incomplete depth information of the scene and may include noise. To integrate these noisy depths from each reference image view and build a

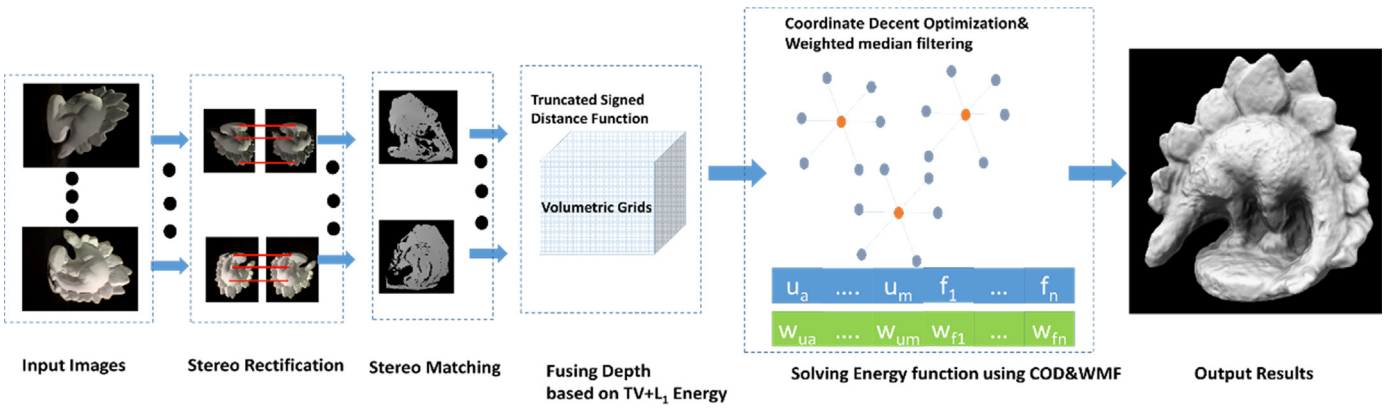

Fig. 1. Overview of the proposed CoD-Fusion framework.