Image and Vision Computing 32 (2014) 202–211

Contents lists available at ScienceDirect

Image and Vision Computing

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

Fast stereo matching using adaptive guided filtering☆

Qingqing Yang [a,b,c], Pan Ji [a,b], Dongxiao Li [a,b,*], Shaojun Yao [a,b], Ming Zhang [a,b]

[a] *Institute of Information and Communication Engineering, Zhejiang University, Hangzhou 310027, China*
[b] *Zhejiang Provincial Key Laboratory of Information Network Technology, Hangzhou 310027, China*
[c] *School of Information Science and Engineering, Ningbo Institute of Technology, Zhejiang University, Ningbo 315100, China*

A R T I C L E I N F O

Article history:
Received 16 August 2012
Received in revised form 10 October 2013
Accepted 9 January 2014

Keywords:
Stereo vision
Local method
Adaptive guided filtering
Parallel integral image
Weighted propagation

A B S T R A C T

Dense disparity map is required by many great 3D applications. In this paper, a novel stereo matching algorithm is presented. The main contributions of this work are three-fold. Firstly, a new cost-volume filtering method is proposed. A novel concept named "two-level local adaptation" is introduced to guide the proposed filtering approach. Secondly, a novel post-processing method is proposed to handle both occlusions and textureless regions. Thirdly, a parallel algorithm is proposed to efficiently calculate an integral image on GPU, and it accelerates the whole cost-volume filtering process. The overall stereo matching algorithm generates the state-of-the-art results. At the time of submission, it ranks the 10th among about 152 algorithms on the Middlebury stereo evaluation benchmark, and takes the 1st place in all local methods. By implementing the entire algorithm on the NVIDIA Tesla C2050 GPU, it can achieve over 30 million disparity estimates per second (MDE/s).

1. Introduction

Stereo matching is one of the key problems in computer vision, and it plays a significant role in many 3D applications. However, the binocular stereo correspondence problem itself is ill-posed. The captured stereo scenes are disturbed by noise from the surrounding environment, e.g. light variations and sensor noise in image formation. Some inherent ambiguities in stereo scenes, such as textureless region and occlusion, also make the problem much more challenging. Various methods have been proposed, but the problem is still not fully solved.

Many methods are proposed under certain assumptions and constraints, among which, the local smoothness assumption is the most widely used. According to the exhaustive review [1], most stereo matching algorithms can be categorized into two major classes: global methods and local methods. Global methods explicitly incorporate smoothness assumption into an energy function and formulate the problem in an energy-minimization framework, and the disparity map is determined by finding a solution that minimizes the global energy. Although global optimization techniques currently produce the best results, almost all the state-of-the-art global methods utilize image segmentation to ensure sharp depth edges [2]. In addition, iteration is inevitable in these algorithms, so they are relatively slow and do no scale well to high-resolution images.

In local methods, data term plays a dominant role, and cost aggregation is performed in local support windows. One implicit assumption embedded in cost aggregation is that, pixels in the support window are of the same disparity. An alternative view of cost aggregation in local methods is to regard it as filtering on the cost-volume. *Data costs* are locally smoothed by some specific filtering approaches [3,4]. Well performed edge preserving filters [5,6] (and their variations) can perform proper cost filtering as well as keep fine edges. It is believed that local stereo matching will play a more significant role in the future, since with the development of cameras, it will be much easier to obtain high-quality images with high resolution. Nevertheless, existing cost filtering based local algorithms usually utilize fixed-size kernel windows, which are not scalable to objects with different sizes in the same scene. They perform even worse in textureless regions due to the restriction of the fixed kernel size. Moreover, many cost filtering approaches are time-consuming [7].

In this paper, we propose a new cost-volume filtering method, whose weight kernel is a more general form of the one proposed in [6]. A novel concept named "two-level local adaptation" is introduced to guide the proposed filtering approach. Not only are the assigned support weights locally adaptive to the local patches, the sizes of local patches are also adjusted adaptively. A novel post-processing method is also proposed to handle occlusions and textureless regions. The proposed stereo matching algorithm ranks the 10th among about 152 algorithms on the Middlebury stereo evaluation benchmark, and takes the 1st place in all local methods. Moreover, a novel algorithm based on *parallel scan* is proposed to efficiently calculate an integral image

☆ This paper has been recommended for acceptance by Michael Goesele.

* Corresponding author at: Institute of Information and Communication Engineering, Zhejiang University, Hangzhou 310027, China. Tel./fax: +86 571 87952017.
E-mail addresses: qqyang@zju.edu.cn (Q. Yang), zhedajipan@zju.edu.cn (P. Ji), lidx@zju.edu.cn (D. Li), 06gksyysj@st.zju.edu.cn (S. Yao), zhangm@zju.edu.cn (M. Zhang).

0262-8856/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.imavis.2014.01.001

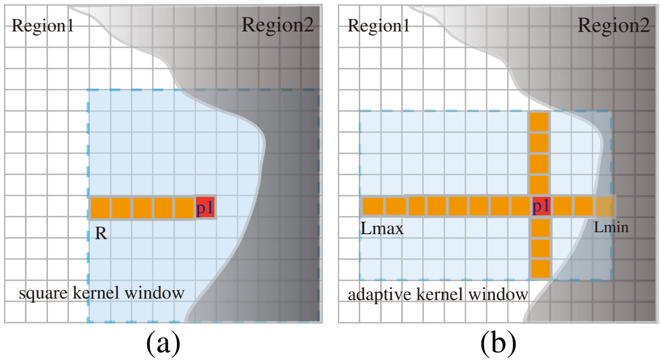

Fig. 1. Kernel windows of Eqs. (2) and (3). Pixels in shaded area represent outliers. (a) Square kernel window used in Eq. (2) and outliers presented in shaded area; (b) proposed adaptive kernel window used in Eq. (3) and outliers presented in shaded area, where fewer outliers are included.

on GPU. And the overall stereo matching algorithm can achieve over 30 million disparity estimates per second (MDE/s).[1]

The rest of this paper is organized as follows. In Section 2, related works on local stereo methods are reviewed. In Section 3, we present the proposed adaptive guided filtering approach and the overall stereo matching algorithm proposed in this paper is described in Section 4. At the end of each section, the computational complexity is analyzed. In Section 5, a hardware implementation of the cost filtering with CUDA is introduced. We report experimental results and give some discussions in Section 6, and conclusions are drawn in Section 7.

2. Related works

In local stereo matching algorithms, the disparity map is determined by selecting the value with the smallest matching cost from disparity candidates, which is also well known as winner-take-all (WTA) optimization. Thus, cost aggregation becomes the most important step in local stereo algorithms. However, it is not a trivial task as it appears to be. The most straightforward aggregation scheme is through simple low-pass filtering in the square support region. Filters with fixed-size convolution kernel, e.g. uniform (box filters), binomial or Gaussian, can be used. However, these simple methods result in poor disparity maps with fattened edges. Fixed-size kernel windows will easily overlap object boundaries, and matching costs in different regions are incorrectly aggregated. To overcome the edge fattening effect, various algorithms are proposed. Efforts on improving cost aggregation can be classified into two categories: variable support window (VSW) based approaches and adaptive support weight (ASW) based approaches.

Methods in the first category try to find support windows that fit the region size and/or shape, while preventing them from crossing object boundaries. In these methods, the piecewise smoothness assumption is utilized more explicitly, since all pixels in the support window are assumed to have the same weight. The simplest way is to filter the cost volume with a set of support windows of different sizes [1]. Its modification is shiftable windows schemes [8,9], whose window centers are anchored at different points in a set of fixed-size windows. A proper window with the most appropriate displacement is selected, which is useful at discontinuity jumps.

An alternative idea is to build a support window with variable size and/or shape tailored to the image content [10–12]. Variable window approach proposed by Veksler [12] performs well when only rectangular support windows are used. For every pixel in the reference image, a square support window is determined by minimizing the local window cost. Although the cost aggregation can be sped up by utilizing the integral image technique [13], the window adjustment step is quite time-consuming. Zhang et al. [14] proposed a fast algorithm in which non-regular support windows are used. A non-regular support window is first decomposed into horizontal and vertical slices, and cost aggregation is then performed in two directions separately.

One advantage of variable support window based approaches is that integral image technique can be utilized to speed up the aggregation procedure, which makes these cost aggregation schemes relatively efficient. However, a rectangular support window is inappropriate for pixels near object boundaries with arbitrary shapes, and a simple discontinuity reasoning method is also not strict enough to conserve edges.

Adaptive support weight based local methods, which are first introduced by Yoon and Kweon [7], adjust support weights for pixels in a local support window. In [7], the local smoothness assumption is constrained under the rules of similarity and proximity. Variations were also proposed to improve the accuracy. In [15], the authors explicitly deployed a smoothness constraint within local objects. Pixels inside the same segment in which the center pixel lies have the full weight 1, and the weights for pixels outside the segment are measured by the proximity term. Hosni et al. [16] proposed to compute the support weights by the geodesic distances, which enforce the foreground connectivity and prevent high weights from being wrongly assigned to background objects. Despite their outstanding performance, one common shortage is the high complexity. Some fast approximations [17,18] were proposed, but at the price of performance degradation. Recently, Yang [19] proposed a non-local approach, in which the cost values are aggregated adaptively on a minimum spanning tree. The support weight between two vertices is determined by their shortest distance on the tree. This literature reported better results than existing ASW based algorithms while offering extremely low computational complexity.

Few or no work tried to combine both VSW and ASW. The main reason may be: the combination of both VSW and ASW is computationally expensive with incommensurate result improvement. Readers are encouraged to refer to [20,21] for more performance study and evaluation of recent aggregation algorithms.

In recent years, cost aggregation is conducted by filtering on the cost-volume. In local stereo matching algorithms, edge preserving filters are frequently used, among which the bilateral filter [5] is the most widely used. But the brute force implementations are of high computational complexity when the kernel window is large. Many approximations [22–25] were proposed for fast implementation, but the accuracy are sacrificed due to quantization. Recently, Yang [26] proposed a recursive implementation of bilateral filtering by confining the range kernel and spatial kernel. The implementation demonstrated better accuracy than traditional bilateral filter with low complexity. But the performance of recursive bilateral filtering is still affected by the kernel confinements, this is reflected from the performance comparison with top-ranking local stereo methods. To overcome the shortages of bilateral filtering, He et al. [6] introduced the concept of guided image filtering, which has better behavior near edges. More importantly, it can be implemented exactly under linear complexity. Local methods that deployed it directly reported excellent results [3,4,27]. Based on the same filtering technique, the proposed method improves the performance by remodeling the weight kernel, following the novel concept "two-level local adaptation".

3. Adaptive guided filtering

3.1. Weight kernel remodeling

In stereo matching, the cost volume C is built in the cost computation stage. It is a three dimensional array which stores the matching costs for all possible disparity candidates. Let $C_{i,d}$ represent the cost value when pixel at $i = (x,y)$ is assigned to disparity d. Given the cost volume C and the reference image I, for each disparity candidate d, we apply

[1] The measurement of million disparity estimates per second (MDE/s) corresponds to the product of the number of pixels times the disparity range times the obtained framerate.