2011 International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission

Real-Time Accurate Stereo Matching using Modified Two-Pass Aggregation and Winner-Take-All Guided Dynamic Programming

Xuefeng Chang[1] Zhong Zhou[1] Liang Wang[2] Yingjie Shi[1] Qinping Zhao[1]
[1]State Key Laboratory of Virtual Reality Technology and Systems,Beihang University, Beijing 100191, China
cxf476@gmail.com, {zz,shiyj,zhaoqp}.vrlab.buaa.edu.cn
[2]University of Kentucky, Lexington, KY, USA

***Abstract*—This paper presents a real-time stereo algorithm that estimates scene depth information with high accuracy. Our algorithm consists of two novel components. First, we apply a modified two-pass aggregation to the adaptive cost aggregation process, use color similarity to calculate support weight, and introduce a credibility estimation mechanism to reduce accuracy loss during two-pass aggregation. Second, we present an amended scan-line optimization technique, which combines winner-take-all and dynamic programming. Our algorithm runs at 20 fps on 320×240 video with a disparity search range of 24. The experimental results are evaluated on the Middlebury benchmark data sets, showing that our method achieves the best reconstruction accuracy among all real-time stereo algorithms.**

***Keywords*-stereo; real-time; color similarity; dp;**

I. Introduction

Stereo matching estimates scene depth information from multiple images and it is one of the fundamental computer vision problems. Stereo research has recently experienced somewhat of a new era because of public available performance testing such as the most widely adopted Middlebury benchmark data set [1], which allows researchers to compare their algorithms against ground truth data and the state-of-the-art algorithms.

For the Middlebury evaluation system [1], [2], [3], the quality of stereo algorithms is evaluated by the reconstruction accuracy, that is, the average percentage of incorrect disparity estimates. Currently, stereo matching algorithm can be divided into three categories: good quality (error rate below 7%), average quality (error rate between 7% and 11%), and poor quality (error rate above 11%). Almost all top-performing stereo methods are based on the global Markov RandomField (MRF) formulation that solves stereo matching by minimizing certain energy function [4]. The lowest energy corresponding to the optimal disparity assignment can be approximately achieved using energy minimization techniques such as belief propagation [5] and graph-cut [6]. Recent development in MRF stereo has significantly advanced the state of the art in terms of accuracy. However, in terms of speed, global stereo algorithms typically take from several seconds to several minutes to compute a disparity map, limiting their applications to off-line processing. There are many interesting applications, such as robot navigation and augmented reality, in which high-quality depth estimation at video frame rate is required.

We in this paper present a real-time stereo algorithm, which is being evaluated as the top real-time performer on the Middlebury evaluation table. The algorithm is built upon the popular scanline-based optimization framework which is the basis for many real-time stereo algorithms on the Middlebury top-list. The algorithm presented in this paper introduces two novel ideas which significantly improve the reconstruction accuracy over existing scanline-based approaches. First, we amend the traditional adaptive color-weighted aggregation by using a modified two-pass aggregation. We compute support weight in a soft fashion using color similarity, analysis the possible accuracy loss generated by two-pass aggregation, and then we use a credibility estimation mechanism to solve this problem. Second, we improve the dynamic programming (DP) optimization technique by leveraging the winner-take-all (WTA) result as a prior, which improves the depth estimation at occlusion boundaries and better preserves depth discontinuities. Furthermore, we implement the cost aggregation scheme on the graphics hardware to facilitate real-time processing speed. The aggregated cost volume is transferred back to CPU for final disparity optimization using DP. In this way our approach makes use of both CPU and GPU in parallel and is able to produce a high-quality depth map at video rate. Experimental results evaluated using ground truth data demonstrate the effectiveness of our method.

The rest of this paper is organized as follows: After reviewing the related work in Section 2, we in Section 3 introduce our cost aggregation method using two-pass aggregation with credibility estimation and the WTA guided dynamic programming algorithm for disparity selection. Section 4 contains experimental results and we conclude in Section 5 with planned future work.

II. Related Work

In general, stereo algorithms can be categorized into local and global methods. Local algorithms are based on correlation and can have very fast implementation [7], [8], [9]. The central problem of local window-based algorithms is how to determine the size and shape of the aggregation window. That is, a window must be large enough to cover

978-0-7695-4369-7/11 $26.00 © 2011 IEEE
DOI 10.1109/3DIMPVT.2011.17

sufficient texture variation while small enough to avoid crossing depth discontinuities. This inherent ambiguity leads to problems such as noisy disparities in textureless region and fattened object boundaries. Global methods assume the disparity map is piecewise smooth except pixels near depth discontinuities and minimize certain cost functions. A popular global method is DP [10]. DP can offer optimized solution for independent scanlines in an efficient manner. Due to its one dimensional optimization solution and good speed performance, DP is the algorithm of choice for many real-time stereo applications [11], [12], [13]. Besides DP, advanced global optimization methods such as Belief Propagation (BP) and Graph Cut (GC) have attracted much attention [14], [5], [15], [16], [17], [18]. Although more accurate results are obtained, both BP and GC are typically computationally expensive. Yang *et al.* [19] implemented BP on GPU and thanks to the parallel processing capability of modern graphics hardware, real-time performance were achieved.

Our paper is also related to a local stereo method presented by Yoon and Kweon [20], in which a cost aggregation scheme that uses a fix-sized support window with per-pixel varying weight is introduced. The support weight is computed based on color similarity and geometric distance to the center pixel of interest. Very strong results are obtained using winner-take-all, without relying on global optimization. It is similar in spirit to many segmentation-based stereo algorithms, but it avoids image segmentation by using a continuous weighting function. Unfortunately, the aggregation process is very computationally expensive. As reported in the paper, it takes about one minute to process a 384×288 images on CPU.

Our approach is very much inspired by a recent paper by Wang *et al.* [12]. To achieve real-time performance, the authors propose a simplified approach. They integrate [20]'s adaptive weight scheme into a DP framework and the per-pixel matching cost is only aggregated in a one dimensional vertical window. This approach can achieve over 50 million disparity evaluations per second (MDE/s) when using the graphics hardware to accelerate the computation. Compared to traditional DP-based algorithms, their approach reduces the typical "streaking" artifacts without using multiple DP passes [21]. However, the quality of this algorithm is not quite satisfactory (error percentage 9.82%). Compared to [12], we adopt a two-pass aggregation strategy instead of the one-pass aggregation. And a layered weighting function is used to improve accuracy. We also use the results from winner-take-all to guide the DP optimization process. Our proposed algorithm significantly improves the accuracy over [12], especially in depth discontinuity regions.

III. Our Approach

In this section, we present our basic stereo model. Note that for notation clarity, our derivation will focus on rectified two-frame stereo images. However, it would be relatively easy to generalize our method to handle multi-view stereo. Given a stereo image pair $\{I, \overline{I}\}$, where $I, \overline{I}$ are the reference and target images respectively, the goal of stereo matching is to compute the dense disparity map of the reference view. Our stereo algorithm consists of two steps: 1) adaptive cost aggregation using two-pass aggregation with credibility estimation; 2) disparity selection by winner-take-all guided dynamic programming. In the following, we will present our algorithm in detail.

A. Weight computation by color similarity

We first compute the initial pixel-wise matching costs using the absolute difference method as

$$C(p,d) = \sum_{c \in \{r,g,b\}} |I_c(p) - \overline{I}_c(\overline{p})| \tag{1}$$

where $I_c, \overline{I}_c$ are the values of color channel c in the left and right images, $p = (x, y)$ is a pixel in the left image. Given a disparity hypothesis d, the corresponding pixel in the target image is $\overline{p} = (x + d, y)$. We define a square window W_p of predefined size centered at pixel p. For each pixel $q \in W_p$, we compute a weighting function $w(p, q)$ that encodes the likelihood that pixel q lies on the same surface with p. In [20], the adaptive support weight is computed based on the color similarity (Δc_{pq}) and geometric similarity (Δg_{pq}) between two pixels as:

$$w(p,q) = \exp(-(\frac{\Delta c_{pq}}{\gamma_c} + \frac{\Delta g_{pq}}{r_g})) \tag{2}$$

In equation (2) γ_c and γ_g are parameters that control grouping by similarity and proximity. However, we experimentally found that the geometric similarity has little effect on the final result if color information is used properly. Therefore, we ignore Δg_{pq} and define a different weighting function as:

$$w(p,q) = \exp(-\frac{\Delta c_{pq}}{\gamma_c}) \tag{3}$$

Figure 1 shows the support weights computed from equations (2) and (3) respectively with $\gamma_c = 36$ and $\gamma_g = 68$. As can be seen from Figure 1 (c) we know $q1$, $q2$ and q locate on different surfaces but Δg_{pq1} and Δg_{pq2} are almost equal and color information is distinctive for layer separation. Additionally, Δg_{pq3} is larger than Δg_{pq4}, and according to equation (2) $w(p, q3)$ should be smaller than $w(p, q4)$, as shown in Figure 1(b). However, p, $q3$ and $q4$ have similar depths which implies $q3$ and $q4$'s contribution to the p should be almost equal. In this example, including the geometric similarity even reduces the power of support weight computation.

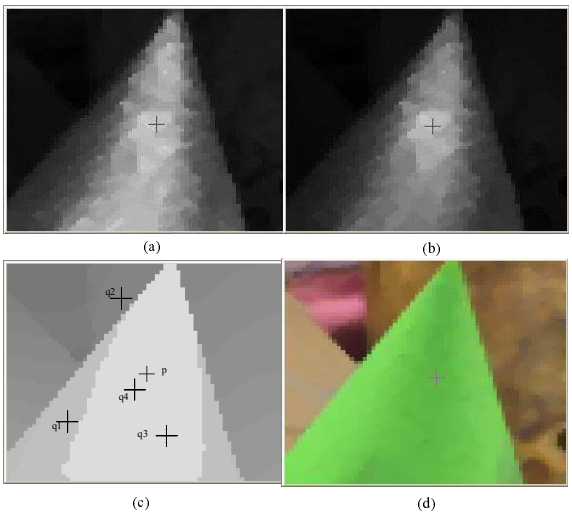


Figure 1. A comparison of weight masks computed with and without geometric similarity. (a) shows the weight mask by equation (3). (b) shows the weight mask by equation 2. (c) displays depth of pixels in this area. We adjust the scale factor so that the depth change is more clear. (d) shows the close-up views at a pixel location in the Cones image. Crosses mark pixels on the Cones image, and p is the center pixel.

B. *Two-pass aggregation based on credibility estimation*

After computing support weight by color similarity, we aggregate matching cost as:

$$C_w(p,d) = \frac{\sum_{q \in N_p} w(p,q) \cdot w(\overline{p},\overline{q}) \cdot C(q,d)}{\sum_{q \in N_p} w(p,q) \cdot w(\overline{p},\overline{q})} \quad (4)$$

where N_p is the set of all pixels covered by the support window. The complexity of this aggregation approach is $O(S^2)$, where S is the width of the support window. This high computational cost makes it less suitable to real-time application. To accelerate the aggregation step, we adopt a two-pass aggregation method proposed in [12] which approximates the full 2D aggregation with two separate 1D windows (one horizontal and one vertical). In essence, we first aggregate costs along the horizontal direction, and then sum the horizontal aggregated results along the vertical direction. This two-pass strategy reduces the aggregation complexity from $O(S^2)$ to $O(S)$. However, the weighted sum function in equation (4) is not separable in theory and this two-pass approach will inevitably introduce accuracy loss.

From Figure 2 we know $w(p,c)$ is not calculated directly during two-pass aggregation, but with the aid of c', that is:

$$w'(c,p) = w(c,c')w(c',p) \quad (5)$$

By combining equations (3) and (5) we can have a clear view of the difference between $w'(c,p)$ and $w(c,p)$, as shown in

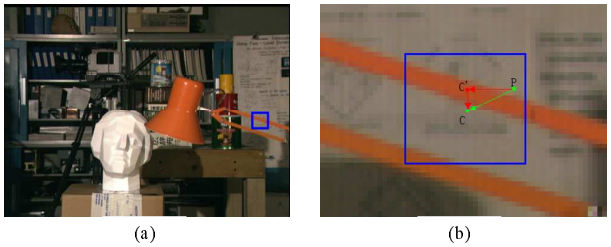


Figure 2. Support weight calculation in two-pass aggregation. Figure2 (b) is the close-up view at pixels in the red rectangle in Figure 2(a). The black rectangle in Figure 1(b) is the support window of pixel c. p is one pixel in the support window, c' is the center of all pixels which locate on the same line with p.

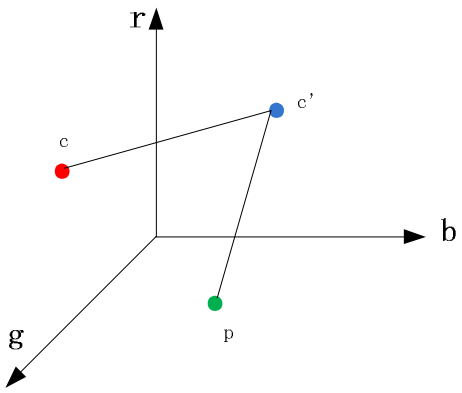


Figure 3. c, p and c' in color space, they construct a triangle in rgb color space.

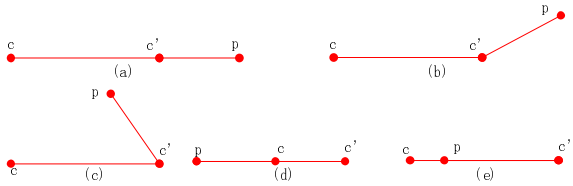


Figure 4. Several location examples of c, p and c' in color space. In (a) $\Delta c_{pc} = \Delta c_{pc'} + \Delta c_{cc'}$. In (b) and (c) $\Delta c_{pc} < \Delta c_{pc'} + \Delta c_{cc'}$ and in (d) and (e),$\Delta c_{pc} = |\Delta c_{pc'} - \Delta c_{cc'}|$.

equation (6).

$$w(c,p) = \exp(-\frac{\Delta c_{pc}}{r_c})$$
$$w'(c,p) = \exp(-\frac{\Delta c_{pc'} + \Delta c_{cc'}}{r_c}) \quad (6)$$

$\Delta c_{pc'}$ is the distance between p and c' in color space, $\Delta c_{cc'}$ is the distance between c and c'. Figure 3 shows the three pixels and their relative position in color space.

In Figure 3, $\Delta c_{pc'}$, $\Delta c_{cc'}$ and Δc_{pc} construct a triangle in color space. We know that in a triangle the sum of two sides is larger than the third one. So $w'(c,p)$ and $w(c,p)$ could not be equal unless c, p and c' are collinear, as shown in Figure 4(a).

From Figure 4, we know $\Delta c_{pc'} + \Delta c_{cc'} - \Delta c_{pc}$ varies from

0 to $2\min(\Delta c_{pc'}, \Delta c_{cc'})$, that is the larger the $\Delta c_{pc'}$ and $\Delta c_{cc'}$, the larger the accuracy loss. Therefore, we introduce a simple but effective aggregation credibility mechanism in two-pass aggregation to reduce the accuracy loss. We compute the support weight $w'(c,p)$ as well as its credibility according to equation (7).

$$R(c,p) = T(\exp(-\frac{\Delta C_{pc'}}{K}))T(\exp(-\frac{\Delta C_{cc'}}{K})) \quad (7)$$

Where $T(x)$ compute credibility of $w'(c,p)$ by judging the length of $\Delta c_{pc'}$ and $\Delta c_{cc'}$ .It is defined as:

$$T(x) = \begin{cases} 0, x < T_{w1} \\ 0.5, x < T_{w2} \\ 1, else \end{cases} \quad (8)$$

K, T_{w1}, T_{w2} are predefined parameters. If $\Delta c_{pc'}$ and $\Delta c_{cc'}$ are too large that $\exp(-\frac{\Delta_{c'p}}{K})$ and $\exp(-\frac{\Delta_{cc'}}{K})$ are less than T_{w1}, then we will assign the credibility to 0,to exclude w'(p, c) from aggregation. If $\Delta c_{pc'}$ and $\Delta c_{cc'}$ are of moderate length, then we will assign their credibility to 0.5. Otherwise their credibility is set to be 1.0, which means difference between $w'(p,c)$ and $w(p,c)$ is definitely small even in the worst condition. In horizontal aggregation, $\Delta c_{cc'}$ is inaccessible while in vertical aggregation the value of $\Delta c_{pc'}$ is unknown, so we judge $w'(c,p)$ as well as its credibility in two steps:

$$\begin{aligned} w_h(c,p) &= T(\exp(-\tfrac{\Delta c_{pc'}}{K}))w(c',p) \\ w_v(c,p) &= T(\exp(-\tfrac{\Delta c_{cc'}}{K}))w(c,c') \end{aligned} \quad (9)$$

In addition, in our modified two-pass approach the aggregated costs are calculated using:

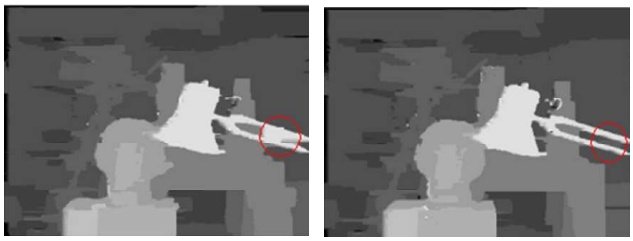
$$\begin{aligned} H_w(c',d) &= \frac{\sum_{p\in H_{c'}} w_h(c',p)w_h(\overline{c'},\overline{p})C(c',p)}{\sum_{p\in H_{c'}} w_h(c',p)w_h(\overline{c'},\overline{p})} \\ V_w(c,d) &= \frac{\sum_{v\in V_c} w_v(c,c')w_v(\overline{c},\overline{c'})C(c,c')}{\sum_{v\in V_c} w_v(c,c')w_v(\overline{c},\overline{c'})} \end{aligned} \quad (10)$$

Where $H_{c'}$ is the set of all pixels locate on the same line with c', V_c is the set of all pixels locate on the same column with c. In Figure 5 we show the depths maps resulting from our modified two-pass approach and compare it with the result from original two-pass approach.

C. Winner-take-all guided DP

We adopt an amended scan-line optimization technique which combines winner-take-all (WTA) and dynamic programming (DP). DP is one of the most widely used algorithms in stereo matching; it is formulated in an energy-minimization framework [4]. The objective is to find a disparity function d, which minimizes the following global energy.

$$E(d) = E_{data}(d) + E_{smooth}(d) \quad (11)$$



(a) (b)

Figure 5. A comparison of two-pass aggregation approach with and without credibility estimation. Figure(a) is generated without credibility estimation, Figure(b) is generated with credibility estimation. Note we have analyzed the possible error occurrence in red circle area at figure 1, and experiment result corresponds with our prediction.

Where $E_{data}(d)$ is the data term defined by equation (7), measures how well the disparity function d agrees with the input image pairs. $E_{smooth}(d)$ is the smoothness term which encodes the smoothness assumptions made by the algorithm. In our implementation, it is formulated as:

$$E_{smooth} = \sum_{x=1}^{Width} \gamma|d(x) - d(x-1)| \quad (12)$$

Here γ is a constant used to penalize depth discontinuities, and Width is the image width. We traverse the aggregated costs along each scanline from left to right. For pixel $p = (x,y)$ we need to traverse all the disparities $D(p')$ and calculate the minimum energy. The corresponding formula is:

$$\begin{aligned} F(p,d(p)) = C'(p,d(p)) + \\ \min_{d=d\min}^{d=d\max}\{F(p',d) + \gamma|(d(p) - d|\} \end{aligned} \quad (13)$$

Where $p' = (x-1, y)$. From equation (13) we know the complexity of this algorithm is $O(D^2)$, D is the disparity search range. This is not suitable for real-time system. To reduce the complexity of this approach, we import disparity smoothness constrain into this DP process. The disparity smoothness constraint means the disparity of a pixel is usually similar to disparities of the surrounding pixels. So for pixel p it is only necessary to consider $d(p) - 1$, $d(p)$, $d(p) + 1$ as the possible disparity candidates of $d(p')$. The modified formula is:

$$\begin{aligned} F(p,d(p)) = C'(p,d(p)) + \\ \min_{d\in[d(p)-1,d(p)+1]}\{F(p',d) + \gamma|(d(p) - d|\} \end{aligned} \quad (14)$$

The algorithm complexity of equation (14) is $O(D)$. But disparities computed by this simplified approach change slowly at depth discontinue areas, which may probably blur the occlusion borders. To avoid this "over-smooth" condition, we use the result of WTA to guide this scanline DP optimization process. Local WTA is simple and efficient but since WTA is based on greedy strategy and computes disparity of each pixel separately, disparity maps from WTA

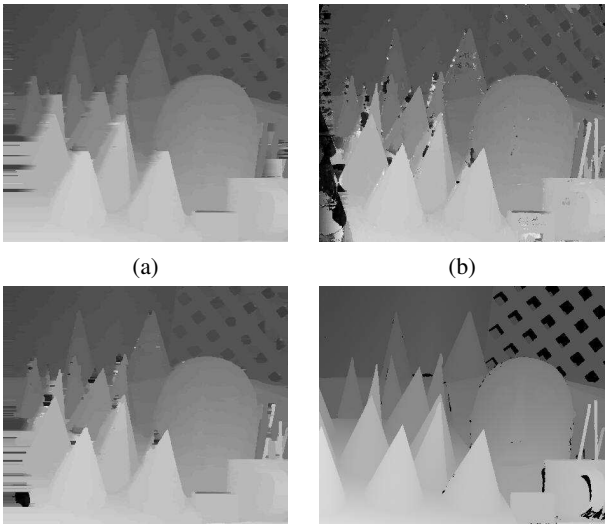


Figure 6. Comparison of different scanline optimization technology. The depth results are all aggregated according to equation 10. (a) is DP method(equation 14). (b) is generated by WTA. (c) is our combined approach(equation 15). (d) is the ground-truth.

are not as smooth as global methods. In our approach, we use WTA to provide an extra disparity candidate for DP. That is for $p = (x, y)$, we compute dx by equation (16), and take d_{x-1} as the fourth disparity candidate.

$$F(p, d(p)) = C'(p, d(p)) + \min_{d \in \{d(p) \pm 1, d(p), d_{x-1}\}} \{F(p', d) + \gamma |(d(p) - d|\} \quad (15)$$

$$d_{x-1} = \arg\min_{d} C'(p', d) \quad (16)$$

By combining WTA and scanline DP, our approach can better handle in depth discontinuity areas. For each column, we only need to compute d_{x-1} at the beginning, so algorithm complexity of this approach is still $O(D)$. Figure 6 shows the depth maps produced by our combined approach and compare it with the results of DP or WTA respectively. As can been seen our combined approach produces more accurate depth maps than using either DP or WTA alone. Indeed, our combined approach inherits the advantages of both DP and WTA. It generates smoother depth maps than WTA, and compared to DP it improves the depth estimation at occlusion boundaries therefore better preserves depth discontinuities.

IV. Experimental Results

We first evaluate the reconstruction quality of our proposed approach using the benchmark Middlebury stereo data set [1]. The second test is performed on a live captured dynamic video sequence in order to assess the performance of our stereo algorithm in real-time applications.

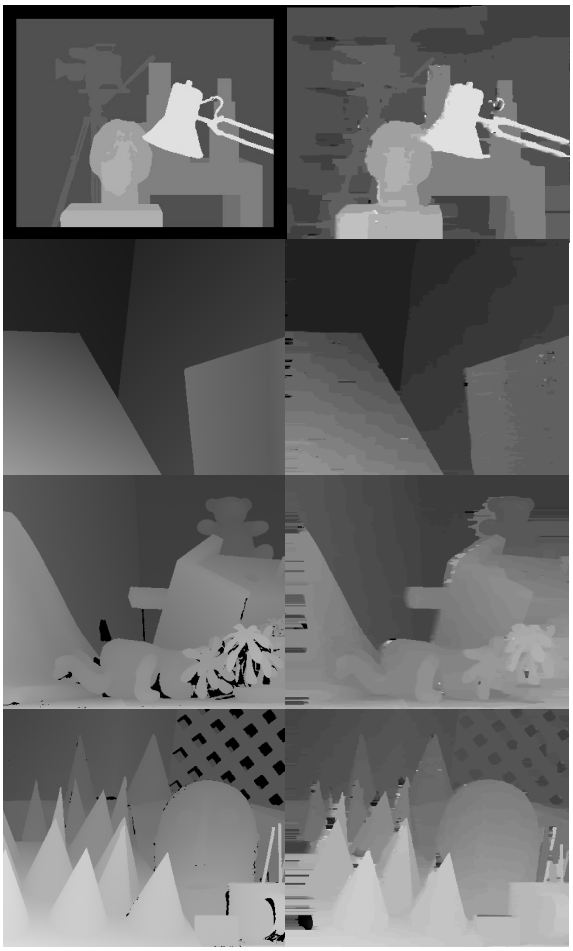


Figure 7. Resulting depth maps from Middlebury stereo data set. (left column) ground truths; (right column) our results.

A. Experiment on Middlebury stereo data set

We run the proposed algorithm on an Intel W3350 CPU with 3.0 GHZ and use a Geforce GTX 285 graphics card with 1GB memory manufactured by NVIDIA. Our cost aggregation is implemented using CUDA [25] on the GPU. All parameters are set to be identical across all experiments. They are: support window (35*35), two parameters in equation 5(K=2, γ_c=36), and the discontinuity cost (γ=3.25).The results from all four test sequences are shown in figure 7.

Table 1 shows quantitative results from the Middlebury evaluation table. Our approach currently ranks the first among all real-time stereo algorithms and ranks the 34th out of all 91 submissions (as of December 18th, 2010). Yoon's algorithm [20] is a slow approach, and ranks the 33rd. The average error rate of our approach (6.65%) is slightly lower than Yoon's method (6.67%) and our approach runs in real-time. RealTimeGPU represents the evaluation results of [12]'s algorithm, which uses one-pass vertical aggregation and traditional DP optimization. As we can be seen, even

Algorithm	Avg. Error	Avg. Rank	Tsukuba nonocc	Tsukuba all	Tsukuba disc	Venus nonocc	Venus all	Venus disc	Teddy nonocc	Teddy all	Teddy disc	Cones nonocc	Cones all	Cones disc
Ours	6.56$_1$	34.6$_2$	1.40$_3$	3.07$_3$	5.86$_1$	0.73$_3$	1.74$_3$	3.86$_2$	6.81$_2$	14.0$_4$	15.4$_1$	3.99$_2$	11.8$_4$	10.1$_2$
Adapt. Weight [20]	6.67$_2$	33.9$_1$	1.38$_2$	1.85$_2$	6.90$_3$	0.71$_2$	1.19$_2$	6.13$_3$	7.88$_5$	13.3$_2$	18.6$_5$	3.97$_1$	9.79$_2$	8.26$_1$
RealTimeABW [22]	7.90$_4$	41.2$_3$	1.26$_1$	1.67$_1$	6.83$_2$	0.33$_1$	0.65$_1$	3.56$_1$	10.7$_7$	18.3$_7$	23.3$_7$	4.81$_6$	12.6$_6$	10.7$_3$
RealTimeBP [19]	7.69$_3$	45.2$_4$	1.49$_4$	3.40$_4$	7.87$_4$	0.77$_4$	1.90$_4$	9.00$_4$	7.78$_4$	14.9$_6$	17.3$_2$	4.58$_4$	12.4$_5$	10.7$_3$
RealTimeVar [23]	9.05$_5$	53.1$_5$	3.36$_6$	5.48$_6$	16.8$_6$	1.15$_5$	2.35$_5$	12.8$_5$	6.18$_1$	13.1$_1$	17.3$_2$	4.66$_5$	11.7$_3$	13.7$_6$
RTCensus [24]	9.73$_6$	58.0$_5$	5.08$_7$	6.25$_7$	19.2$_7$	1.58$_6$	2.42$_6$	14.2$_6$	7.96$_6$	13.8$_3$	20.3$_6$	4.10$_3$	9.54$_1$	12.2$_5$
RealTime GPU [12]	9.82$_7$	59.1$_7$	2.05$_5$	4.22$_5$	10.6$_5$	1.92$_7$	2.98$_7$	20.3$_7$	7.23$_3$	14.4$_5$	17.6$_4$	6.41$_7$	13.7$_7$	16.5$_7$

Table I
PERFORMANCE COMPARISON OF THE PROPOSED METHOD WITH OTHER HIGH-QUALITY APPROACHES

Algorithm	Avg. Error	Avg. Rank	Tsukuba nonocc	Tsukuba all	Tsukuba disc	Venus nonocc	Venus all	Venus disc	Teddy nonocc	Teddy all	Teddy disc	Cones nonocc	Cones all	Cones disc
Without Credibility Estimation	8.52$_2$	51.3$_2$	3.10$_2$	4.70$_2$	13.3$_2$	1.89$_2$	2.86$_2$	11.0$_2$	7.00$_2$	14.0$_1$	16.9$_2$	4.22$_2$	12.0$_2$	11.2$_2$
With Credibility Estimation	6.56$_1$	34.6$_1$	1.40$_1$	3.07$_1$	5.86$_1$	0.73$_1$	1.74$_1$	3.86$_1$	6.81$_1$	14.0$_1$	15.4$_1$	3.99$_1$	11.8$_1$	10.1$_1$

Table II
PERFORMANCE COMPARISON OF DIFFERENT TWO-PASS AGGREGATION APPROACHES

Algorithm	Avg. Error	Avg. Rank	Tsukuba nonocc	Tsukuba all	Tsukuba disc	Venus nonocc	Venus all	Venus disc	Teddy nonocc	Teddy all	Teddy disc	Cones nonocc	Cones all	Cones disc
DP	7.27$_2$	42.0$_2$	1.54$_2$	3.30$_2$	6.68$_2$	0.79$_2$	1.95$_2$	5.15$_2$	6.89$_2$	14.2$_2$	15.6$_2$	5.02$_2$	13.1$_2$	13.0$_3$
WTA	9.32$_3$	61.2$_3$	3.20$_3$	5.21$_3$	7.04$_3$	2.49$_3$	3.93$_3$	9.66$_3$	10.3$_3$	18.0$_3$	18.5$_3$	5.92$_3$	15.4$_3$	12.2$_3$
DP+WTA	6.56$_1$	34.6$_1$	1.40$_1$	3.07$_1$	5.86$_1$	0.73$_1$	1.74$_1$	3.86$_1$	6.81$_1$	14.0$_1$	15.4$_1$	3.99$_1$	11.8$_1$	10.1$_1$

Table III
PERFORMANCE COMPARISON OF DIFFERENT DISPARITY COMPUTATION METHODS

with 1D vertical aggregation and scanline DP optimization, the accuracy of RealTimeGPU is not satisfactory (avg. error=9.82%).

To better illustrate the advantages of our credibility estimation mechanism, we list the results generated by different two-pass approach in table 2. The first row is two-pass aggregation without credibility estimation and the second row is our modified approach with credibility estimation. The improvement is obvious and after using credibility estimation the average error rate is reduce from 8.52% to 6.56%.

Table 3 shows results of different disparity computation methods. As we can see, the precision of DP method is better than WTA, and our combined approach generates better results than either using DP or WTA alone, especially near depth discontinuities regions.

B. Experiment on dynamic scene

Our algorithm is also tested on live videos captured by a bumblebee XB3 camera manufactured by Point Gray Research. The algorithm can achieve 20 fps when handing stereo image pairs of 320×240 pixels with 24 disparity levels. This is equivalent to 36.87 MDE/s [12]. The disparity maps generated from two stereo images captured in this experiment are shown in Figure 8. As can be seen our approach is able to produce detailed and accurate depth maps with clean object boundaries.

Figure 8. Two sample frames captured in our experiment and their corresponding disparity maps.

V. CONCLUSION

In this paper we present a high quality real-time stereo algorithm. We compute support weight for each pixel using color similarity and aggregate matching cost by a modified

two-pass aggregation, which is based on a new credibility estimation. We also improve the dynamic programming optimization technique by leveraging WTA disparity map as a prior guide, which improves the depth estimation at occlusion boundaries and better preserves the depth borders. We use CPU and GPU in parallel and produce high-quality depth map at video frame rate. We demonstrate the effectiveness of our algorithm by applying it to virtual reality applications that require accurate real-time depth estimates. Experimental results evaluated using ground truth data on Middlebury evaluation system show that our approach currently achieves the best reconstruction accuracy among all real-time stereo algorithms.

Acknowledgements: This work is supported by the Natural Science Foundation of China under Grant No.61073070, the National 973 Program of China under Grant No. 2009CB320805, the 2008 China Next Generation Internet Application Demonstration sub-Project under Grant No. CNGI2008-123, and Fundamental Research Funds for the Central Universities of China.

REFERENCES

[1] Vision.middlebury.edu/stereo/.

[2] D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2003.

[3] D. Scharstein and C. Pal, "Learning conditional random fields for stereo," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2007.

[4] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. of Computer Vision*, vol. 47, no. 1, pp. 7–42, May 2002.

[5] J. Sun, N.-N. Zheng, and H.-Y. Shum, "Stereo matching using belief propagation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 25, no. 7, pp. 787–800, July 2003.

[6] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 23, no. 11, pp. 1222–1239, November 2001.

[7] R. Yang and M. Pollefeys, "Multi-resolution real-time stereo on commodity graphics hardware," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2003.

[8] O. Veksler, "Fast variable window for stereo correspondence using integral images," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2003.

[9] L. Wang, M. Gong, M. Gong, and R. Yang, "How far can we go with local optimization in real-time stereo matching," in *Intl. Symposium on 3D Data Processing, Visualization and Transmission*, 2006, pp. 129–136.

[10] A. F. Bobick and S. S. Intille, "Large occlusion stereo," *Int. J. of Computer Vision*, vol. 33, no. 3, p. 181?200, 1999.

[11] M. Gong and Y.-H. Yang, "Fast unambiguous stereo matching using reliability-based dynamic programming," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 998–1003, June 2005.

[12] L. Wang, M. Liao, M. Gong, and R. Yang, "High-quality real-time stereo using adaptive cost aggregation and dynamic programming," in *Intl. Symposium on 3D Data Processing, Visualization and Transmission*, 2006, pp. 798–805.

[13] S. Forstmann, J. Ohya, Y. Kanou, A. Schmitt, and S. Thuering, "Real-time stereo by using dynamic programming," in *CVPR Workshop on Real-Time 3D Sensors and Their Use*, 2004.

[14] V. Kolmogorov and R. Zabih, "Multi-camera scene reconstruction via graph cuts," in *Proc. of Europ. Conf. on Computer Vision*, 2002.

[15] A. Klaus, M. Sormann, and K. Karne, "Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure," in *Proc. of Intl. Con. on Pattern Recognition*, 2006.

[16] Q. Yang, L. Wang, R. Yang, H. Stewenius, and D. Nister, "Stereo matching with color-weighted correlation, hierachical belief propagation and occlusion handling," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2006.

[17] O. J. Woodford, P. H. S. Torr, I. D. Reid, and A. W. Fitzgibbon, "Global stereo reconstruction under second order smoothness priors," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2008.

[18] L. Wang, H. Jin, and R. Yang, "Search space reduction for mrf stereo," in *Proc. of Europ. Conf. on Computer Vision*, 2008.

[19] Q. Yang, L. Wang, R. Yang, S. Wang, and D. Nister, "Real-time global stereo matching using hierarchical belief propagation," in *British Machine Vision Conf.*, 2006.

[20] K.-J. Yoon and I.-S. Kweon, "Locally adaptive support-weight approach for visual correspondence search," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2005, pp. 924–931.

[21] J. C. Kim, K. M. Lee, B. T. Choi, and S. U. Lee, "A dense stereo matchings uing two-pass dynamic programming with generalized ground control points," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2005.

[22] R.Gupta and S.-Y. Cho., "Real-time stereo matching using adaptive binary window," in *Intl. Symposium on 3D Data Processing, Visualization and Transmission*.

[23] S. Kosov, T.Thormahlen, and H.-P. Seidel., "Accurate real-time disparity estimation with variational methods," in *International Symposium on Visual Computing*.

[24] M. Humenberger, C. Zinner, M. Weber, W. Kubinger, , and M. Vincze, "A fast stereo matching algorithm suitable for embedded real-time systems," in *Computer Vision and Image Understanding*.

[25] Www.nvidia.com/page/technologies.html.