A Complementary Aggregation Approach for Local Stereo Matching Using Color and Correlation Cues

Ran Ju, Yang Yang, Xiangyang Xu, Chunrong Xia, and Gangshan Wu

State Key Laboratory for Novel Software Technology, Nanjing University, China
juran@smail.nju.edu.com, charlie.yang.nju@gmail.com,
xiangyang_xu@smail.nju.edu.cn, chunrong.xia@gmail.com, gswu@nju.edu.com

Abstract. Existed local stereo methods usually choose the support region for aggregation using correlation or color information independently. The correlation cue works well with high textures but has a poor performance near depth discontinuities while the color cue plays the complementary role. In this paper we first propose a new soft segmentation approach for correlation-based aggregation. Then we make a combination of the two cues and adopt the advantages of them to overcome the limitation of each other. Our approach performs a two stage aggregation based on correlation and color respectively. Each stage is operated by a bilateral filter on the cost volume. The combination is simple and effective, which enables our approach to achieve a better performance in both highly textured areas and depth discontinuities than existed methods. The experimental results conform to our expectation and do make improvements to state-of-the-art methods.

Keywords: stereo matching, cost aggregation, adaptive weight, bilateral filter, cost volume

1 Introduction

Stereo matching tries to find corresponding pixels in two or more images to reconstruct the 3D structure of a scene. The result is usually given in the form of disparity map, where the intensity of each pixel indicates the horizontal displacement to its corresponding pixel in the other view. The process of searching corresponding pixels may be confused by noise, poor texture etc. To overcome such difficulties there are generally two categories of solutions [1]: global and local methods. Global methods [2,3,4] use explicit smoothness assumptions to get global optimal solutions, which can make very accurate results but have a very high computational cost. On the contrary, local methods only rely on local information and thus work much faster than global methods. However, the accuracy and robustness can't fulfill the requirement of real applications and still need to be improved.

Generally, local methods need to take an aggregation, which sums or averages the matching cost over a local support region such that the ambiguity can be

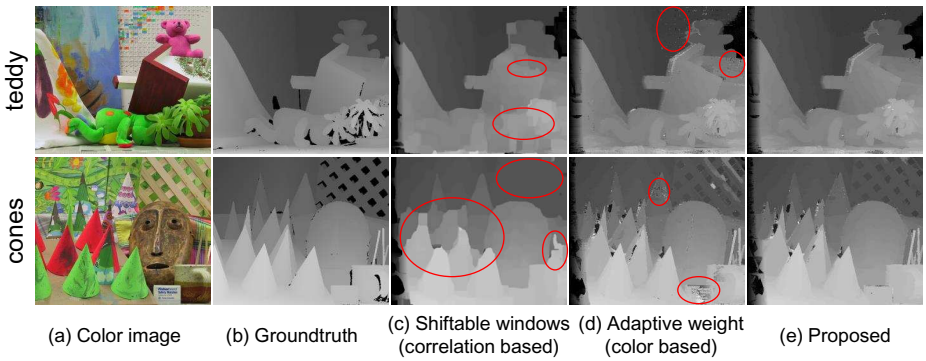


Fig. 1. Correlation-based and color-based methods. (a) (b) The two images with groundtruth disparity maps are from the Middlebury stereo benchmark [5]. (c) (d) The disadvantages in dealing with depth borders and high textures of correlation and color based methods respectively are emphasized using red ellipses. (e) Our approach makes a combination of the two approaches and thus achieves a satisfactory balance between smoothness in highly textured areas and edge-preserving property in depth discontinuities.

reduced effectively. However, how to choose the size and shape of the support region is a challenging problem and may influence the matching results directly. Correlation-based methods used to be the most successful solution. They search for the surrounding area with highest correlation on all disparity hypothesizes as the support region [1,6,7]. The basic prior is, a local set of pixels with similar disparities get the highest correlation value (or lowest matching cost) given the correct disparity hypothesis. These methods work well in highly textured areas but has a poor performance near low textured depth borders. Besides, existed correlation-based methods are constrained by the shape of rectangle windows, which may lead to an edge-fattening problem. Recently, color-based methods have made great advances in local stereo matching and thus represent state-of-the-art. They [8,9,10,11] choose the support region based on color similarity and space proximity. The basic prior is, pixels with similar colors tend to have similar disparities. These methods are good at handling depth discontinuities but easily confused in highly textured areas for insufficient aggregation support. The disadvantages of the two kinds of methods can be seen in Fig. 1. Obviously, correlation and color based methods play complementary roles and thus can be combined to cover the shortage of each other.

Our contribution lies in two aspects. First, we propose a new correlation-based aggregation method using soft segmentation to overcome the edge-fattening problem. Second, we make a combination of correlation and color cues by a two stage aggregation. The proposed approach is evaluated using the Middlebury dataset [5], computer-generated stereo images and photos taken in real life. The results show that our approach works well in both highly textured areas and depth discontinuities. The rest of this paper is organized as follows: First, we

Table 1. Typical local stereo methods

Algorithm	Basis	Segmentation
Shiftable window [1]	Correlation	Hard
Multiple windows [6]	Correlation	Hard
Variable window [7]	Correlation	Hard
Adaptive weight [8]	Color	Soft
Segment support [9]	Color	Soft
Geodesic support [10]	Color	Soft
Cost-volume filter [11]	Color	Soft
Proposed	Both	Soft

give a review of related works in Sect. 2. Then in Sect. 3 we describe the proposed approach. After that, we present our experiment and analysis in Sect. 4. At last, we make a conclusion and give some remarks on the future work in Sect. 5.

2 Related Works

Table 1 lists some typical local methods and their basic attributes. Adaptive window methods like [1,6,7] choose the sub window with best matching cost or some best surrounding windows around the central pixel as support region. However, besides the inherent disadvantage in low textured depth borders, these methods are constrained by the shape of rectangle windows, which is not suitable for irregular borders. The color-based methods use the local color distribution to guide aggregation. Typically, the adaptive weight method [8] employs the color similarity and geometric proximity to measure the contribution of each pixel while [9] and [10] take connectivity into consideration. To compute the aggregation cost more efficiently, some latest methods use edge-preserving filters on the cost volume. For example, Yang et al. [12] use a joint bilateral filter and Rhemann et al. [11] use the guided image filter instead. These methods work very well near depth discontinuities. However, they are easily confused by high textures for insufficient support. Generally speaking, color-based methods are unsuitable for the case that color distribution is inconsistent with disparity distribution. An extreme example is the random-dot stereogram. In this case color-based methods will get very poor results because color cue fails to work.

In the last column of Table 1, we list the segmentation type of the mentioned methods. Hard segmentation indicates the pixels are simply accepted or rejected for aggregation, while to soft segmentation the aggregation support of a pixel can be partially accepted. For two reasons soft segmentation is more suitable than hard segmentation. First, in depth borders the pixels may be mixed by foreground and background. Second, soft segmentation works in a probabilistic way which is more robust than hard constraints. The performances of the listed methods also show that soft segmentation is preferable and thus taken by our approach.

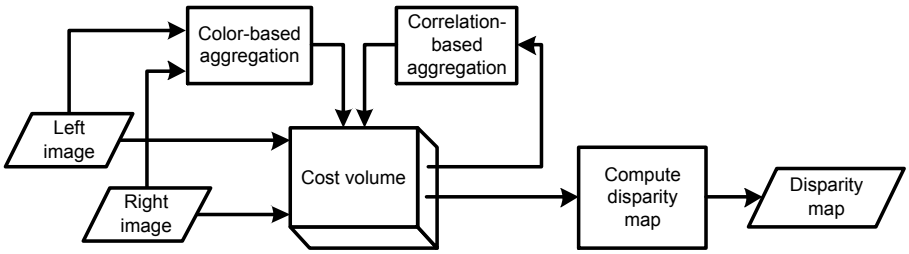


Fig. 2. Overview of the proposed approach.

3 Approach

We show an overview of our approach in Fig. 2. First we use the two color images (also called the left and right image) to compute the cost volume. Then we utilize an independent bilateral filter slice by slice on the cost volume as a correlation-based aggregation. Next we use a joint bilateral filter to perform a color-based aggregation. At last we generate the disparity map by trivially choosing for each pixel the disparity with minimum aggregated matching cost.

3.1 Construction of the Cost Volume

We first construct the cost volume by computing the matching cost of each pixel at all disparity hypothesizes. There have been a lot of studies on stereo matching cost [13,14] and Hirschmuller et al. made an excellent survey [15]. However up to now there still doesn't exist a perfect matching cost function. For example, NCC (normalized cross correlation) is insensitive to illumination changes but time-consuming, while AD (absolute difference) is simple and fast but easy to confused by illumination changes. In this paper, as we focus mainly on aggregation approaches, we choose the absolute difference of color intensities as our matching cost function for simplicity. Note for the other methods we also use AD to compute matching cost to give a fair comparison on aggregation. Suppose p is a pixel in the left image. Given a disparity hypothesis d, the corresponding pixel in the right image is $p - d$. The matching cost of pixel (p, d) is:

$$C(p,d) = \frac{1}{3} \sum_{c \in \{R,G,B\}} |I_l^c(p) - I_r^c(p-d)| \tag{1}$$

where $I_l^c(p)$ indicates the intensity of pixel p in the left image, and c is the color channel. As d varies from 0 to a maximum value d_{MAX}, we get a 3 dimensional cost volume, the intensity of each voxel is assigned with $C(p, d)$. In Fig. 3 we show a few slices of the cost volume for the image called "Teddy" (as shown in the first row of Fig. 1). The sequence number of each slice indicates the disparity value.

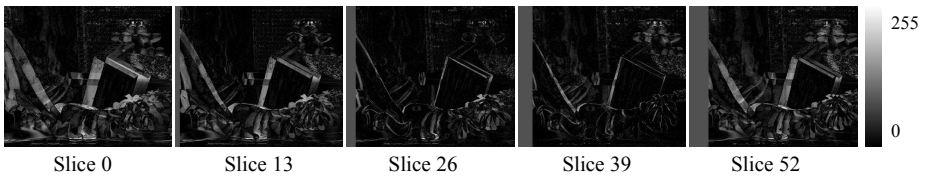


Fig. 3. A few slices of the cost volume for Teddy.

3.2 Correlation-based Aggregation

To improve the performance of classical adaptive window methods [1,6,7] near depth discontinuities, we use soft segmentation in aggregation. Generally, in the cost volume the pixels with lower matching costs are more likely to be correctly matched and thus should be assigned with higher weights for aggregation. Meanwhile, the pixels with higher matching costs should also be highlighted for rejection. The bilateral filter [16] achieved this goal very well. First, it adaptively aggregates the matching costs according to the cost distribution. Second, it performs a soft segmentation, which is more robust than hard segmentation and also handles depth discontinuities finely. Furthermore, the bilateral filter can also eliminate noises and strengthen edges. Thus we utilize a bilateral filter on the cost volume to break through the limitation of window shape constraints. The matching cost of each pixel (p, d) in the cost volume is updated as:

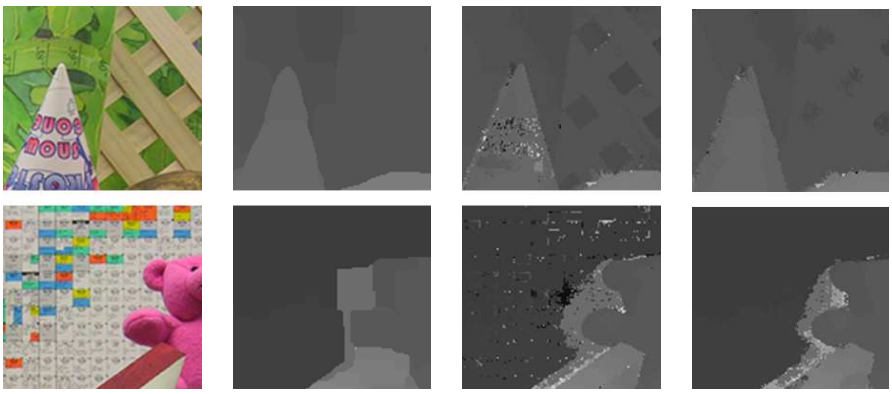
$$C_A(p,d) = \frac{\sum_{q \in N(p)} w_o(p,q) C(q,d)}{\sum_{q \in N(p)} w_o(p,q)} \tag{2}$$

where $N(p)$ indicates the neighboring pixels of p, usually a square window centered at p. $C(q,d)$ is the pixel-wise matching cost. $w_o(p,q)$ is the support weight of pixel q for p, which can be expressed as:

$$w_o(p,q) = \exp\left(-\left(\frac{|C(q,d) - C(p,d)|}{\gamma_o} + \frac{\|p-q\|_2}{\eta_o}\right)\right) \tag{3}$$

where $|C(q,d) - C(p,d)|$ indicates the intensity difference and $\|p-q\|_2$ is the Euclidean distance between the two pixels. γ_o and η_o are two parameters can be adjusted by users to control the power of color and space influence respectively.

To show the effect of the proposed correlation-based aggregation, we compute the intermediate disparity results by choosing the disparity with minimum aggregated matching cost for each pixel. We compare the results with the shiftable windows method [1] and adaptive support weight method [8], which represents the correlation-based and color-based method respectively. As shown in Fig. 4, adaptive window method generates smoother disparity maps but has serious edge-fattening problem. On the contrary, adaptive weight method has very accurate results near depth discontinuities but gets a poor performance in highly textured areas. The proposed approach shows more balanced results in both the two areas. However, there still exists edge-fattening problem due to the inherent disadvantage of correlation cue. Thus, we follow with a color-based aggregation.



(a) (b) (c) (d)

Fig. 4. The close-up disparity results. (a) Left color image. (b) Shiftable windows [1]. (c) Adaptive weight [8]. (d) Proposed method.

3.3 Color-based Aggregation

According to the adaptive support weight method [8], the aggregated matching cost is computed as:

$$C'_A(p,d) = \frac{\sum_{q\in N(p),q_d\in N(p_d)} w_c(p,q)w_c(p_d,q_d)C_A(q,d)}{\sum_{q\in N(p),q_d\in N(p_d)} w_c(p,q)w_c(p_d,q_d)} \tag{4}$$

where p and q are the pixels in the left image and p_d and q_d are the corresponding pixels in the right image. $w_c(p,q)$ is the support weight of pixel q for p. It is expressed as:

$$w_c(p,q) = \exp\left(-\left(\frac{\Delta_{pq}}{\gamma_c} + \frac{\|p-q\|_2}{\eta_c}\right)\right) \tag{5}$$

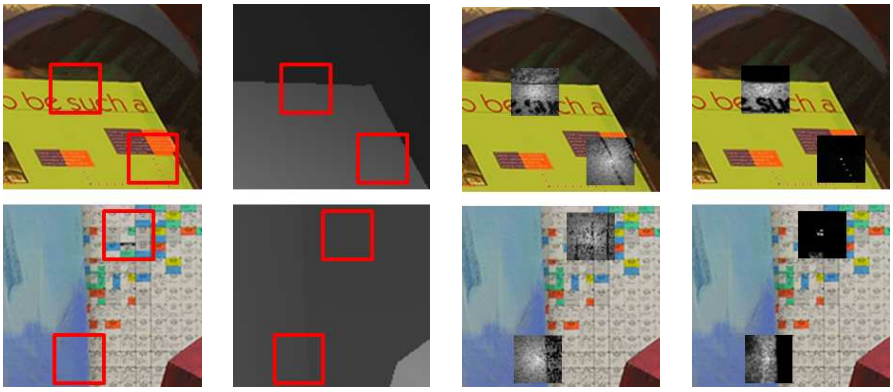
Δ_{pq} is the Euclidean distance between pixel p and q in the RGB color space. $\|p-q\|_2$ is the Euclidean distance between the two pixels.

We give a comparison of the two kinds of aggregation in Fig. 5. It is obviously that the correlation-based aggregation has more sufficient supports in highly textured areas than the color-based aggregation. Meanwhile near depth discontinuities the color-based aggregation is more accurate. As we make a combination, the goal of dealing with more complicated scenes is achieved by the complementary approach.

3.4 Disparity Computation and Time Complexity

We use the WTA (winner-take-all) strategy to compute final disparity map, where the disparity for each pixel is computed as:

$$d(p) = \arg\min_{0\le d\le d_{MAX}} C'_A(p,d) \tag{6}$$



(a) (b) (c) (d)

Fig. 5. Local supports of correlation-based and color-based methods. (a) Color image. (b) Truth depth map. (c) Local supports of correlation-based aggregation. (d) Local supports of color-based aggregation. The red windows in the first and second column indicate example aggregation windows. The gray windows in the third and fourth column indicate local support weights corresponding to the red windows. The pixels appear brighter indicate higher support weights and vice versa.

where d_{MAX} is maximum disparity hypothesis. It can be seen that the time complexity for each pixel is $O(d_{MAX}W)$, where W is the size of the local support window. The complexity is the same with [8] and the approach can be easily accelerated by parallel processing.

4 Experiments

4.1 Datasets and Parameter Settings

We test our algorithm using the Middlebury dataset [5], computer-generated stereo images and real world stereo photos. Four images used as benchmark in the Middlebury website are shown in the first row of Fig. 6. The benchmark uses percent of bad pixels as evaluation standard and we also take this way. To show the improvement of our approach we compare with state-of-the-art methods. The parameters in our algorithm are set as $\{\gamma_o, \eta_o, \gamma_c, \eta_c\} = \{10, 24, 15, 50\}$. The window size for correlation-based aggregation and color-based aggregation are 13×13 and 35×35 respectively. The parameters of the competitors are set according to their papers.

4.2 Experimental Results and Analysis

We choose four state-of-the-art methods for comparison, i.e. “AdaptWeight” [8], “SegmentSupport” [9],“GeoSup” [10] and “CostFilter” [11]. In their papers, the results have preformed different matching cost computation and post-processings.

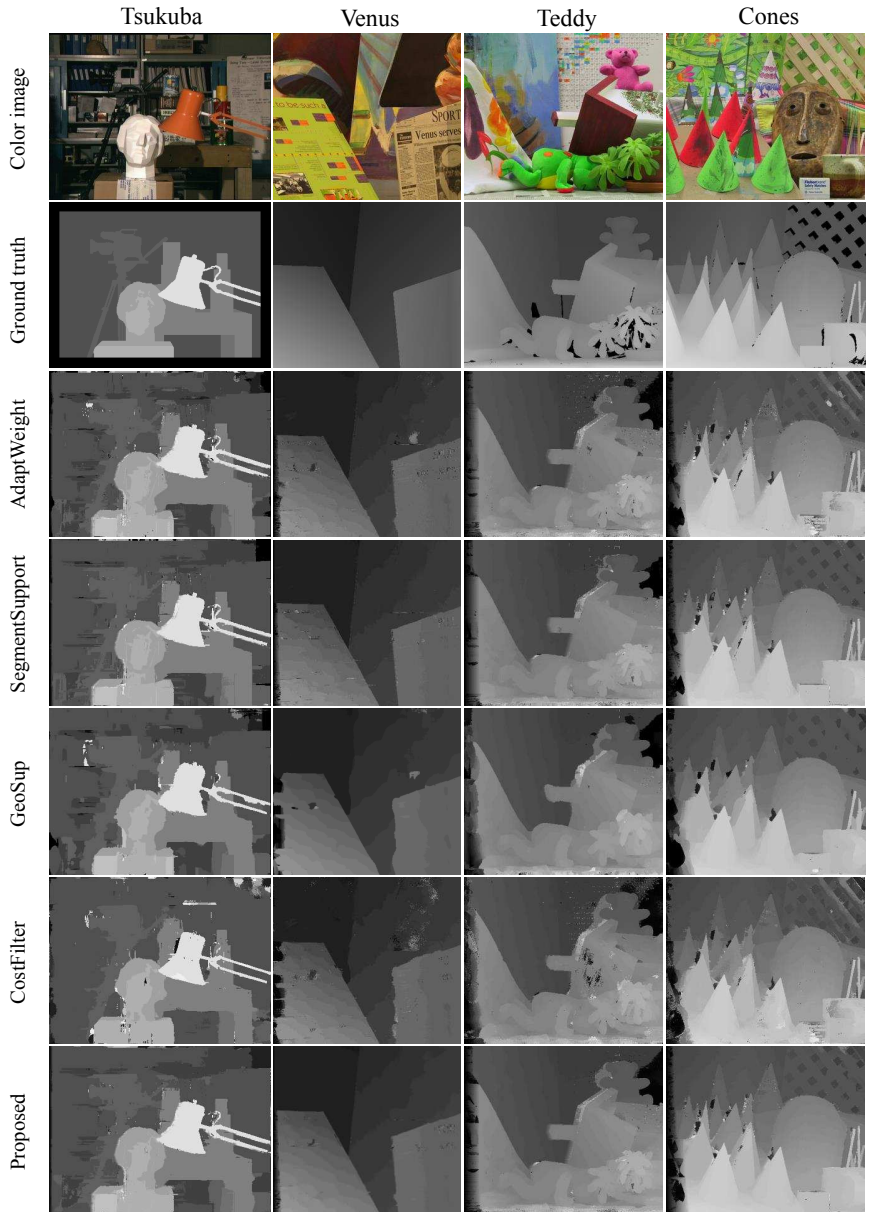


Fig. 6. Results comparison on the Middlebury dataset.

To give a fair comparison, we use TAD (truncated absolute difference) as matching cost function and compute the results without any post-processing for all methods. The results on the Middlebury dataset are shown in Fig. 6. It can be

seen that for the images with larger area of high textures like Venus and Cones our algorithm has an obvious improvement compared to the other competitors. AdaptWeight and CostFilter generate false matches in highly textured areas for insufficient aggregation support. GeoSup has edge-fattening problems, especially for the image named "Venus". SegmentSupport also has an excellent performance in highly textured areas because it aggregates more support by clustering small color pieces. However, it is still strongly dependent on color distribution while our algorithm uses correlation cue, which is more insensitive to color distribution and thus performs more robust as will be shown in the following.

Table 2. Bad pixels in non-occluded regions

Algorithm	Tsukuba	Venus	Teddy	Cones	Average
AdaptWeight [8]	2.82	2.76	12.1	9.66	6.84
SegmentSupport [9]	2.05	1.47	10.8	5.08	4.85
GeoSup [10]	3.16	2.74	11.6	5.11	5.65
CostFilter [11]	3.23	4.53	13.7	15.3	9.19
Proposed	2.01	1.25	11.1	4.93	4.82

Table 3. Bad pixels in discontinuous regions

Algorithm	Tsukuba	Venus	Teddy	Cones	Average
AdaptWeight [8]	7.38	10.4	21.4	15.9	13.77
SegmentSupport [9]	7.14	10.5	21.7	12.5	12.96
GeoSup [10]	10.1	17.3	22.9	12.3	15.65
CostFilter [11]	9.16	20.8	24.1	22.1	19.04
Proposed	7.07	5.86	21.2	11.4	11.38

We also use the evaluation standard by calculating the percent of bad pixels as the Middlebury benchmark does. In Table 2 and Table 3 we give the percent of bad pixels in non-occluded regions and near depth discontinuities respectively. The bad pixels in non-occluded regions give an overall evaluation on the disparity results except the occluded area because we didn't perform any post-processing to handle occlusion. The bad pixels in discontinuous regions give a targeted evaluation on the performance of handling depth discontinuities. As shown in the tables, in both the two areas our algorithm has a better performance than the other competitors in average. However, the Middlebury benchmark contains four images is still limited and may not reflect the overall performance of the local methods. And thus we give more experiments to show the power of our approach.

We give a few more examples in Fig. 7. As shown in the first row, we first use computer-generated random-dot stereo images to test the performance of handling the case that color distribution is independent of disparity distribution. The competitors all generate poor results because they all depend on the color prior, that is, pixels with similar colors tend to have similar disparities. Although

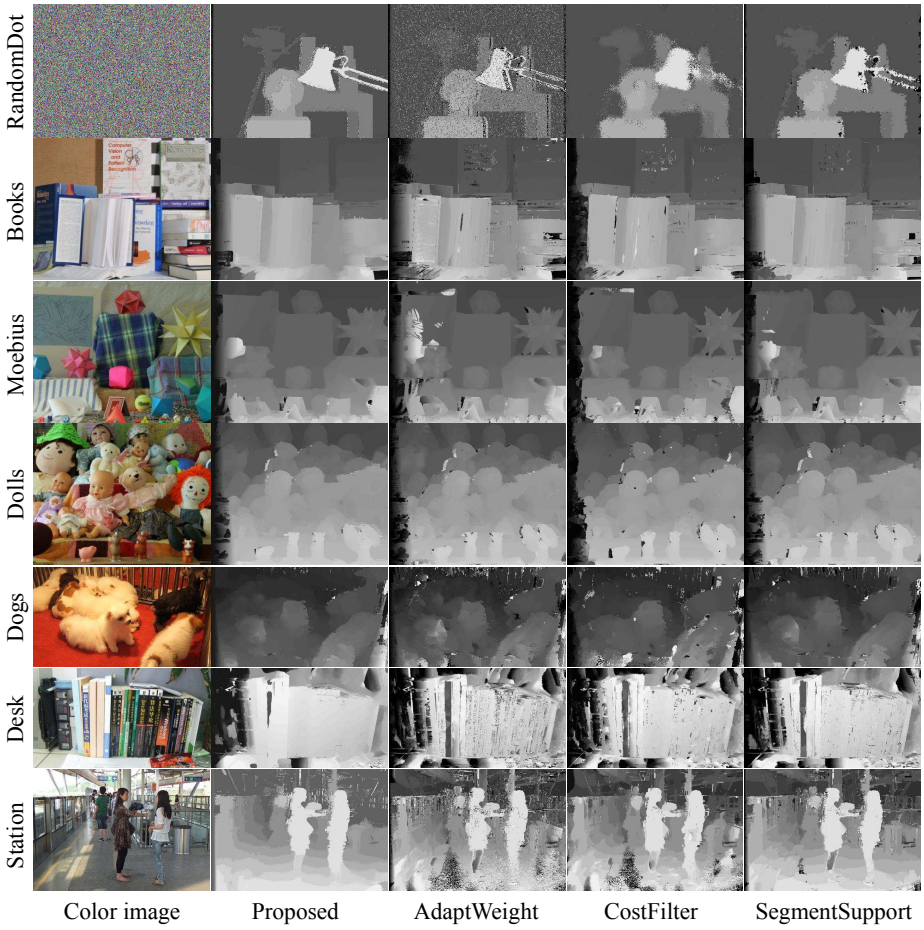


Fig. 7. More results tested on random-dot images (1st row), the Middlebury dataset (2nd to 4th row) and photos taken by a Fujifilm 3D camera (5th to 7th row). Color images (1st column), disparity maps generated by our approach (2nd column), AdaptWeight (3rd column), CostFilter (4th column) and SegmentSupport (5th column).

the prior works well in most cases, it is not a "one size fits all" standard. This can also be seen in the other results in Fig. 7. More or less, an image contains the parts do not meet the color prior. In comparison, our approach still works well even for random-dot stereo images because we combined the correlation cue for aggregation. The other images are from the Middlebury dataset (from the 2nd row to the 4th row) and taken by a FujiFilm W3 stereo camera (from the 5th row to the 7th row). It can be seen that our algorithm performs more robust than the other algorithms, especially for the real world stereo photos. This can be explained that the Middlebury images are taken under ideal illumination and

of high quality, also the scenes are artificially designed. In comparison, the real world photos contain more noise and the power of color prior may be reduced for natural scenes. However, our approach made a combination of color and correlation cues and thus works more robust.

5 Conclusion and Future Work

In this paper we have proposed a new aggregation approach using both correlation and color cues for local stereo matching. By integrating the advantages of the two cues, the proposed approach can easily deal with high textures and depth borders. The experimental results indicate our approach gives a better solution for local stereo matching no matter the color distribution of the image is consistent or inconsistent with the disparity distribution and performs more robust than state-of-the-art.

In the future, we plan to make efforts in the following two directions. First, we will incorporate our approach with more robust matching cost measurement [15] to deal with illumination changes and non-ideal Lambertian reflectance. Second, we are going to supply a high level description of the disparity map, e.g. we may apply the surface fitting technique [17] to get a few geometric primitives of the disparity map. We believe the disparity maps generated by our approach can be very useful for further applications like reconstruction, object segmentation, instance recognition and scene analysis etc.

Acknowledgments. This work is supported by the Natural Science Foundation of China (No. 61021062), the 863 Program of China (No. 2011AA01A202) and National Special Fund (No 2011ZX05035-004-004HZ).

References

1. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International journal of computer vision **47**(1) (2002) 7–42
2. Klaus, A., Sormann, M., Karner, K.: Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. In: Pattern Recognition, 2006. ICPR 2006. 18th International Conference on. Volume 3., IEEE (2006) 15–18
3. Yang, Q., Wang, L., Yang, R., Stewénius, H., Nistér, D.: Stereo matching with color-weighted correlation, hierarchical belief propagation, and occlusion handling. Pattern Analysis and Machine Intelligence, IEEE Transactions on **31**(3) (2009) 492–504
4. Wang, Z.F., Zheng, Z.G.: A region based stereo matching algorithm using cooperative optimization. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE (2008) 1–8
5. Scharstein, D., Szeliski, R.: Middlebury stereo vision page `http://vision.middlebury.edu/stereo`.

6. Hirschmüller, H., Innocent, P., Garibaldi, J.: Real-time correlation-based stereo vision with reduced border errors. International Journal of Computer Vision **47**(1) (2002) 229–246
7. Veksler, O.: Fast variable window for stereo correspondence using integral images. In: Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on. Volume 1., IEEE (2003) I–556
8. Yoon, K., Kweon, I.: Adaptive support-weight approach for correspondence search. Pattern Analysis and Machine Intelligence, IEEE Transactions on **28**(4) (2006) 650–656
9. Tombari, F., Mattoccia, S., Di Stefano, L.: Segmentation-based adaptive support for accurate stereo correspondence. Advances in Image and Video Technology (2007) 427–438
10. Hosni, A., Bleyer, M., Gelautz, M., Rhemann, C.: Local stereo matching using geodesic support weights. In: Image Processing (ICIP), 2009 16th IEEE International Conference on, IEEE (2009) 2093–2096
11. Rhemann, C., Hosni, A., Bleyer, M., Rother, C., Gelautz, M.: Fast cost-volume filtering for visual correspondence and beyond. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE (2011) 3017–3024
12. Yang, Q., Yang, R., Davis, J., Nistér, D.: Spatial-depth super resolution for range images. In: Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, IEEE (2007) 1–8
13. Birchfield, S., Tomasi, C.: A pixel dissimilarity measure that is insensitive to image sampling. Pattern Analysis and Machine Intelligence, IEEE Transactions on **20**(4) (1998) 401–406
14. Heo, Y.S., Lee, K.M., Lee, S.U.: Robust stereo matching using adaptive normalized cross-correlation. Pattern Analysis and Machine Intelligence, IEEE Transactions on **33**(4) (2011) 807–822
15. Hirschmuller, H., Scharstein, D.: Evaluation of stereo matching costs on images with radiometric differences. Pattern Analysis and Machine Intelligence, IEEE Transactions on **31**(9) (2009) 1582–1599
16. Tomasi, C., Manduchi, R.: Bilateral filtering for gray and color images. In: Computer Vision, 1998. Sixth International Conference on, IEEE (1998) 839–846
17. Cohen-Steiner, D., Alliez, P., Desbrun, M.: Variational shape approximation. In: ACM Transactions on Graphics (TOG). Volume 23., ACM (2004) 905–914