Full-image guided method based on confidence coefficient for fast stereo matching

Xiaodong Wang and Binming Xie
NARI Group Corporation Information System Integration Company; Nanjing, China

1 E-mail: Wangxiaodong5@sgepri.sgcc.com.cn

Abstract. Stereo matching is a challenging problem and is widely used in many applications, including autonomous navigation, 3D-reconstruction and etc. Therefore, we propose a new non-local method for stereo matching in this paper, which is mainly improved in two aspects. On the one hand, all the elements are employed during the step of cost aggregation aiming to handle the textureless regions. Besides, the calculation of aggregation weight is decreased greatly as a result of summing up the penalty terms to propagate weight. On the other hand, different aggregated costs are combined by defining the confidence coefficient instead of assigning weight in conventional methods. We evaluate the performance of our algorithm on the Middlebury dataset. Our model shows excellent results in particular for repetitive textures.

1. Introduction
Generation of dense disparity map from a pair of stereo images is a popular topic, which is also one of the most important and challenge subjects in computer vision. In general, we assume that the stereo pair has already been rectified such that corresponding points lie on the same horizontal scanlines in left and right views. As Scharstein recommends [1], most stereo methods consist of four steps: (1) matching cost computation; (2) cost aggregation; (3) disparity optimization; (4) disparity refinement. Generally speaking, the stereo algorithms can be divided into two categories: global and local algorithms. Their main differences lie in step (3). The global methods usually apply the hypothesis of explicit smoothness, and decide disparities of all the pixels simultaneously by minimizing the energy function. The local methods utilize “Winner Take All” (WTA) strategy pixel by pixel, directly choosing the disparity where the aggregated cost is minimum. Although global algorithms usually generate accurate disparity maps, they are relatively slow due to the inevitable iterations, and do not scale well to high-resolution images. Besides, it’s difficult to determine the optimal parameters sometimes. On the other hand, local methods perform well in real time implementation, and could produce good results on par with those generated by global methods with the continual improvement.

One of the most important steps in local stereo matching algorithms is cost aggregation [1]. The initial matching costs of each pixel at every possible disparities is obtained in the first step, which compose a three-dimensional data structure called cost volume. For a given pixel in reference image, the corresponding matching costs at every candidate disparities are shown in figure 1(a), which is irregular. Hence, it’s essential to exploit the information of neighborhood pixels. Typically, the conventional local methods sum up the weighted cost within a fixed size window, just like the process of filtering. As shown in figure 1(b), the curve is smoothed, and the disparity correspondence to the minimum cost is more evident. However, it’s well known that the filter usually fails in protecting object borders and tends to blur the depth boundaries. In addition, the other major problem lies in the
textureless regions, which is also called the aperture problem. Most of the background area in an image is occupied by weak textures or the repetitive textures, like the sky and grass. Therefore, the performance of current state-of-the-art matching algorithms handling the images with weak textures and so on we have mentioned above, still calls for improvement.

Aiming to the issues mentioned above, many edge-preserving filters [2, 3] were proposed to keep the object border distinct, which achieve better performance of local algorithms. Nevertheless, those methods still only exploit the features in local circumambient regions. Yang and etc. [4] proposed a full-image guided filtering, which was implemented on a four-connected grid of the guide image. However, the weight between two pixels was computed by consecutive multiplication along the path, where the amount of computation is enormous. On the other hand, the matching costs in texture less regions are usually equal or close to zero, which performs little contributions in local cost aggregation. Hence, with the WTA strategy, any candidate disparity may be selected for these pixels. Therefore, it’s necessary to propagate the information from textured regions.

In this paper, a novel non-local matching method is proposed. Since the effect of aggregation is similar with that of filtering, we firstly handle the raw images to obtain sharp boundaries. In regard to the regions with few texture, the aggregated costs of those are still not distinct. To this end, we introduce a horizontal-tree based non-local algorithm, which requires extremely low computational amount, to aggregate the initial costs. Besides, the confidence coefficient is proposed in the step of disparity computing to combine different methods of cost measuring. Finally, we test our algorithm on the Middlebury stereo matching platform [5], and make a comparison with other state-of-the-art methods, the detail of which is shown in result section.

2. Proposed method

2.1. Initial cost computing

Traditionally, the raw image pairs are directly used in the stereo matching or simply with the median filter to suppress noise. However, as stated in the first section, it tends to blur the object borders, which has a negative effect on the initial cost computing and aggregation in the region where depth is discontinuous. According to experiment results, the filter-based guidance image model proposed in [6], which considers the spatiality and the color difference effect, is rather effective for preserving edges and smoothing the homogeneous areas. Hence, we obtain the guide image on the basis of the model. In contrast to the traditional methods, we extract the gradient from the guide image instead of the raw image due to more distinct boundaries and less noises in guide image. Firstly, the matching cost based color-gradient is computed as:

![Figure 1](image_url)

**Figure 1.** Curves of initial cost and aggregated cost along disparity of one pixel. The cost aggregation makes the optimal disparity of the pixel clearer.
\[
C_{c-g}(p, p') = \alpha \cdot \min (\|r_p - I_p\|, r_{\text{color}}) + (1-\alpha) \cdot \min (\|\nabla I_p - \nabla I_p\|, r_{\text{grad}})
\]

(1)

where \(p\) and \(p'\) denote the pixels in reference image \(I\) and target image \(I'\) respectively. \(\|r_p - I_p\|\) represents the Euclidean distance between \(p\) and \(p'\) in the RGB color space. \(\|\nabla I_p - \nabla I_p\|\) represents the absolute difference between \(p\) and \(p'\) on gray-value gradients. In addition, \(\alpha\) balances the influence of color and gradient term, \(r_{\text{color}}\) and \(r_{\text{grad}}\) truncate the costs to limit the negative impact of outliers.

Moreover, Census transformation [7] has been reported performing better while coping with radiometric variations. Taking pixel \(p\) as an example, compute the census transform result, which is a string vector consisting of one and zero, within its neighboring support window \(w_p\).

\[
\text{cen}(p) = \otimes_{q \in w_p} c(p, q)
\]

(2)

where \(\otimes\) represents concatenation, and \(c(p, q)\) denotes a binary function that can be expressed as follows:

\[
c(p, q) = \begin{cases} 
1, & \text{if } I(p) > I(q) \\
0, & \text{otherwise}
\end{cases}
\]

(3)

Then, the initial census transform cost of pixel \(p\) and pixel \(p'\) is obtained.

\[
C_{\text{cen}}(p, p') = \text{cen}(p) \oplus \text{cen}(p')
\]

(4)

Here, denotes the Minkowski sum.

The combination of initial costs based color-gradient and census transform will be discussed in the step of disparity computing at length.

2.2. Non-local weighted aggregation

For the given reference image \(I\), it can be regarded as a connected, undirected graph \(G = (N, E)\), where \(N\) denotes the set of nodes, which correspond to the pixels in image \(I\). And each edge in \(E\) connects a pair of neighbored nodes. For an edge connecting node \(i\) and node \(j\), let the penalty term on the edge be:

\[
p(i, j) = \|I(i) - I(j)\|
\]

(5)

which is greatly inspired by the explicit assumption in local stereo algorithm. The greater intensity difference between two pixels is, the smaller possibility at the same disparity level is, and thus the corresponding penalty term value is larger. Specially, it is computed as the maximum difference of RGB color space to diminish the computation complexity for the color image.

![Figure 2. The sketch of horizontal tree structure. Here, \(i\) and \(j\) denote two adjacent nodes, and \(s, r\) denote two nonadjacent nodes. The penalty between nodes \(s\) and \(r\) is determined by penalties of edges on the horizontal-first path that connects \(s\) and \(r\).](image-url)
into our method. In the defined graph G, weight between nonadjacent nodes is propagated through the horizontal tree as shown in figure 2. Specifically, the penalty between two nonadjacent nodes s and r is obtained by summing up the penalties of edges on the path that connects nodes s and r in the horizontal tree. Apparently, there are so many paths that can connect nodes s and r in graph G. In principle, the best path should be selected to achieve better results, the penalty of which is the lowest. However, the computational complexity will be too high to find the best path. Thus we choose the horizontal-first strategy [4], namely the path of horizontal tree. Therefore, there is the only definite path to connect any two nodes on the horizontal tree structure. In addition, the penalty between nodes s and r is defined as follows.

\[ p(s, r) = \sum_{(i, j) \in P(s, r)} p(i, j) \]  

(6)

Here, nodes i and j denote two adjacent nodes located on the path that connects two nonadjacent nodes s and r. Clearly, the value of penalty term between two nodes is large if there is a long distance, even if their intensity values are similar. Furthermore, when the distance is certain, nodes at the same disparity level is more likely if their intensity values are similar, and thus the penalty is smaller.

Similar to traditional local algorithms, we define a weight in the step of cost aggregation. In our paper, we take the factors of intensity value and distance into consideration when computing the aggregation weight.

\[ \omega(s, r) = \exp\left(-\frac{p(s, r)}{\sigma}\right) \]  

(7)

where sigma is a user-specified parameter, and p(s, r) represents the penalty term between nodes s and r as mentioned above. The weight is defined by an exponential function, which dexterously forms the monotone decreasing relationship between weight \( w(s, r) \) and penalty \( p(s, r) \). And the range of weight is (0, 1).

After determining the weight, initial costs might be aggregated just like the conventional local methods. As mentioned above, some important information would be missed if the costs only in neighbor regions are used. Besides, the costs in textureless areas are still not distinct and close to zero after aggregation. It’s hard to determine disparities in these regions. Thus, we propose a non-local aggregation method, which can be expressed as:

\[ C_d^a(s) = \sum_{r \in I} \omega(s, r)C_d(r) \]  

(8)

where \( C_d(r) \) signifies the initial matching cost of pixel r at disparity level d. \( C_d^a(s) \) means the aggregated cost of pixel s at disparity level d. On the basis of (8), r covers every pixel in image I, namely we exploit the information of the entire image, which is different from conventional local methods.

2.3. Disparity computing based confidence coefficient

We improve our method in the step of cost aggregation as well as disparity computing. Generally speaking, the raw image would be affected by the illumination variations [9] and the noises, especially in the textureless regions. Therefore, the combination of different costs is applied. In our paper, the confidence coefficient is introduced. In contrast to conventional local methods, we combine the aggregated costs in the step of disparity computing, instead of in the step of cost computing by assigning weights to the initial costs.

As elaborated above, we utilize costs based color-gradient and census transform. These two costs are aggregated separately. And then we obtain two individual cost volumes. Take one pixel as an example, two curves of cost variation along disparity d for the pixel are acquired from two cost volumes respectively.

Let confidence coefficient be
\[ \rho = \frac{\min_d C_{A,d}^A(x)}{\min_d C_{A,d}^C(x)} \]  

(9)

Here, \( C_{A,d}^A(x) \) means the aggregated cost based census transform of pixel \( x \) with the disparity \( d \), \( \min_1 \) signifies the global minimum value, and \( \min_2 \) signifies the second local minimum value. \( R \) is the range of candidate disparity values. Apparently, the higher confidence coefficient is, the more confident the global minimum of \( C_A^A(x) \) is.

Obviously, the cost based color-gradient contains more information including color and gradient, and census transformation performs better while coping with radiometric variations. Therefore, for one pixel, we first select the disparity corresponding to the lowest aggregated cost based color-gradient according to the strategy of WTA, and then compute the confidence coefficient of aggregated census transform cost. We choose the disparity corresponding to lowest aggregated census transform cost only if the confidence coefficient obtained is greater than the threshold value \( \tau \). Namely, the disparity can be expressed as:

\[
D(p) = \begin{cases} 
\arg \min_d (C_{A,d}^{c-g}(p)) & \rho \geq \tau \\
\arg \min_d (C_{A,d}^A(p)) & \rho < \tau 
\end{cases}
\]  

(10)

In the experiment, we have tried computing the confidence coefficients of two aggregated costs respectively and selecting the disparity corresponding to cost of higher confidence coefficient. However, the result shows there is no benefit to accuracy enhancement but an increase of the computation complexity.

![Figure 3](image)

Figure 3. (a) The curve of aggregated cost based color-gradient along disparity of one pixel; (b) the curve of aggregated cost based census transform along disparity.

According to the WTA strategy, the best disparity of the pixel corresponding to cost based color-gradient is \( d_{c-g} \) in the figure 3, and that corresponding to cost based census transform is \( d_c \). If the confidence coefficient \( \rho = \frac{\min_1}{\min_2} \geq \tau \), let the best disparity be \( d_c \), otherwise be \( d_{c-g} \).

2.4. Disparity refinement
The algorithm runs in turn with both left and right images as reference image to obtain two disparity maps. To search pixels that are occluded or mismatched, we use the left-right consistency constraint: matched pixels in the stereo pair should have the same disparity.
\[ |d_{LR}(p) - d_{RL}(p) - d_{LR}(p)| \leq T_{LR} \]  

(11)

where \( T_{LR} \) denotes the threshold value of disparity difference. Label the pixels that satisfy the condition above with “stable”, otherwise label those with “unstable”. Let \( D \) be the current obtained disparity map. Then define the cost for each pixel at disparity \( d \):

\[
C'_{D}(p) = \begin{cases} 
|d - D(p)|, & \text{if } p \text{ is stable} \\
0, & \text{else} 
\end{cases}
\]

(12)

It’s noted that for all unstable pixels, the new cost volume \( C'_{D}(p) \) will be zero for all disparity levels. The cost aggregation algorithm and WTA are performed on this new cost again. And, the updated cost volume \( C'_{D}(p) \) would propagate the values from stable to unstable pixels.

3. Experimental results

In this section, the performance of the proposed stereo matching algorithm is evaluated on the Middlebury stereo benchmark [5]. All the experiments were performed on a computer containing a 3.30GHz, Intel Core i5 CPU, 8GB RAM and 64bit OS. Meanwhile, all the parameters are set to the same values in all experiments. According to the ground-truth disparity maps in Middlebury benchmark, set the error threshold is one pixel and compute the mismatch rate. The experiment results show that our proposed method achieves high accuracy while keeping the lower computational complexity. Specially, the matching accuracy in textureless regions is improved greatly, which is hard to accomplish for other local algorithms.

**Figure 4.** Results of benchmark stereo images on the Middlebury website. From top to bottom are: “Teddy”, “Pipes”, “Vintage” and “Plastic”. (a) Left stereo image. (b) The ground truth. (c) Result of the proposed method. (d) Result of TF-MST. (e) Result of LSE. Apparently, there are subtle differences between the disparity maps of “Teddy”, and all the methods perform well, which benefits from the abundant texture information. However, there are a large numbers of mismatched pixels in (d) and (e), such as in the white wall region of “Vintage” and the yellow plane region of “Plastic” marked with red rectangles.
Our approach is qualitatively evaluated using the Middlebury standard data sets [5] with accurate ground truth. Besides, disparity maps generated by other two non-local methods (Tree Filtering in minimum spanning tree (TF-MST) [10] and Local Smoothness Enforced (LSE) [11]) together with the proposed method are compared. TF-MST method adopts the structure of minimum spanning tree, and LSE also uses full image information while increasing the computational complexity sharply to compute the weights. Figure 4 shows the result of the subjective comparison. According to the figure 4, all the methods perform well in “Teddy”, since the abundant texture information makes great achievements to the high accuracy. However, the results of other three pairs of stereo images, generated by TF-MST and LSE, are not competitive with that of image “Teddy”. Specifically, disparity maps show the poor performance in the white featureless regions of “Pipes” and “Vintage” as well as the yellow plane region of “Plastic”.

Besides, accuracy comparison with current top-ranking local and non-local methods (TF-MST, LSE, Cross-Scale: CS-ST and CS-Guided Filter(CS-GF) [12], CostFilter [13]) are presented in table 1. The CS-ST uses local cost computing method, and aggregates cost volume cross scale, thus it’s also a non-local method in a certain sense. The proposed stereo algorithm achieves high accuracy while keeping low computational complexity.

Table 1. Performance evaluations on six algorithms in non-occluded regions. Error threshold is 1.0 pixel.

| Stereo Pairs | Our method | TF-MST | LSE | CS-ST | CS-GF | CF |
|--------------|------------|--------|-----|-------|-------|----|
| Teddy        | 3.57       | 6.79   | 3.93| 4.87  | 5.88  | 6.99|
| Cones        | 2.72       | 2.75   | 2.53| 3.31  | 2.90  | 3.23|
| Venus        | 0.27       | 0.28   | 1.04| 1.30  | 0.22  | 1.09|
| Tsukuba      | 1.24       | 1.87   | 1.61| 1.57  | 2.07  | 2.30|
| Pipes         | 6.73       | 8.98   | 9.67| 8.32  | 8.95  | 9.52|
| Vintage      | 9.24       | 14.62  | 29.24| 17.34 | 18.92 | 21.40|
| Plastic      | 7.46       | 17.54  | 15.86| 14.93 | 13.34 | 15.35|

4. Conclusions
In the paper, a non-local stereo matching algorithm is proposed. Our method is efficient based on these aspects. First, the computation complexity has been decreased greatly on the aggregation weight calculation as a result of summing up the penalty terms to propagate weight, while multiplication operations are usually required in other non-local stereo algorithms. Second, we combine the different aggregated costs by defining the confidence coefficient instead of assigning weight in conventional methods. Experiments demonstrate that our algorithm offers excellent high accuracy and speed performance in both old and new databases on Middlebury.

Acknowledgments
Thanks for the financial support from Technological projects of State Grid Corporation of China.

References
[1] Scharstein R Szeliski 2002 A taxonomy and evaluation of dense twoframe stereo correspondence algorithms Int. J. Compute. Vis. 47(1-3) 7–42
[2] He J Sun and X Tang 2013 Guided image filtering IEEE Trans. Patt. Anal. Mach. Intell. 35(6) 1397-1409
[3] Gastal, M M Oliverira 2011 Domain transform for edge-aware image and video processing ACM Trans. Graph. 30(4) 1-12
[4] Yang, D Li, L Wang, and et al. 2013 Full-Image guided filtering for fast stereo matching IEEE Signal Process. Lett. 20(3) 237-240
[5] Scharstein R Szeliski 2002 Middlebury Stereo Evaluation\ Version 2
[6] Zhan, Y Gu, K Huang, C Zhang and K Hu. 2015 Accurate image-guided stereo matching with efficient matching cost and disparity refinement IEEE Transactions on Circuits and Systems for Video Technology
[7] Zabih, J Woodfill 1994 Non-parametric local transforms for computing visual correspondence. European Conference on Computer Vision pp 151-158
[8] Bleyer, M Gelautz 2008 Simple but effective tree structures for dynamic programming-based stereo matching. Proc. Int. Conf. Comput. Vis. Theory Appl. pp 415-422
[9] Bleyer, C Breiteneder 2013 Stereo Matching-State-of-the-art and research challenges. Advanced Topics in Computer Vision, Springer pp 143-179
[10] Yang Stereo 2015 Matching using tree filtering in TPAMI 37(4) pp 834-846
[11] Qingqin Yang 2015 Local smoothness enforced cost volume regularization for fast stereo correspondence. IEEE Signal Process. Lett. 22(9)
[12] Zhang, Y Fang, D Min et al 2014 Cross-Scale cost aggregation for stereo matching in Proc. IEEE Conf. CVPR pp 407–414
[13] Rhemann, A Hosni, M Bleyer, C Rother and M Gelautz 2011 Fast cost-volume filtering for visual correspondence and beyond in Proc IEEE Conf. CVPR pp 3017–3024