Energy-Based Iterative Cost Aggregation in Depth Estimation with a Stereo Camera

Nguyen Xuan Truong
Electrical and Computer Engineering
Seoul National University
Seoul, Korea
truongnx@capp.snu.ac.kr

Huyk-Jae Lee
Electrical and Computer Engineering
Seoul National University
Seoul, Korea
hjlee@capp.snu.ac.kr

Abstract— **This paper presents a novel algorithm for performing an efficient cost aggregation in stereo vision. The cost aggregation is re-formulated under an iterative framework with a perspective of an energy model. The convergence of global energy is exploited to calculate the number of the iterations in cost aggregation. Experimental results show that the proposed method improves the quality of the disparity.**

I. Introduction

DEPTH estimation from a pair of stereo images has been one of the most important problems in computer vision ([1]). In general, stereo matching algorithms are classified into global and local ones according to the strategies used for estimation. In general, local algorithms are much faster and more compatible to practical applications than global ones. However, leading local algorithms which generate high-quality disparity maps still have high complexity. In this paper, we explore the convergence of an iterative cost aggregation stereo matching algorithm; then propose a novel method to perform an efficient cost aggregation.

The procedure of iterative local approaches is as follows. When a truncated absolute difference (TAD) is used to estimate a left disparity map, a per-pixel cost $C^0(p,d)$ for disparity d is first estimated by using the left and "f_p" -shifted right images, where I_l and I_r are left and right images, respectively. Aggregated cost $C^k(p,d)$ of pixel *p* at iteration *k* is then recursively computed via an adaptive summation of the cost at the previous iteration on the supportive window $N(p)$. Finally, the Winner-Takes-All (WTA) technique is performed for seeking the best one among all the disparity hypotheses.

$$\begin{aligned}
&C^0(p,d) = \min(|I_l(x,y) - I_r(x-d,y)|, \sigma) \\
&for\ k = 1 : T \\
&\qquad C^k(p,d) = \sum_{q \in N(p)} w(p,q)\, C^{k-1}(q,d) \qquad (1) \\
&\qquad d(p) = \arg\min_{d} C^T(p,d)
\end{aligned}$$

The per-pixel cost is truncated with a threshold σ to limit the influence of outliers to the dissimilarity measure. When the number of iteration is one, the algorithm is exactly same as the common non-iterative stereo matching algorithm.

II. Energy-Based Iterative Cost Aggregation

In the aggregation step (1), the weighting function can play an important role for gathering the information of neighboring pixels where disparity values are likely to be similar. Yoon and Kweon [2] proposed an adaptive (soft) weight approach which leverages the color and spatial similarity measures with the corresponding color images. The weighting function (or correlation) between pixels *p* and *q* is defined as follows:

$$w(p,q) = exp\left(-\sqrt{\left(I_p - I_q\right)^2}/\sigma_I - \sqrt{(p-q)^2}/\sigma_S\right) \quad (2)$$

Where σ_I and σ_S are color and spatial regularization constants. As the color similarity is measured by using a corresponding color image, it can be interpreted as a variant of joint bilateral filtering [3]. Note that a bilateral filter is an edge-preserving and noise-reduction smoothing filtering for images which preserves sharp edges by systematically looping through each pixel and adjusting weights to the adjacent pixels accordingly. Then the aggregated cost, which is iteratively computed by the filter, still preserves edges while reducing the noise.

The problem is how to define the number of cost aggregation steps in the iterative framework. In this paper, an energy-based method is used to define the number. Remind that general stereo matching algorithms can be defined as the energy minimization problem [4]. Let P be the set of pixels in an image and D be a finite set of disparities. The problem is to find a labeling that assigns a label $d \in D$ to each pixel $p \in P$. The subject is to minimize the labeling cost given by an energy function:

$$E = \sum_{p \in P} C(p,d) + \sum_{(p,q) \in N} V(d(p) - d(q)) \qquad (3)$$

where *N* are the (undirected) edges in the four-connected image grid graph. $C(p,d)$ is the cost of assigning disparity *d* to pixel *p*, and is referred to as the data cost. $V(d(p) - d(q))$ measures the cost of assigning disparities $d(p)$ and $d(q)$ to two neighboring pixels, and is often referred to as the discontinuity cost. If the discontinuity term is ignored, the energy function only includes the data term. It is obvious that

978-1-5090-3219-8/16/$31.00 ©2016 IEEE

the simple local solution gives an optimal solution in such case. This observation implies that an aggregation strategy can be actually interpreted into an iterative energy minimization problem in which the energy is reduced iteratively. The proposed energy-based iterative cost aggregation is based on this observation.

The procedure of the algorithm is presented in Figure 1. At first, the data cost is computed by the per-pixel cost, followed by the computation of a disparity map and its corresponding energy. As the weights only depend on the stereo images, they can be computed before the main loop to reduce the redundant computation. At each iteration the aggregated cost is computed by the variant of the bilateral filtering followed by the disparity and energy computation. The loop is terminated as the energy change is smaller than a threshold.

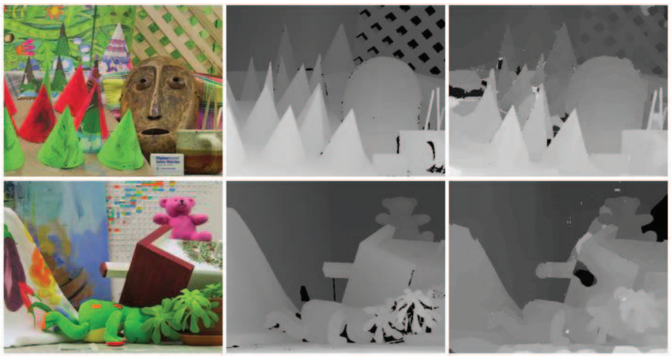
```
$C^0(p,d) = \min(|I_l(x,y) - I_r(x-d,y)|, \sigma)$
$d^0(p) = \arg\min_d C^0(p,d)$
$E^0 = \sum_{p \in P} C^0(p,d) + \sum_{(p,q) \in N} V(d^0(p) - d^0(q))$
$Compute\ w(p,q) = exp\left(-\sqrt{(I_p - I_q)^2}/\sigma_I - \sqrt{(p-q)^2}/\sigma_S\right)$
$Repeat$
    $k = k + 1$
    $C^k(p,d) = \sum_{q \in N(p)} w(p,q) C^{k-1}(q,d)$
    $d^k(p) = \arg\min_d C^k(p,d)$
    $E^k = \sum_{p \in P} C^0(p,d) + \sum_{(p,q) \in N} V(d^k(p) - d^k(q))$
$Until\ |(E^k - E^{k-1})/E^{k-1}| < th$
```

Figure 1: Pseudocode of the proposed algorithm

III. EXPERIMENTAL RESULTS

We have implemented the proposed method and evaluated with the Middlebury test bed: 'Cones', 'Teddy', 'Venus' and Sawtooth stereo images [5]. The parameters are set as follows $\sigma = 15.0, \sigma_I = 5.0, \sigma_S = RAD + 0.5, \lambda = 0.07$ and number of disparities is fixed by 64. The loop is terminated if the change of energy is less than 1% ($th = 0.01$). Table I shows the detail results. The first column is the radius of aggregation window (RAD) which is varying among 2, 3, 4, 8, 12, 16, and 32. For each dataset, results of the energy after the 1st iteration, the final energy, the number of iterations and the running time are reported. Column 2, 6, 10 and 14 shows the energy values after the 1st iteration for 'Cones', 'Teddy', 'Venus' and 'Sawtooth' image pairs, respectively. As the aggregation window increases, the energy decreases. The window 5x5 gives the largest initial energy, while the one 65x65 gives the smallest initial energy. Meanwhile, the window 17x17 often gives the smallest final energy as shown in the column 3, 7, 11 and 15.

The number of iterations are reported in column 4, 8, 12 and 16 while running time values are indicated in column 5, 9, 13 and 17. For each aggregation window, the running time is proportional to the number of iterations. For example, if RAD is 2 (or window is 5x5), each iterations of 'Cones' takes about 0.165s. On average, each iteration takes 0.158, 0.204, 0.300, 0.548, 0.874, 1.150 and 2.273 seconds when RAD is 2, 3, 4, 8, 12, 16 and 32, respectively. The time is proportional to RAD as we use two-pass approximation of the bilateral filter in each iteration. The horizontal aggregation followed by vertical ones is performed to make the complexity of the filter linear to the window size. It implies that the proposed algorithm is robust as a fast and simple cost aggregation method can output a high-quality disparity map (Figure 2).



ACKNOWLEDGMENTS

This research was supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2016-H8601-16-1008) supervised by the IITP(Institute for Information & communications Technology Promotion)

REFERENCES

[1] Scharstein, D. and R. Szeliski (2002). "A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms." International Journal of Computer Vision **47**(1): 7-42.

[2] Kuk-Jin, Y. and K. In So (2006). "Adaptive support-weight approach for correspondence search." IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(4): 650-656.

[3] Kopf, J., et al. (2007). Joint bilateral upsampling. ACM SIGGRAPH 2007 papers. San Diego, California, ACM**:** 96.

[4] Boykov, Y., et al. (2001). "Fast approximate energy minimization via graph cuts." IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(11): 1222-1239.

[5] http://vision.middlebury.edu/stereo/data/

TABLE I: RESULTS OF 'CONES', 'TEDDY', 'VENUS' AND 'SAWTOOTH' FOR TH = 0.01

RAD	Cones				Teddy				Venus				Sawtooth			
	1st	final	iters	time (s)	1st	final	iters	time (s)	1st	final	iters	time (s)	1st	final	iters	time (s)
2	10.074	1.488	41	6.78	7.546	1.257	38	6.24	6.142	0.640	27	4.11	4.697	0.562	28	4.21
3	8.245	1.342	30	6.31	5.918	1.155	27	5.62	4.039	0.317	43	8.64	3.078	0.401	32	6.29
4	7.020	1.261	25	6.57	4.974	0.940	32	8.02	2.964	0.268	39	9.22	2.350	0.417	23	5.72
8	4.669	**1.236**	15	8.45	3.455	0.864	20	11.08	1.458	**0.268**	20	10.65	1.435	**0.386**	13	7.02
12	3.784	1.253	12	10.71	2.873	**0.775**	18	15.45	1.099	0.272	13	11.21	1.135	0.400	9	7.96
16	3.341	1.265	12	13.82	2.531	0.779	16	18.26	0.929	0.285	10	11.56	1.014	0.423	7	8.05
32	2.911	1.485	7	15.62	2.122	0.848	10	21.99	0.665	0.386	6	13.71	0.805	0.510	4	9.51