A Revisit to Cost Aggregation in Stereo Matching: How Far Can We Reduce Its Computational Redundancy?

Dongbo Min[†] Jiangbo Lu[†] Minh N. Do[§]
Advanced Digital Sciences Center (ADSC), Singapore[†]
University of Illinois at Urbana-Champaign, IL, USA[§]
dbmin99@gmail.com, Jiangbo.Lu@adsc.com.sg, minhdo@illinois.edu

Abstract

This paper presents a novel method for performing an efficient cost aggregation in stereo matching. The cost aggregation problem is re-formulated with a perspective of a histogram, and it gives us a potential to reduce the complexity of the cost aggregation significantly. Different from the previous methods which have tried to reduce the complexity in terms of the size of an image and a matching window, our approach focuses on reducing the computational redundancy which exists among the search range, caused by a repeated filtering for all disparity hypotheses. Moreover, we also reduce the complexity of the window-based filtering through an efficient sampling scheme inside the matching window. The trade-off between accuracy and complexity is extensively investigated into parameters used in the proposed method. Experimental results show that the proposed method provides high-quality disparity maps with low complexity. This work provides new insights into complexity-constrained stereo matching algorithm design.

1. Introduction

Depth estimation from a stereo image pair has been one of the most important problems in the field of computer vision [1]. Generally, stereo matching methods can be classified into two approaches (global and local) according to the strategies used for estimation. It has been generally known that local approaches are much faster and more compatible to a practical implementation than global approaches. However, the complexity of the leading local approaches which provide high-quality disparity maps is still huge. In this paper, we explore the computational redundancy of cost aggregation in the local approaches and propose a novel method for performing an efficient cost aggregation.

Local approaches measure correlation between intensity values inside a matching window $N(p)$ of a reference pixel p, based on the assumption that all the pixels in the matching window have similar disparities. The performance highly depends on how to find an optimal window for each pixel. The general procedure of the local approaches is as follows. For instance, when a truncated absolute difference (TAD) is used to estimate a left disparity map, a per-pixel cost $e(p,d)$ for disparity hypothesis d is first calculated by using the left and 'd'-shifted right images. An aggregated cost $E(p,d)$ is then computed via an adaptive summation of the per-pixel cost. This process, which causes a huge complexity, is repeated for all the disparity hypotheses. The Winner-Takes-All (WTA) technique is finally performed for seeking the best one among all the disparity hypotheses as:

$$\begin{aligned} e(p,d) &= \min(|I_l(x,y) - I_r(x-d,y)|, \sigma) \\ E(p,d) &= \frac{\sum\limits_{q \in N(p)} w(p,q)e(q,d)}{\sum\limits_{q \in N(p)} w(p,q)} \\ d(p) &= \mathop{\arg\min}_{d \in [0,\cdots,D-1]} E(p,d), \end{aligned} \tag{1}$$

where I_l and I_r are left and right color images, respectively. The per-pixel cost is truncated with a threshold σ to limit the influence of outliers to the dissimilarity measure. Note that other dissimilarity measures such as Birchfield-Tomasi dissimilarity [2], rank/census transform [3] or normalized cross correlation (NCC) can also be used.

2. Previous work and motivation

For obtaining high-quality disparity maps, a number of local stereo matching methods have been proposed by defining the weighting function $w(p,q)$ which can implicitly measure the similarity of disparity values between pixel p and q. Yoon and Kweon [4] proposed an adaptive (soft) weight approach which leverages the color and spatial similarity measures with the corresponding color images, and it can be interpreted as a variant of joint bilateral filtering [10]. It is easy to implement and provides high accuracy, but has huge complexity due to its nonlinearity from the computation of the weighting function. The color segmentation based cost aggregation [5] was also presented with the assumption that pixels inside the same segment

978-1-4577-1102-2/11/$26.00 ©2011 IEEE

are likely to have similar disparity values. Cross-based approaches [6][7] used a shape-adaptive window which consists of multiple horizontal line segments spanning several neighboring rows. The shape of the matching window $N(p)$ is estimated based on the color similarity and an implicit connectivity constraint, and a hard weighting value (1 or 0) is finally used.

In general, the complexity of the cost aggregation can be characterized as $O(HWBD)$, where H and W are the size of an image, and B and D represent the size of the matching window and the search range, namely, the number of disparity hypotheses. In order to reduce the complexity of the cost aggregation, many algorithms have been proposed in terms of the size of the image HW and the matching window B. Min and Sohn [8] proposed a new multiscale approach for ensuring reliable cost aggregation. They tried to reduce the complexity by using smaller matching windows on the coarse image and cost domains. Richardt *et al.* [9] reduced the complexity of the adaptive support weight approach [4] by using an approximation of a bilateral filter [11]. The complexity is independent of the size of the matching window, but a grey image used in the bilateral grid causes some loss of quality, because it cannot preserve the discriminative power of color vectors completely when the weighting function $w(p,q)$ is computed.

In this paper, we extensively explore the principles behind the cost aggregation and propose a novel approach for performing the cost aggregation in an efficient manner. Different from the conventional approaches which have tried to reduce the complexity in terms of the size of the image and the matching window by using a multiscale scheme [8] or a signal processing technique [9], our approach focuses on reducing the redundancy which exists among the search range D, caused by the repeated calculation of $E(p,d)$ for all the disparity hypotheses in Eq. (1). Moreover, the redundancy which exists in the window-based filtering is exploited as well. We will show that the proposed spatial sampling scheme inside the matching window $N(p)$ can lead to a significant reduction of the complexity. Finally, the tradeoff between accuracy and complexity is extensively investigated over the parameters used in the proposed method.

3. Efficient cost aggregation

3.1. New formulation for cost aggregation

For local approaches, cost aggregation is the most important yet time-consuming part. In this paper, we re-formulate the cost aggregation problem of Eq. (1) as:

$$e^h(p,d) = \max(\sigma - |I_l(x,y) - I_r(x-d,y)|, 0)$$

$$E^{'}(p,d) = \frac{\sum\limits_{q\in N(p)} w(p,q)e^h(q,d)}{\sum\limits_{q\in N(p)} w(p,q)} \quad (2)$$

$$d(p) = \underset{d\in[0,\cdots,D-1]}{\arg\max}\ E^{'}(p,d)\ .$$

After applying the same procedure, the output disparity value $d(p)$ is estimated by seeking the *maximum* value of $E^{'}(p,d)$, which is the same to the solution of Eq. (1). The re-defined $e^h(p,d)$ is likely to have a large value as the disparity hypothesis d approaches a true disparity value. In this paper, we define $e^h(p,d)$ as a likelihood (evidence) function, since it represents a probability that the pixel p has for a specific disparity hypothesis d. We further modify the formulation of the cost aggregation by omitting the normalization term $\sum w(p,q)$ in Eq. (2). This modification does not affect the accuracy of the cost aggregation, since the disparity value $d(p)$ is estimated for each pixel independently where this normalization term is fixed for all ds. The aggregated likelihood $E^h(p,d)$ is then defined as follows.

$$E^h(p,d) = \sum_{q\in N(p)} w(p,q)e^h(q,d) \quad (3)$$

It has a similar formulation to a histogram which represents a probability distribution of continuous (or discrete) values in a given data. In general, each bin of the histogram can be calculated by counting the number of corresponding observations in the set of data. Similarly, given the data set of the neighboring pixels q, the d^{th} bin of the reference pixel p is computed by counting the bin with the corresponding $e^h(q,d)$. Since a single pixel q is associated with a set of multiple data (*i.e.* $e^h(q,d)$ for all bin ds), the aggregated likelihood function $E^h(p,d)$ can be referred to as a *relaxed* histogram.

Another characteristic of the proposed histogram-based aggregation is the use of the weighting function $w(p,q)$. As previously mentioned, the weighting function can play an important role for gathering the information of neighboring pixels where disparity values are likely to be similar. In this paper, we use a similarity measure based on the color and spatial distances as follows [4][8]:

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

Since the color similarity is measured by using a corresponding color image, it shares the similar principle to the joint bilateral filtering [10], where the weight is computed with a signal different from the signal to be filtered. This characteristic enables the joint histogram to be extended into a weighted filtering with the support of color discriminative power. In the following section, we will describe two methods for reducing the complexity of building the joint histogram $E^h(p,d)$.

3.2. First approximation: compact representation of likelihood for search range

Recently, several methods have been proposed using a compact representation of the data that consists of a com-

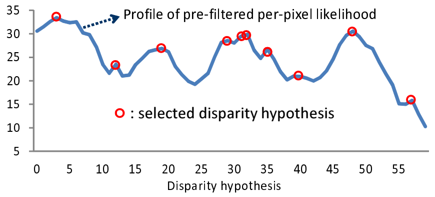


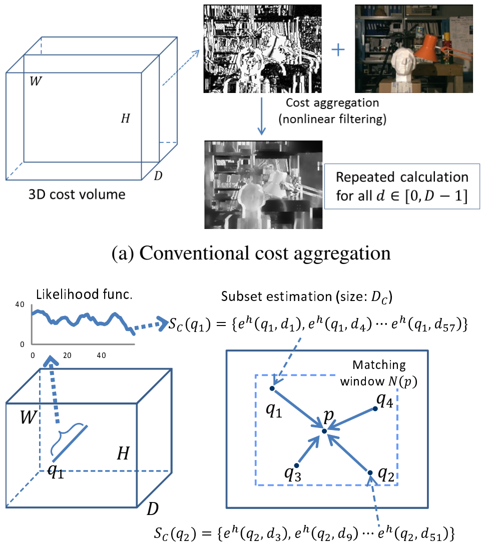
Figure 1. Disparity candidate selection with local/global maxima.

plex form in stereo matching. Yu *et al.* [12] proposed a novel envelope point transform (EPT) method by applying a principal components analysis (PCA) to compress messages used in belief propagation [15]. Wang *et al.* [13] estimated the subset of disparity hypotheses for reliably matched pixels and then propagated them on MRF formulation for estimating the subset of unreliable pixels. Yang *et al.* [14] proposed the method for reducing the search range and applied it into hierarchical belief propagation [16]. PCA or Gaussian Mixture Model (GMM) can be used for the compact representation, but the compression for all pixels is time-consuming.

The weighting function $w(p,q)$ based on the color and spatial distances have been used to obtain accurate disparity maps as in Eq. (2). The cost aggregation hence becomes a non-linear filtering, whose complexity is very high. In this paper, we propose a new approach for reducing the complexity from a perspective of the relaxed joint histogram. Our key idea is *to find a compact representation of the per-pixel likelihood* $e^h(p,d)$, based on the assumption that $e^h(p,d)$ with low values do not provide really informative support on the histogram-based aggregation.

In this paper, we extract the subset of local maxima at the per-pixel likelihood $e^h(p,d)$ for the compact representation. The per-pixel likelihood for each pixel is pre-filtered with a 5×5 box window for suppressing noise. The pre-filtering is done for all disparity hypotheses, but its complexity is trivial in case of using a spatial sampling method, which will be described in the next section. The local maximum points are calculated by using the profile of the pre-filtered likelihood function. They are then sorted in a descending order and a pre-defined number of disparity candidates $D_c(\ll D)$ are finally selected. If the number of the local maxima is less than D_c, the values corresponding to the 2^{nd}, 3^{rd} (and so on) highest likelihood are selected. Fig. 1 shows an example of the disparity candidate selection for 'Teddy' stereo images, where the number of the disparity hypotheses is 60. The new aggregated cost $E^h(p,d)$ is defined with the subset of disparity hypotheses only.

$$\begin{aligned} E^h(p,d) &= \sum_{q \in N(p)} w(p,q) e_1^h(q,d) o(q,d) \\ o(q,d) &= \begin{cases} 1 & d \in S_C(q) \\ 0 & otherwise \end{cases} \end{aligned} , \tag{4}$$



(a) Conventional cost aggregation

Likelihood func.
40
0
0
20
40
Subset estimation (size: D_C)
$S_C(q_1) = \{e^h(q_1,d_1), e^h(q_1,d_4) \cdots e^h(q_1,d_{57})\}$
W
H
D
q_1
Matching
window $N(p)$
q_1
p
q_4
q_3
q_2
$S_C(q_2) = \{e^h(q_2,d_3), e^h(q_2,d_9) \cdots e^h(q_2,d_{51})\}$

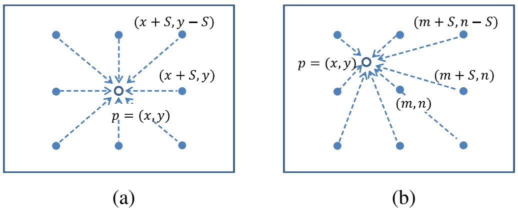
(b) Proposed method

Figure 2. Cost aggregation: (a) conventional approaches perform nonlinear filtering with (or without) a color image for all disparity hypotheses: $O(HWBD)$. (b) Proposed method estimates the subset of disparity hypotheses, whose size is $D_c(\ll D)$, and then performs joint histogram-based aggregation: $O(HWBD_c)$.

where $S_C(q)$ is a subset of disparity hypotheses whose size is D_c. Note that $S_C(q)$ varies for all pixels. e_1^h represents the prefiltered likelihood with 5×5 box window. Fig. 2 explains the difference between the conventional cost aggregation and the proposed method. When the size of the matching window is set to B, the conventional method performs the non-linear filtering for all pixels (HW) and disparity hypotheses (D), so the complexity is $O(HWBD)$. In contrast, the proposed method votes the subset of informative per-pixel likelihoods (whose size is D_c) into $E^h(p,d)$ with the complexity of $O(HWBD_c)$. Moreover, since the normalization term $\sum w(p,q)$ is not used in the joint histogram $E^h(p,d)$, the complexity has been further reduced. We will show in the experimental results that the compact representation by the subset of local maxima is helpful for reducing the complexity while maintaining the accuracy.

3.3. Second approximation: spatial sampling of matching window

Another source for reducing the complexity is on the spatial sampling inside the matching window. There is a trade-off between the accuracy and the complexity according to the size of the matching window. In general, using a large matching window and a well-defined weighting function $w(p,q)$ for obtaining a high quality disparity map leads to high computational complexity [4][8]. In this paper, we handle this problem with a spatial sampling scheme inside



(a) (b)

Figure 3. Spatial sampling of matching window: (a) reference pixel p-dependent, (b) reference pixel p-independent sampling.

the matching window, different from the previous work that used the signal processing technique [9].

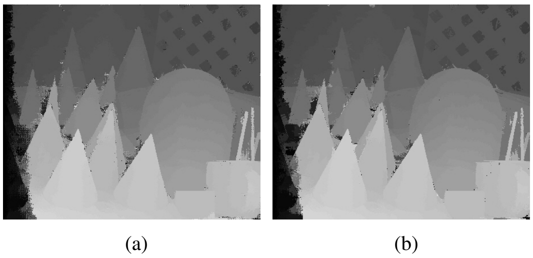
Many approaches have used a smoothness assumption that disparities inside an object vary smoothly, except near the boundaries. A large window is generally needed for reliable matching, but *it does not mean that all the pixels inside the matching window, whose disparity values are likely to be similar in case of being located in the same object, should be used altogether*.

This observation suggests that the spatial sampling inside the matching window can reduce the complexity of the window-based filtering. More specifically, the sparse samples inside the matching window could be enough to gather reliable information. Ideally, the pixels can be classified according to their likelihoods. It is, however, impossible to classify the pixels inside the matching window according to their disparity values, which should be finally estimated. Color segmentation may be a good choice for grouping the pixels, but the segmentation is time-consuming and not feasible for a practical implementation.

In this paper, a simple but powerful way for the spatial sampling is proposed. The pixels inside the matching window are regularly sampled, and then only the sampled ones are used for the joint histogram-based aggregation in Eq. (4). The neighboring pixels which are close to each other are likely to have similar disparity values, so that the regularly-sampled data is sufficient for ensuring reliable matching so long as the pixels at a distance are used. As shown in Fig. 3, there are two ways for spatial sampling: reference pixel *p-dependent* and *independent* sampling. The p-dependent sampling can be defined as follows:

$$\begin{aligned} &E^h(p,d) = \sum_{q \in N(p)} w(p,q) e_1^h(q,d) o(q,d) s_1(p,q) \\ &s_1(p,q) = \begin{cases} 1 & |p-q|\%S = 0 \\ 0 & otherwise \end{cases} \end{aligned} , \quad (5)$$

where $s_1(p,q)$ is a binary function capturing the regularly-sampled pixels inside the matching window for a sampling ratio S. As previously mentioned, the prefiltering with 5×5 window is applied for suppressing noise in the disparity candidate selection. Since the likelihood profile for all disparity hypotheses is saved for estimating the local maxima, a $3D$ volume of $e^h(p,d)$ should be constructed for performing the efficient prefiltering with the constant time box filtering. However, it causes a huge amount of memory (for a $3D$ volume). For instance, a pair of HD (1920×1080) images with 300 disparity candidates need 4.8GB to store a floating point 3D cost volume, which make it difficult to implement the algorithm efficiently on a GPU or an embedded system. We hence calculate the dissimilarity measure every time, not saving the precalculated per-pixel likelihoods. In other words, the constant time or separable box filtering methods are not used. However, this leads to *relatively* high complexity, compared to that of the joint histogram-based aggregation. For instance, when $S = 3$ and $D_c = 6$ for the 'Cone' image in Fig. 4 (a), the processing time $(3.01s)$ of the disparity candidate selection, which consists of dissimilarity measure, box filtering, and local maxima estimation/sorting, is much longer than that $(0.57s)$ of the joint histogram-based cost aggregation.



(a) (b)

Figure 4. Examples of the disparity maps estimated by two sampling methods on the 'Cone' image when $S = 3$ and $D_c = 6$: (a) p-dependent sampling, (b) p-independent sampling. The processing times are $3.58s(= 3.01s + 0.57s)$ and $0.91s(= 0.34s + 0.57s)$.

The reference pixel p-independent sampling can handle this problem. As shown in Fig. 3 (b), our new sampling scheme can be defined as follows:

$$\begin{aligned} &E^h(p,d) = \sum_{q \in N(p)} w(p,q) e_1^h(q,d) o(q,d) s_2(q) \\ &s_2(q) = \begin{cases} 1 & q\%S = 0 \\ 0 & otherwise \end{cases} \end{aligned} , \quad (6)$$

where $s_2(q)$ is also a binary function which is similar to $s_1(p,q)$, but does not depend on the reference pixel p. All the reference pixels are supported by the same regularly-sampled neighboring pixels, so that we can reduce the complexity of the disparity candidate selection with a factor of the sampling ratio $S \times S$. The dissimilarity is first measured and the subset of the disparity hypotheses are then estimated for every S pixel. Note that the sampling ratio S is related to the sampling of the neighboring pixels only. Table 1 shows a pseudo code for the proposed method.

Fig. 4 shows disparity maps estimated by two sampling methods on the 'Cone' image, when $S = 3$ and $D_c = 6$. The post-processing such as median filtering or occlusion handling was not used to evaluate the performance of two sampling methods only. The results are similar, except that Fig. 4 (a) contains some checkered patterns on the object

Table 1. Pseudo code for efficient likelihood aggregation.

```
Parameter definition
$HW$: The size of an image $I$
$B$: The size of matching window $N(p)$ (=$M \times M$)
$S_D$: The set of disparity hypotheses whose size is $D$
$S_C$: The subset of disparity hypotheses whose size is $D_c$
$S$: Sampling ratio inside a matching window

Algorithm: Efficient likelihood aggregation

DISPARITY CANDIDATE SELECTION
Complexity: $O(25HWD/S^2)$
For all pixels p which satisfy $p\%S = 0$ and $p \in I$
    1: Initialize prefiltered likelihood function $e_1^h(p,d)$
       to 0 for all $d$s.
    For all disparity candidates $d \in S_D(p)$
        For all neighboring pixels which satisfy $|p-q|_\infty \leq 2$
            2: Compute per-pixel likelihood $e^h(q,d)$ and
               $e_1^h(p,d) += e^h(q,d)$ ($5 \times 5$ box filtering)
        End
    End
    3: Estimate $S_C(p)$ with the local maxima on $e_1^h(p,d)$
End

JOINT HISTOGRAM-BASED AGGREGATION
Complexity: $O(HWBD_c/S^2)$
For all reference pixels $p \in I$
    4: Initialize likelihood function $E^h(p,d)$ to 0 for all $d$s.
    For neighboring pixels which satisfy $|q_1|_\infty \leq M/2S$
        5: Compute weight $w(p,q)$ with color and spatial
           distances between two neighboring pixels
           $p$ and $q = ((int)(p/S) + q_1) \times S$.
           (Reference pixel $p$-independent sampling)
        For all disparity candidates $d_q \in S_C(q)$
            6: $E^h(p,d_q) += w(p,q) \times e_1^h(q,d_q)$
        End
    End
    7: $d(p) = \underset{d \in [0,\cdots,D-1]}{\arg\max} E^h(p,d)$
End
```

boundaries, while the processing times are $3.58s (= 3.01s + 0.57s)$ and $0.91s (= 0.34s + 0.57s)$, respectively.

4. Comparative study

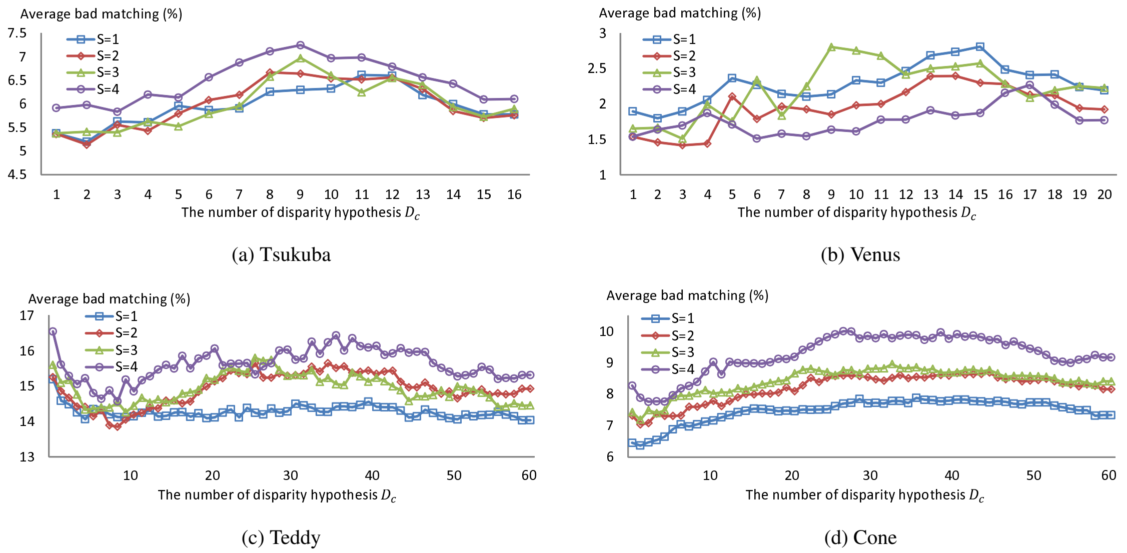
We have implemented the proposed method and compared the performance with state-of-the-arts methods in the Middlebury test bed: 'Tsukuba,' 'Venus,' 'Teddy,' and 'Cone' stereo images [19]. The estimated disparity maps are evaluated by measuring the percent of bad matching pixels (where the absolute disparity error is larger than 1 pixel) for three subsets of an image: nonocc (the pixels in the nonoccluded region), all (all the pixels), and disc (the visible pixels near the occluded regions).

The proposed method has been tested using the same parameters, except for two parameters: the number of disparity candidates D_c and the spatial sampling ratio S. We investigated the effects of these two parameters for the accuracy and the complexity. The CIELab color space is used for calculating the weighting function $w(p,q)$, where σ_I and σ_S are 5.0 and 17.5, respectively. The size of the matching window $N(p)$ is set to 31 × 31, and the census transform [3], which is robust against photometric distortion, is used for measuring the per-pixel likelihood $e^h(p,d)$. Occlusion is also handled to evaluate the overall accuracy of the estimated disparity maps. The occluded pixels are detected by a cross-checking technique and the disparity value of background regions is then assigned to the occluded pixels.

Fig. 5 shows an objective evaluation according to the number of depth candidate D_c and the spatial sampling ratio S. The average percent (%) of bad matching pixels for 'nonocc', 'all' and 'disc' regions is shown for each sampling ratio S. Note that when S is set to 1 and all disparity hypotheses are used (e.g. $D_c = 60$ for 'Teddy'), the proposed method is equivalent to the conventional cost aggregation, except that the joint histogram-based aggregation is used. We could find that the bad matching percent does not converge (or sometimes it increases) as the number of disparity hypothesis D_c increases. It indicates that using the information of all the disparity hypotheses does not necessarily guarantee to obtain the accurate disparity maps. In other words, *unnecessary candidates with low likelihood (evidence) values may contaminate the likelihood aggregation process*. In terms of the spatial sampling ratio S, we found that the quality of the disparity maps is gradually degenerated as S increases, but the results of $S = 1, 2, 3$ are similar.

Next, we investigated the trade-off between the accuracy and the complexity by comparing processing times in Fig. 6. Due to the lack of space, we showed the results of 'Tsukuba' and 'Teddy' only. Note that the proposed method was implemented on the CPU only, but it is easy to implement on the GPU or FPGA thanks to its efficient memory utilization and compatibility to parallel processing. The processing time was measured for the calculation of a single (left or right) disparity map and did not include the occlusion detection/handling in order to compare the complexity of the likelihood aggregation only. As expected, the processing time is proportional to the number of disparity hypotheses D_c and the sampling ratio $S \times S$. Interestingly, when the number of disparity hypotheses D_c is small (e.g. $D_c = 1 \sim 10$ for 'Teddy' or 'Cone'), the processing times for $S = 3$ and 4 are almost similar. The trade-off in Fig. 6 (b) and (d) shows that the accuracy is not monotonically increasing as the processing time (D_c) increases.

Fig. 7 shows the accuracy of the disparity candidate selection in 'nonocc' regions of 'Teddy' image according to the number of disparity hypotheses D_c. It was calculated



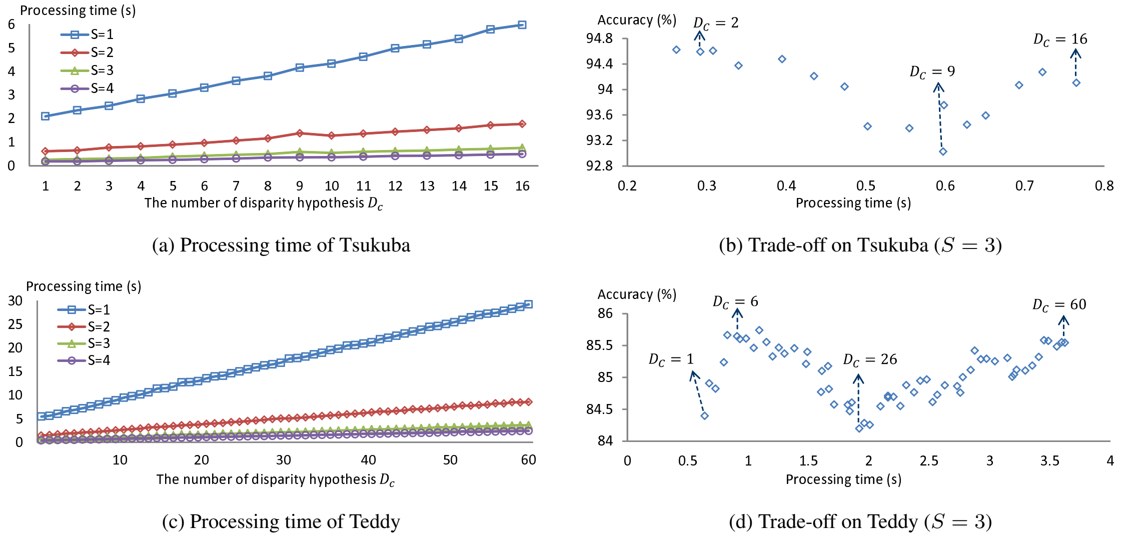
(a) Tsukuba

(b) Venus

(c) Teddy

(d) Cone

Figure 5. Objective evaluation: average percent (%) of bad matching pixels for 'nonocc', 'all' and 'disc' regions according to D_c and S.



(a) Processing time of Tsukuba

(b) Trade-off on Tsukuba ($S = 3$)

(c) Processing time of Teddy

(d) Trade-off on Teddy ($S = 3$)

Figure 6. Processing times (a,c) and trade-off (b,d) of the proposed method according to D_c and S. The results of 'Tsukuba' and 'Teddy' images only are shown due to the lack of space. One can find that the accuracy is not monotonically increasing as the processing time (D_c) increases.

by counting the number of pixels whose subsets actually include a ground truth disparity value. When $D_c = 60$, namely the same to the original size, the subsets of all pixels include the ground truth disparity value. Interestingly, when $D_c = 6$, only 89.5% pixels contain the ground truth disparity values in their subsets, but the accuracy of the estimated disparity map (91.69%) is almost similar to these of the best one $(91.75\%$, when $D_c = 11)$ or the disparity map estimated with all the disparity hypotheses $(91.67\%$, when $D_c = 60)$. This shows that the joint histogram based aggregation can reliably handle errors of the initial candidate selection by gathering the information appropriately from the subsets of the neighboring pixels.

Fig. 8 shows the examples of the disparity maps estimated by the proposed method when the number of disparity hypotheses D_c is 10% of the original search range and the spatial sampling ratio S is fixed to 3. Namely, D_c is set to 2 for 'Tsukuba', 2 for 'Venus', 6 for 'Teddy', and 6 for 'Cone', respectively. One could find that the proposed method provides high-quality disparity maps, even though a small number of disparity hypotheses are used. The processing times including the occlusion detection/handling are $0.34s$ for 'Tsukuba', $0.54s$ for 'Venus', $0.98s$ for 'Teddy', and $1.01s$ for 'Cone', respectively.

The objective evaluation is shown in Table 2 by reporting a comparison with other state-of-the-art methods. The

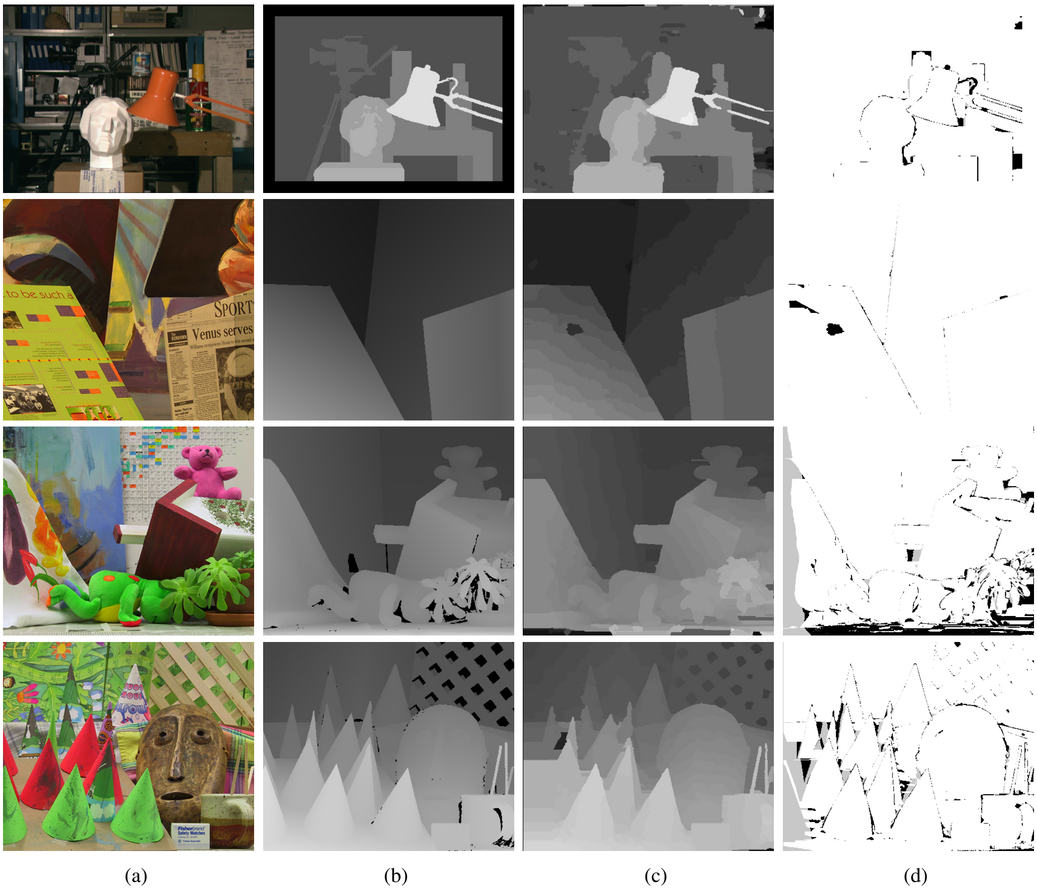


Figure 8. Results for (from top to bottom) ‘Tsukuba’, ‘Venus’, ‘Teddy’ and ‘Cone’ image pairs: (a) original images, (b) ground truth maps, (c) our results, (d) error maps. The number of disparity hypotheses D_c is set to 10% of the original search range and the spatial sampling ratio S is set to 3.

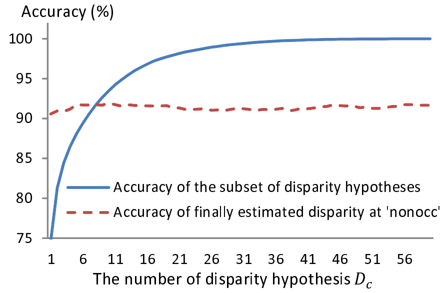


Figure 7. Accuracy of the disparity candidate selection and the finally estimated disparity map in ‘nonocc’ regions of ‘Teddy’ according to D_c.

methods were sorted with APBP (Average Percent Bad Pixels). ‘**Proposed method 1**’ represents an objective evaluation of Fig. 8. ‘**Proposed method 2**’ is the result when the parameters (S and D_c) that provide the disparity maps with the best accuracy are used.

For the comparison of the complexity, we referred to the results reported in the previous work. The processing time of the ‘AdaptWeight’ method [4] is $60s$ for ‘Tsukuba’, in contrast to $0.34s$ of the proposed method. According to [17], the processing time of ‘FastBilateral’ method is $32s$ for ‘Teddy’, while that of the proposed method is only $0.98s$. Note that the processing time of the proposed method also includes the occlusion detection/handling, while the previous works consider the cost aggregation only.

One interesting fact is that two methods for reducing the complexity of the joint histogram-based aggregation can be combined with other cost aggregation methods as well. A number of local approaches have been proposed by defining the weighting function $w(p, q)$ with hard or soft values. After re-formulating these methods into the histogram-based scheme, the compact representation of per-pixel likelihoods and the spatial sampling of the matching window can be used for an efficient implementation. Moreover, the trade-off between the accuracy and the complexity presented here can be taken into account in the complexity-constrained algorithm design.

Table 2. Objective evaluation for the proposed method with the Middlebury test bed [19]

Algorithm	*Tsukuba*			*Venus*			*Teddy*			*Cone*			APBP (%)
	nocc	all	disc	nocc	all	disc	nocc	all	disc	nocc	all	disc	
CostAggr+occ [8]	1.38	1.96	7.14	0.44	1.13	4.87	6.80	11.9	17.3	3.60	8.57	9.36	6.20
AdaptWeight [4]	1.38	1.85	6.90	0.71	1.19	6.13	7.88	13.3	18.6	3.97	9.79	8.26	6.67
Proposed method 2	2.37	2.67	10.4	0.72	0.94	2.59	7.97	13.4	20.2	2.91	8.10	8.10	6.69
FastBilateral [17]	2.38	2.80	10.4	0.34	0.92	4.55	9.83	15.3	20.3	3.10	9.31	8.59	7.31
Proposed method 1	2.47	2.71	11.1	0.74	0.97	3.28	8.31	13.8	21.0	3.86	9.47	10.4	7.33
VariableCross [6]	1.99	2.65	6.77	0.62	0.96	3.20	9.75	15.1	18.2	6.28	12.7	12.9	7.60
ESAW [18]	1.92	2.45	9.66	1.03	1.65	6.89	8.48	14.2	18.7	6.56	12.7	14.4	8.21
FastAggreg	1.16	2.11	6.06	4.03	4.75	6.43	9.04	15.2	20.2	5.37	12.6	11.9	8.24
AdaptPolygon	2.29	2.88	8.94	0.80	1.11	3.41	10.5	15.9	21.3	6.13	13.2	13.3	8.32
DCBGrid [9]	5.9	7.26	21.0	1.35	1.91	11.2	10.5	17.2	22.2	5.34	11.9	14.9	10.9
SSD+MF	5.23	7.07	24.1	3.74	5.16	11.9	16.5	24.8	32.9	10.6	19.8	26.3	15.7

5. Conclusion

In this paper, we have presented a novel approach for the efficient cost aggregation in stereo matching. We reformulated the problem in the perspective of the relaxed joint histogram, given the per-pixel likelihood (evidence) function. Some algorithms were then proposed for reducing the complexity of the joint histogram-based aggregation. Different from the conventional local approaches, we could reduce the complexity in terms of the search range by estimating a subset of informative disparity hypotheses. The experimental results showed that the reliable disparity maps were obtained even when the number of disparity hypotheses D_c was less than 10% of the original search range. Moreover, the complexity of the window-based processing was dramatically reduced while keeping a similar accuracy through the reference pixel-independent sampling of the matching window. In further research, we will investigate more elaborate algorithms for selecting the subset of disparity hypotheses. As shown in Fig. 5, the optimal number of disparity hypotheses is dependent on the characteristics of input images and the spatial sampling ratio S, even though the proposed method can provide excellent results with a fixed number of disparity hypotheses (e.g. 10% of the original search range). We plan to devise an efficient method for estimating the optimal number D_c adaptively for different input images.

References

[1] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *IJCV*, vol. 47, no. 13, 7-42, Apr. 2002. 1

[2] S. Birchfield and C. Tomasi, "Depth discontinuities by pixel-to-pixel stereo," *IJCV*, vol. 35, no. 3, Dec. 1999. 1

[3] R. Zabih and J. Woodfill, "Non-parametric local transforms for computing visual correspondence," *ECCV*, 1994. 1, 5

[4] K. Yoon and I. Kweon, "Adaptive support-weight approach for correspondence search," *IEEE Transactions on PAMI*, vol. 28, no. 4, pp. 650-656, Apr. 2006. 1, 2, 3, 7, 8

[5] F. Tombari, S. Mattoccia, L. D. Stefano, and E. Addimanda, "Near real-time stereo based on effective cost aggregation," *IEEE Proc. ICPR*, 2008. 1

[6] K. Zhang, J. Lu, and G. Lafruit, "Cross-based local stereo matching using orthogonal integral images," *IEEE Trans. on CSVT*, vol. 19, no. 7, pp. 1073-1079, Jul. 2009. 2, 8

[7] K. Zhang, J. Lu, G. Lafruit, R. Lauwereins, and L. Van Gool, "Real-time accurate stereo with bitwise fast voting on CUDA," *IEEE Proc. ICCVW*, 2009. 2

[8] D. Min and K. Sohn, "Cost aggregation and occlusion handling with WLS in stereo matching," *IEEE Trans. on Image Processing*, vol. 17, no. 8, pp. 1431-1442, Aug. 2008. 2, 3, 8

[9] C. Richardt, D. Orr, I. Davies, A. Criminisi, and N. A. Dodgson, "Real-time spatiotemporal stereo matching using the dual-cross-bilateral grid," *Proc. ECCV*, 2010. 2, 4, 8

[10] J. Kopf, M. F. Cohen, D. Lischinski and M. Uyttendaele, "Joint bilateral upsampling," *ACM SIGGRAPH* 2007. 1, 2

[11] S. Paris and F. Durand, "A fast approximation of the bilateral filter using a signal processing approach," *Proc. ECCV*, pp. 568-580, 2006. 2

[12] T. Yu, R.-S. Lin, B. S., and B. Tang, "Efficient message representations for belief propagation," *IEEE ICCV*, 2007. 3

[13] L. Wang, H. Jin, and R. Yang, "Search space reduction for MRF stereo," *Proc. ECCV*, 2008. 3

[14] Q. Yang, L. Wang, and N. Ahuja, "A constant-space belief propagation algorithm for stereo matching," *IEEE Proc. CVPR*, 2010. 3

[15] J. Sun, N. Zheng, and H. Y. Shum, "Stereo matching using belief propagation," *IEEE Trans. on PAMI*, vol. 25, no. 7, pp. 787800, 2003. 3

[16] P. Felzenszwalb and D. Huttenlocher, "Efficient belief propagation for early vision," *IJCV*, vol. 70, no. 1, 2006. 3

[17] S. Mattoccia, S. Giardino, and A. Gambini, "Accurate and efficient cost aggregation strategy for stereo correspondence based on approximated joint bilateral filtering," *Proc. ACCV*, 2009. 7, 8

[18] W. Yu, T. Chen, F. Franchetti, and J. Hoe, "High performance stereo vision designed for massively data parallel platforms," *IEEE Trans. on CSVT*, 2010. 8

[19] http://vision.middlebury.edu/stereo. 5, 8