Joint Histogram Based Cost Aggregation for Stereo Matching

Dongbo Min, *Member, IEEE*, Jiangbo Lu, *Member, IEEE*, and Minh N. Do, *Senior Member, IEEE*

Abstract—This paper presents a novel method for performing efficient cost aggregation in stereo matching. The cost aggregation problem is re-formulated from a perspective of a histogram, giving us a potential to reduce the complexity of the cost aggregation in stereo matching significantly. Different from 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 the 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 by varying the parameters used in the proposed method. Experimental results show that the proposed method provides high-quality disparity maps with a low complexity and outperforms existing local methods. This work also provides new insights into complexity-constrained stereo matching algorithm design.

Index Terms—Cost aggregation, stereo matching, disparity hypotheses, joint histogram.

1 Introduction

Depth estimation from a stereo image pair [1] has been one of the most fundamental tasks in the field of computer vision. It aims at estimating a pair of corresponding points between two (or more) consecutive images taken from different viewpoints. Stereo matching can be classified into two categories (global and local) according to the strategies used for estimation. Global approaches generally define an energy model with various constraints (using smoothness or uniqueness assumptions) and solve it using global optimization techniques such as belief propagation or graph cut. Local approaches obtain a disparity map by measuring correlation of color patterns in local neighboring windows. It has been generally known that the local approaches are much faster and more suitable for a practical implementation than global approaches. However, the complexity of the leading local approaches which provide high-quality disparity maps is still huge. This paper explores the computational redundancy of cost aggregation in the local approaches and proposes a novel method for performing an efficient cost aggregation.

Local approaches measure a correlation between intensity values inside a matching window $\mathcal{N}(p)$ of a reference pixel p, based on the assumption that all the pixels in the matching window have similar disparities. The performance depends heavily on how to find an optimal window for each pixel. The general procedure of local approaches is as follows. For instance, suppose that a truncated absolute difference (TAD) is used to estimate a left disparity map D_l. A per-pixel raw matching cost $e(p,d)$ for disparity hypothesis d is first calculated by using the left and 'd'-shifted right images as follows:

$$e(p,d) = \min(||I_l(x,y) - I_r(x-d,y)||, \sigma), \tag{1}$$

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. 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, stepping from 0 to $D-1$:

$$E(p,d) = \frac{\sum_{q \in \mathcal{N}(p)} w(p,q) e(q,d)}{\sum_{q \in \mathcal{N}(p)} w(p,q)}. \tag{2}$$

The Winner-Takes-All (WTA) technique is finally performed for seeking the best one among all the disparity hypotheses as:

$$D_l(p) = \underset{d \in [0,\cdots,D-1]}{\arg\min} E(p,d). \tag{3}$$

2 Related 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

- *D. Min and J. Lu are with the Advanced Digital Sciences Center, Singapore. E-mail: dongbo@adsc.com.sg; jiangbo.lu@adsc.com.sg*
- *M. N. Do is with the University of Illinois at Urbana-Champaign, Urbana, Illinois, USA. E-mail: minhdo@illinois.edu*

color and spatial similarity measures with the corresponding color images, and it can be interpreted as a variant of joint bilateral filtering [5]. 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 [6] was also presented with the assumption that pixels inside the same segment are likely to have similar disparity values. Cross-based approaches [7] used a shape-adaptive window which consists of multiple horizontal line segments spanning several neighboring rows. The shape of the matching window $\mathcal{N}(p)$ 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 $O(NBL)$, where N and B are the size of the input image and the matching window $\mathcal{N}(p)$, and L represents the search range, i.e., the number of discrete labels (e.g. disparity hypotheses). To reduce the complexity of the cost aggregation, a number of algorithms have been proposed in terms of the size of the image N and the matching window B. Min and Sohn [8] proposed a new multiscale approach for ensuring reliable cost aggregation in the stereo matching. They tried to reduce the complexity by using smaller matching windows on the coarse image and cost domain. Richardt *et al.* [9] reduced the complexity of the adaptive support weight approach [4] by using an approximation of the bilateral filter [10]. 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.

An iterative solution [11], inspired by the anisotropic diffusion, was proposed to achieve similar results to the adaptive weight approach [4] with a lower computational load. It was shown that the geodesic diffusion is efficiently performed after a few iterations and produces state-of-the-arts results among the local stereo methods [11]. Rhemann *et al.* [12] formulated several computer vision tasks with a discrete labeling problem, and then performed the cost aggregation with the guided image filtering [13], which allows the constant time implementation regardless of the window size. They demonstrated that this simple and generic framework achieved very competitive results in the stereo matching, optical flow estimation, and interactive segmentation. The complexity increases linearly with an image size (N) and the number of labels (L) only.

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 the multiscale scheme [8] or the constant time filtering techniques [9] [12], our approach focuses on reducing the redundancy which exists among the search range L, caused by the repeated calculation of $E(p,d)$ for all the disparity hypotheses in (2). 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 $\mathcal{N}(p)$ can lead to a significant reduction of the complexity. Finally, the trade-off between accuracy and complexity is extensively investigated over the parameters used in the proposed method.

This paper extends our preliminary work [14] by performing new experiments with a well-established raw matching cost through careful parameter tuning. We also provide an in-depth analysis of some critical parameters of our algorithm and include an accuracy-complexity trade-off study. The reminder of this paper is organized as follows. In Section III, we describe a new formulation for the efficient cost aggregation and its approximation techniques. We then present experimental results in Section IV and summarize conclusion in Section V, respectively.

3 Efficient Cost Aggregation in Stereo

3.1 New Formulation for Likelihood Aggregation

For local approaches, the cost aggregation is the most important yet time-consuming part. Given two images I_l and I_r, we define a function $e^h(p,d)$ which represents how likely a pixel p is to have a specific disparity hypothesis d. For instance, it could be defined using the TAD as $e^h(p,d) = \max(\sigma - |I_l(x,y) - I_r(x-d,y)|, 0)$. Note that other metrics such as the rank/census transform or NCC can also be utilized in a way that it is likely to have a large value as the disparity hypothesis d approaches a true disparity value.

To yield a reliable likelihood function, we implicitly consider a smoothness constraint by utilizing a color-weighted adaptive likelihood aggregation. The aggregated likelihood $E^h(p,d)$ can be then formulated using the matching window $\mathcal{N}(p)$ in a similar manner to (2) as

$$E^h(p,d) = \sum_{q \in \mathcal{N}(p)} w(p,q) e^h(q,d). \tag{4}$$

After applying the same aggregation procedure, the output disparity value $D_l(p)$ is estimated by seeking the *maximum* value of the aggregated likelihood $E^h(p,d)$, which is the same as the solution of (3). Note that in the likelihood aggregation, the normalization term $\sum w(p,q)$ is omitted, unlike (2). This modification does not affect the accuracy of the likelihood aggregation, since the disparity value $D_l(p)$ is estimated for each pixel independently where this normalization term is fixed for all ds [15].

The aggregated likelihood function $E^h(p,d)$ 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

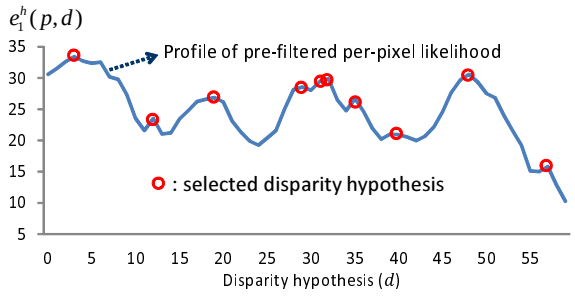

Fig. 1. Disparity candidate selection with local/global maxima.

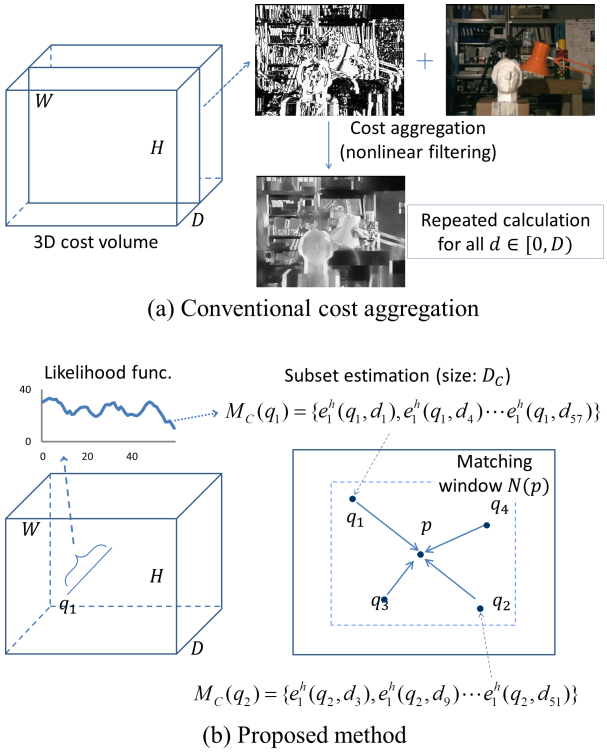

Fig. 2. Cost aggregation: (a) conventional approaches perform nonlinear filtering with (or without) a color image for all disparity hypotheses: $O(NBD)$. (b) The proposed method estimates the subset of disparity hypotheses, whose size is $D_c(\ll D)$, and then performs joint histogram-based aggregation: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

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]:

$$w(p,q) = exp\left(-||I_p - I_q||/\sigma_I - ||p-q||/\sigma_S\right) \ .$$

Since the color similarity is measured by using a corresponding color image, it shares the similar principle to the joint bilateral filtering [5], 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 complex form in stereo matching. Yu *et al.* [16] proposed a novel envelope point transform (EPT) method by applying a principal components analysis (PCA) to compress messages used in belief propagation [17]. Wang *et al.* [18] estimated the subset of disparity hypotheses for reliably matched pixels and then propagated them on an MRF formulation for estimating the subset of unreliable pixels. Yang *et al.* [19] proposed the method for reducing the search range and applied it into hierarchical belief propagation [20]. 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 (4). The likelihood 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 [21]. 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 predefined 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 likelihood $E^h(p,d)$ is defined with the subset of disparity hypotheses only:

$$E^h(p,d) = \sum_{q \in \mathcal{N}(p)} w(p,q) e_1^h(q,d) o(q,d)$$
$$o(q,d) = \begin{cases} 1 & d \in M_C(q) \\ 0 & otherwise \end{cases}, \quad (5)$$

where $M_C(q)$ is a subset of disparity hypotheses whose size is D_c. Note that $M_C(q)$ varies from pixel to pixel. e_1^h represents the prefiltered likelihood with a 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 (N) and disparity hypotheses (D), so the complexity is $O(NBD)$. 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 $O(NBD_c)$. 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.

Fig. 3 shows the accuracy of the disparity candidate selection in the non-occluded region of 'Teddy' image according to the number of disparity hypotheses D_c. It was calculated by counting the number of pixels whose subsets actually include a ground truth disparity value. When $D_c = 60$, namely the same as the original size, the subsets of all pixels include the ground truth disparity value. Interestingly, when $D_c = 6$, only 91.8% pixels contain the ground truth disparity values in their subsets, but the accuracy of the estimated disparity map (94.1%) is almost similar to these of the best one (94.2%, when $D_c = 5$) or slightly better than the disparity map estimated with all the disparity hypotheses (93.7%, 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.

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

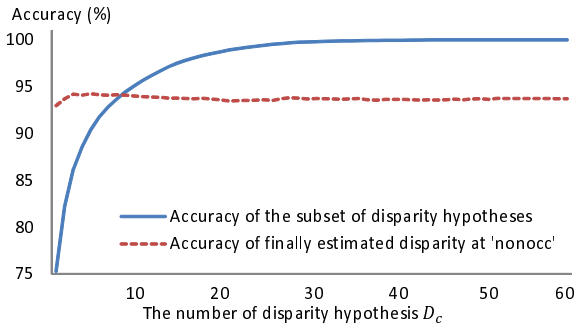

Fig. 3. Accuracy of the disparity candidate selection and the finally estimated disparity map in the non-occluded regions of 'Teddy' according to D_c. The accuracy of the selection process was measured by counting the number of pixels whose subsets actually include a ground truth disparity value.

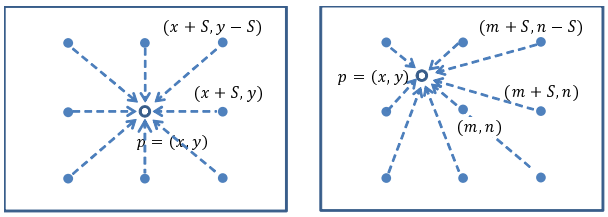

Fig. 4. Spatial sampling of matching window: (a) reference pixel-dependent, (b) reference pixel-independent sampling. A neighboring pixel q=(m,n) is sampled inside an image independently regardless of a reference pixel p=(x,y).

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 (5). 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. 4, there are two ways for spatial sampling: reference pixel-*dependent* and *independent* sampling. The dependent sampling can be defined as follows:

$$E^h(p,d) = \sum_{q \in \mathcal{N}(p)} w(p,q) e_1^h(q,d) o(q,d) s_1(p,q)$$
$$s_1(p,q) = \begin{cases} 1 & ||p-q||\%S = 0 \\ 0 & otherwise \end{cases}, \quad (6)$$

where $s_1(p,q)$ is a binary function capturing the regularly-sampled pixels inside the matching window for a sampling ratio S. $p\%S = 0$ denotes a pixel whose x and y coordinates are both multiples of S. As previously mentioned, the prefiltering with 5×5 window is applied into the per-pixel likelihood function for suppressing noise in the disparity candidate selection. Note that in this sampling strategy, regardless of the spatial sampling ratio S, the likelihood function $e_1^h(p,d)$ for all the pixels should be estimated by using the disparity candidate selection, which consists of dissimilarity measure, box filtering, and local maxima estimation/sorting. It leads to relatively high complexity compared to the joint histogram-based aggregation.

The reference pixel-independent sampling can solve this problem. As shown in Fig. 4 (b), our new sampling scheme can be defined as follows:

$$E^h(p,d) = \sum_{q \in \mathcal{N}(p)} w(p,q) e_1^h(q,d) o(q,d) s_2(q)$$
$$s_2(q) = \begin{cases} 1 & q\%S = 0 \\ 0 & otherwise \end{cases}, \quad (7)$$

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.

TABLE 1
Pseudo code for efficient likelihood aggregation.

Parameter definition

N: The size of an image I
B: The size of matching window $\mathcal{N}(p)$ (=$W \times W$)
M_D: The set of disparity hypotheses whose size is D
M_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: $O(25ND/S^2)$
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 ds.
  For *all disparity candidates $d \in M_D(p)$*
    For *all neighboring pixels which satisfy $||p-q||_\infty \leq 2$*
      2: Compute per-pixel likelihood $e^h(q,d)$ and $e_1^h(p,d) += e^h(q,d)$ (5×5 box filtering)
    End
  End
  3: Estimate $M_C(p)$ with the local maxima on $e_1^h(p,d)$
End

JOINT HISTOGRAM-BASED AGGREGATION
Complexity: $O(NBD_c/S^2)$
For *all reference pixels $p \in I$*
  4: Initialize likelihood function $E^h(p,d)$ to 0 for all ds.
  For *neighboring pixels which satisfy $|q_1|_\infty \leq W/2S$*
    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-independent sampling)
    For *all disparity candidates $d_q \in M_C(q)$*
      6: $E^h(p,d_q) += w(p,q) \times e_1^h(q,d_q)$
    End
  End
  7: $D_l(p) = \arg\max_{d \in [0,\cdots,D-1]} E^h(p,d)$
End

4 EXPERIMENTAL RESULTS

We compared the performance of the proposed method with state-of-the-arts methods in the Middlebury test bed [22]. All the experiments were performed on a computer containing an Intel Xeon 2.8-GHz CPU (using a single core only) and a 6-GB RAM. The proposed stereo matching method is 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 1.5 and 17.0, respectively. The size of the matching window $\mathcal{N}(p)$ is set to 31×31 for the stereo matching. 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. Finally, a weighted median filter (WMF) is applied to the disparity maps for better boundary handling. It is applied across the discontinuities regions only, and thus its computational load is negligible (e.g. 15ms for 'Tsukuba'). We found this post-processing achieves a small improvement on the discontinuities regions.

The per-pixel likelihood function $e^h(p,d)$ was measured by using both the TAD of the color images and their gradient as

$$e^h(p,d) = \alpha \cdot \max(\lambda_c - ||I_l(x,y) - I_r(x-d,y)||, 0) + (1-\alpha) \cdot \max(\lambda_g - |\nabla_x I_l(x,y) - \nabla_x I_r(x-d,y)|, 0) \quad (8)$$

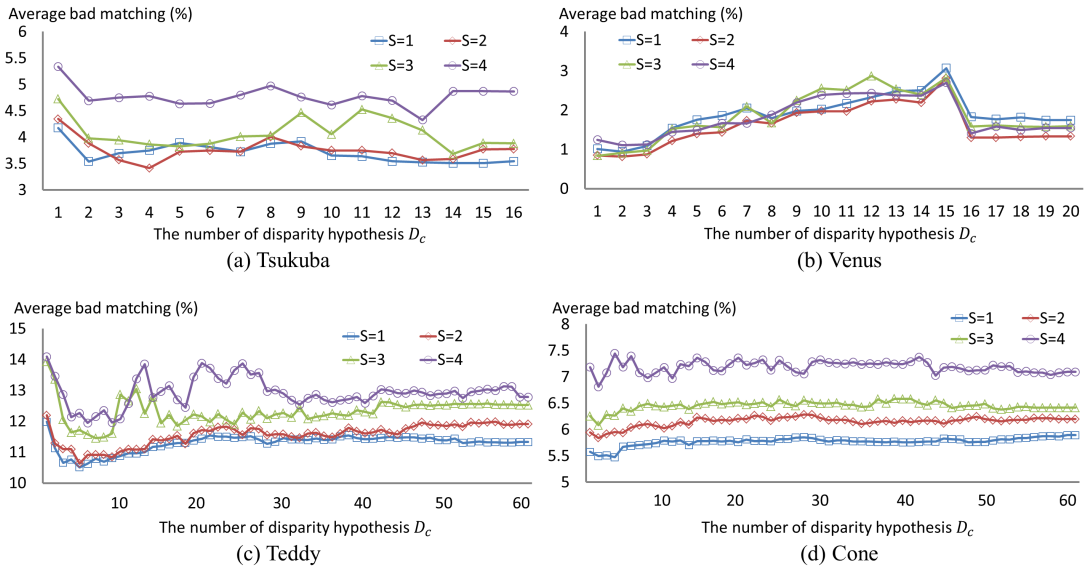

Fig. 5. Performance evaluation: average percent (%) of bad matching pixels for 'nonocc', 'all' and 'disc' regions according to D_c and S.

where α is a parameter controlling the influence of two (color and gradient) terms, which are truncated with λ_c and λ_g, respectively. It has been known that this model is more robust against illumination variation [23]. We will show that this per-pixel likelihood function combining the image gradient significantly improves a depth accuracy over our previous work [14]. For all experiments, we set α, λ_c, and λ_g to 0.11, 13.5, and 2.0, respectively. Note that (8) is likely to become large as d approaches a true disparity value.

Fig. 5 shows an performance evaluation according to the number of depth candidate D_c and the spatial sampling ratio S. The average percent (%) of bad pixels (APBP) 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 hypotheses D_c increases. It indicates that using the information of all the disparity hypotheses does not necessarily guarantee to obtain 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. Interestingly, in the 'Venus' image, the results of using $S = 2$ showed slightly better than those of $S = 1$. The 'Venus' image consists of a few planar surfaces only, which are simple and easy to be estimated compared to the 'Teddy' and 'Cone' images, and thus the effect of the spatial sampling in the joint histogram

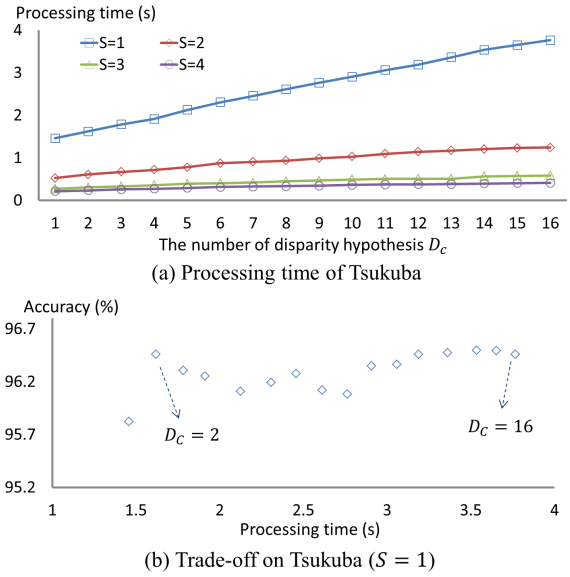

Fig. 6. Processing times (a) and trade-off (b) of the proposed method according to D_c and S. The results of 'Tsukuba' image only are shown due to the lack of space. In (b), the 'Accuracy' means (100%−APBP). One can find that the accuracy is not monotonically increasing as the processing time (D_c) increases.

based aggregation would be relatively marginal.

Next, we investigated the trade-off between the accuracy and the complexity by comparing processing times in Fig. 6. We showed the results of 'Tsukuba' ($S = 1$) only, and other results also show similar behaviors. Note that the proposed method was implemented on the CPU only. The processing time was measured for the calculation of a single (left or right) disparity map. As expected,

TABLE 2
Performance evaluation of disparity accuracy for local stereo matching methods

Algorithm	*Tsukuba*			*Venus*			*Teddy*			*Cone*			APBP (%)
	nocc	all	disc	nocc	all	disc	nocc	all	disc	nocc	all	disc	
Our method (BEST)	1.44	1.78	7.01	0.14	0.38	1.92	5.75	11.2	14.6	2.29	7.62	6.52	5.05
Our method (S=1, D_c=10%)	1.93	2.30	6.39	0.16	0.46	2.22	5.88	11.3	14.7	2.41	7.78	6.89	5.20
PatchMatch [24]	2.09	2.33	9.31	0.21	0.39	2.62	2.99	8.16	9.62	2.47	7.80	7.11	4.59
CrossLMF-1 [25]	2.46	2.78	6.26	0.27	0.38	2.15	5.50	10.6	14.2	2.34	7.82	6.80	5.13
Our method (S=2, D_c=10%)	2.09	2.44	7.12	0.14	0.38	1.92	5.92	11.3	15.5	2.70	7.93	7.48	5.41
CostFilter [12]	1.51	1.85	7.61	0.20	0.39	2.42	6.16	11.8	16.0	2.71	8.24	7.66	5.55
NonLocalFilter [26]	1.47	1.85	7.88	0.25	0.42	2.60	6.01	11.6	14.3	2.87	8.45	8.10	5.48
GeoSup [27]	1.45	1.83	7.71	0.14	0.26	1.90	6.88	13.2	16.1	2.94	8.89	8.32	5.80
P-LinearS [28]	1.10	1.67	5.92	0.53	0.89	5.71	6.69	12.0	15.9	2.60	8.44	6.71	5.68
Our method (S=3, D_c=10%)	2.11	2.38	7.45	0.17	0.39	2.17	6.53	11.9	16.3	2.91	8.16	7.97	5.70
GeoDif [11]	1.88	2.35	7.64	0.38	0.82	3.02	5.99	11.3	13.3	2.84	8.33	8.09	5.49
RecursiveBF [29]	1.85	2.51	7.45	0.35	0.88	3.01	6.28	12.1	14.3	2.80	8.91	7.79	5.68
DistinctSM [30]	1.21	1.75	6.39	0.35	0.69	2.63	7.45	13.0	18.1	3.91	9.91	8.32	6.14
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
FastBilateral [31]	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
HistoAggr [Our prev. work] [14]	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 [7]	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
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

the processing time is proportional to the number of disparity hypotheses D_c, and inversely proportional to the square of the sampling ratio 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) shows that the accuracy ($100\%-$APBP) is not monotonically increasing as the processing time (D_c) increases.

The performance evaluation from the Middlebury test bed is shown in Table 2 by reporting a comparison with other state-of-the-art methods. All the leading local methods were sorted with the average ranking listed in the Middlebury test bed. The number of disparity candidates D_c are all set to 10% of the original search range. We measured the disparity accuracy with the varying spatial sampling ratio S ($1 \sim 3$). We found that the proposed method with $S = 1$ achieved the best accuracy among all the leading local stereo matching methods. In comparison with all stereo matching methods, the average ranking of the current results (S=1, D_c=10%) is 16^{th}, while the ranking of the initial results from our previous work [14] is 76^{th}. As mentioned earlier, it is mainly due to the use of the raw matching cost combining the TAD of the color images and their gradient as in (8). '**Our method (BEST)**' represents the result when the parameters (S and D_c) that provide the disparity maps with the best accuracy are used. Interestingly, using 10% of original search range ($S = 1$) produces the result which is nearly close to the best disparity quality.

For the comparison of the complexity, we referred to the results reported in the recent work [25]. We have optimized our C implementation for both 'CrossLMF-0/1' [25] and 'CostFilter' [12] and achieved accelerated runtime since the publication of [25]. 'CrossLMF-0' and 'CrossLMF-1' represent cross-based local multi-point filtering methods [25] using the zero-order and first-order polynomial models, respectively. To analyze the trade-off between complexity and accuracy, we list both the processing time and the APBP (%) for some representative local stereo matching methods in Table 3. 'CostFilter' and 'P-LinearS' used the guided filter (GF) [13] for efficient cost aggregation. Note that the processing time on a single core CPU was measured for 'Tsukuba', and the average error was calculated for all the test sequences. The processing time of the proposed method also includes the post-processing such as occlusion detection/handling and WMF, while some of the previous works consider the cost aggregation only. As already explained in Fig. 5, the disparity results ($S = 1$, $D_c = 10\%$) estimated using only 10% of original search range are better than those ($S = 1$, $D_c = 100\%$) estimated using all disparity candidates.

TABLE 3
Comparison with other methods (as of Oct. 2012): the runtime was measured for 'Tsukuba'.

Algorithms	Rank	APBP	Runtime
Our method (S=1, D_c=10%)	**16**	**5.20** %	**1.38** s
CrossLMF-1 [25]	18	5.13 %	0.50 s
CrossLMF-0 [25]	19	5.24 %	0.21 s
Our method (S=2, D_c=10%)	**23**	**5.41** %	**0.52** s
CostFilter (GF) [12]	24	5.55 %	0.48 s
P-LinearS (GF) [28]	33	5.68 %	33.0 s
Our method (S=1,D_c=100%)	**36**	**5.63** %	**3.50** s
Our method (S=3, D_c=10%)	**37**	**5.70** %	**0.28** s
AdaptWeight [4]	73	6.67 %	60.0 s

Fig. 7 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 1. 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.

To analyze the effect of prefiltering the likelihood function $e^h(p, d)$ in the disparity candidate selection,

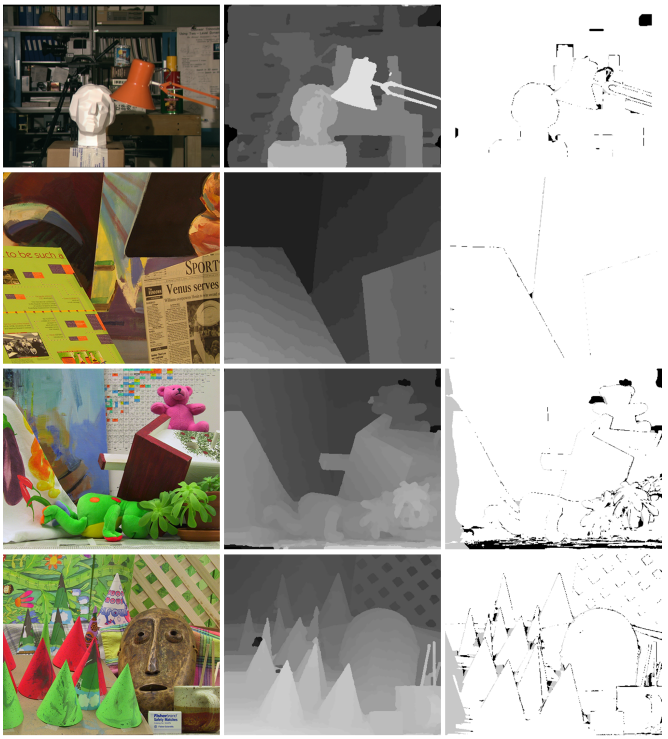

Fig. 7. Results for (from top to bottom) 'Tsukuba', 'Venus', 'Teddy' and 'Cone' image pairs: (a) original images, (b) our results, (c) error maps. The number of disparity hypotheses D_c is set to 10% of the original search range and the spatial sampling ratio S is set to 1.

we measured the accuracy of disparity maps obtained when applying the box filtering with varying window sizes to e^h. Table 4 shows the APBP for the Middlebury test sequences (S=1 and D_c=10%). The result with no prefiltering (1×1) shows serious performance degeneration. As the size of the box filter increases, the method produces better quality but using too large box windows (7×7, 9×9) deteriorates the quality, and incurs more computational overhead. Note that while this pre-filtering can be seen as the first cost aggregation step, it mainly serves the removal of noise from the per-pixel likelihood functions.

One interesting fact is that the proposed 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 4
Effect of the prefiltering in the disparity candidate selection.

Window size	1×1	3×3	5×5	7×7	9×9
APBP (%)	9.56 %	5.36 %	5.20 %	5.46 %	5.79 %

5 Conclusion

In this paper, we have presented a novel approach for the efficient cost aggregation used in the stereo matching. Given the per-pixel likelihood (evidence) function, we re-formulated the problem from the perspective of the relaxed joint histogram. Two algorithms were then proposed for reducing the complexity of the joint histogram-based aggregation. Different from the conventional local approaches, we reduce the complexity in terms of the search range by estimating a subset of informative disparity hypotheses. We showed that the reliable disparity maps were obtained even when the number of labels hypotheses (D_c) was about 10% of the original full search range. In addition, 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 label hypotheses. As shown in Fig. 5, the optimal number of disparity hypotheses may be 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 label 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. Another further research would be to extend the method to an optical flow estimation.

Acknowledgments

This study is supported by the research grant for the Human Sixth Sense Programme at the Advanced Digital Sciences Center from Singapore's Agency for Science, Technology and Research (A*STAR).

References

[1] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47, pp. 7–42, 2001.
[2] S. Birchfield and C. Tomasi, "Depth discontinuities by pixel-to-pixel stereo," *International Journal of Computer Vision*, vol. 35, no. 3, pp. 269–293, 1999.
[3] R. Zabih and J. Woodfill, "Non-parametric local transforms for computing visual correspondence," in *ECCV (2)*, 1994, pp. 151–158.
[4] K.-J. Yoon and I.-S. Kweon, "Adaptive support-weight approach for correspondence search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 650–656, 2006.
[5] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," *ACM Trans. Graph.*, vol. 26, no. 3, p. 96, 2007.
[6] F. Tombari, S. Mattoccia, L. di Stefano, and E. Addimanda, "Near real-time stereo based on effective cost aggregation," in *IEEE Int. Conf. on Pattern Recognition*, 2008, pp. 1–4.

[7] K. Zhang, J. Lu, and G. Lafruit, "Cross-based local stereo matching using orthogonal integral images," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 19, no. 7, pp. 1073–1079, 2009.
[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, 2008.
[9] C. Richardt, D. Orr, I. P. Davies, A. Criminisi, and N. A. Dodgson, "Real-time spatiotemporal stereo matching using the dual-cross-bilateral grid," in *European Conf. on Computer Vision*, 2010, pp. 510–523.
[10] S. Paris and F. Durand, "A fast approximation of the bilateral filter using a signal processing approach," *International Journal of Computer Vision*, vol. 81, no. 1, pp. 24–52, 2009.
[11] L. De-Maeztu, A. Villanueva, and R. Cabeza, "Near real-time stereo matching using geodesic diffusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 2, pp. 410–416, 2012.
[12] C. Rhemann, A. Hosni, M. Bleyer, C. Rother, and M. Gelautz, "Fast cost-volume filtering for visual correspondence and beyond," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2011, pp. 3017–3024.
[13] K. He, J. Sun, and X. Tang, "Guided image filtering," in *European Conf. on Computer Vision*, 2010, pp. 1–14.
[14] D. Min, J. Lu, and M. N. Do, "A revisit to cost aggregation in stereo matching: How far can we reduce its computational redundancy?" in *IEEE Int. Conf. on Computer Vision*, 2011, pp. 1567–1574.
[15] J. Ding, J. Liu, W. Zhou, H. Yu, Y. Wang, and X. Gong, "Real-time stereo vision system using adaptive weight cost aggregation approach," *EURASIP J. Image and Video Processing*, vol. 2011, p. 20, 2011.
[16] T. Yu, R.-S. Lin, B. J. Super, and B. Tang, "Efficient message representations for belief propagation," in *IEEE Int. Conf. on Computer Vision*, 2007, pp. 1–8.
[17] J. Sun, N. Zheng, and H.-Y. Shum, "Stereo matching using belief propagation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 7, pp. 787–800, 2003.
[18] L. Wang, H. Jin, and R. Yang, "Search space reduction for MRF stereo," in *European Conf. on Computer Vision*, 2008, pp. 576–588.
[19] Q. Yang, L. Wang, and N. Ahuja, "A constant-space belief propagation algorithm for stereo matching," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2010, pp. 1458–1465.
[20] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient belief propagation for early vision," *International Journal of Computer Vision*, vol. 70, no. 1, pp. 41–54, 2006.
[21] C. Dima and S. Lacroix, "Using multiple disparity hypotheses for improved indoor stereo," in *IEEE Int. Conf. on Robotics and Automation*, 2002, pp. 3347–3353.
[22] http://vision.middlebury.edu/stereo.
[23] T. Brox and J. Malik, "Large displacement optical flow: Descriptor matching in variational motion estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 500–513, 2011.
[24] M. Bleyer, C. Rhemann, and C. Rother, "Patchmatch stereo - stereo matching with slanted support windows," in *Proc. BMVC*, 2011, pp. 14.1–14.11.
[25] J. Lu, K. Shi, D. Min, L. Lin, and M. N. Do, "Cross-based local multipoint filtering," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2012, pp. 430–437.
[26] Q. Yang, "A non-local cost aggregation method for stereo matching," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2012, pp. 1402–1409.
[27] A. Hosni, M. Bleyer, M. Gelautz, and C. Rhemann, "Local stereo matching using geodesic support weights," in *IEEE Int. Conf. on Image Processing*, 2009, pp. 2093–2096.
[28] L. De-Maeztu, S. Mattoccia, A. Villanueva, and R. Cabeza, "Linear stereo matching," in *IEEE Int. Conf. on Computer Vision*, 2011, pp. 1708–1715.
[29] Q. Yang, "Recursive bilateral filtering," in *European Conf. on Computer Vision*, 2012.
[30] K.-J. Yoon and I.-S. Kweon, "Stereo matching with the distinctive similarity measure," in *IEEE Int. Conf. on Computer Vision*, 2007, pp. 1–7.
[31] S. Mattoccia, S. Giardino, and A. Gambini, "Accurate and efficient cost aggregation strategy for stereo correspondence based on approximated joint bilateral filtering," in *Asian Conference on Computer Vision*, 2009, pp. 371–380.