Computer Vision and Image Understanding 135 (2015) 95–108

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Adaptive stereo similarity fusion using confidence measures ☆

Gorkem Saygili *, Laurens van der Maaten, Emile A. Hendriks

Vision Lab, Pattern Recognition and Bioinformatics Group, Delft University of Technology, Mekelweg 4, 2628CD Delft, The Netherlands

ARTICLE INFO

Article history:
Received 6 May 2014
Accepted 9 February 2015
Available online 17 February 2015

Keywords:
Stereo confidence measures
Stereo similarity measure fusion
Stereo matching

ABSTRACT

In most stereo-matching algorithms, stereo similarity measures are used to determine which image patches in a left–right image pair correspond to each other. Different similarity measures may behave very differently on different kinds of image structures, for instance, some may be more robust to noise whilst others are more susceptible to small texture variations. As a result, it may be beneficial to use different similarity measures in different image regions. We present an adaptive stereo similarity measure that achieves this via a weighted combination of measures, in which the weights depend on the local image structure. Specifically, the weights are defined as a function of a confidence measure on the stereo similarities: similarity measures with a higher confidence at a particular image location are given higher weight. We evaluate the performance of our adaptive stereo similarity measure in both local and global stereo algorithms on standard benchmarks such as the Middlebury and KITTI data sets. The results of our experiments demonstrate the potential merits of our adaptive stereo similarity measure.

1. Introduction

The objective of stereo matching is to estimate the depth of a scene based on two rectified images that are obtained from two cameras. The resulting depth image can be used in algorithms for problems such as 3D reconstruction [1] and virtual view rendering [2].

Stereo matching has been extensively researched over the past decades [3]. There are two types of stereo algorithms: sparse and dense algorithms. Sparse algorithms employ feature-based methods that match key-point locations. The resulting depth map is sparse since there are locations without depth estimation [4,5]. In contrast, dense stereo algorithms produce depth estimations at every pixel in an image using pixel-wise matching between stereo views. Dense stereo matching algorithms can be further grouped into two classes, namely, global and local stereo matching. Global stereo matching algorithms make global smoothness assumptions on the disparity image; this generally leads to high-quality depth estimates but it is computationally expensive [6–8]. By contrast, local stereo matching algorithms do not employ such smoothness constraints; as a result, they are computationally cheap but provide disparity estimates of lower quality [37,38].

Stereo algorithms estimate the disparity of a scene by matching pixels/patches between the two rectified images. In general, they are composed of four steps: (1) cost initialization, (2) cost aggregation, (3) disparity selection, and (4) disparity refinement [3]. The first step is cost initialization, which is performed by matching pixels of the two rectified images. The resulting cost space is called the initial disparity space image (DSI). In the matching, a variety of different similarity measures can be used, each of which may have different characteristics. In the second step, cost aggregation is applied on the initial DSI to filter out noisy matches that may have arisen in the first step. In order to select a disparity for each pixel, most of the algorithms use a winner-take-all approach as the third step. In the winner-take-all approach, the disparity that has the lowest matching cost at a particular image location is selected as the disparity for that location. The fourth (optional) step aims to refine the resulting disparity map by filtering out wrong matches using global smoothness assumptions.

An important step in both global and local stereo matching algorithms is the cost initialization. In both types of algorithms, it is essential to obtain high-quality initial DSIs in order to obtain good disparity maps [10–12]. Since stereo similarity measures may perform differently depending on texture and noise variations, the quality of the initial DSI may be improved by using different stereo similarity measures for different parts of the image. One way in which this can be achieved is by fusing similarities adaptively based on their performance on different regions of the image. (Note that adaptive fusion is very different than the non-adaptive fusion scheme of [10,11,13].) In this paper, we propose a new algorithm that fuses similarity measures adaptively based on their performance on different image regions. Figs. 1 and 2 show two examples of the output of our algorithm, which illustrate

☆ This paper has been recommended for acceptance by Y. Aloimonos.
* Corresponding author.

http://dx.doi.org/10.1016/j.cviu.2015.02.005
1077-3142/

that adaptive fusion of two similarities may enhance the accuracy of final disparity estimations substantially.

The main contribution of this paper is a new algorithm for adaptive fusion of stereo similarity measures. Our adaptive fusion strategy can fuse any number of stereo similarity measures without introducing any similarity measure-dependent parameters. Our fusion algorithm operated by evaluating the local performance of each similarity measure at each image location using stereo confidence measures [15,14,16–18]. The resulting confidence measures are used to determine the weight of each of the stereo similarity measures: more confident similarity measures are assigned a higher weight in the final ensemble. The fusion and confidence estimation are performed on the initial DSI, which contains the pixel-wise matching scores. To exploit spatial correlations and increase the robustness of our fusion algorithm, we use a local consensus between neighboring pixels when computing the weights. Each neighboring pixel votes for the disparity of the center pixel proportional to its confidence. The disparity that has the highest consensus for the center pixel is chosen as the disparity of the center pixel. The DSIs of the similarities that support the consensus disparity are adaptively aggregated based on their confidences in order to obtain the final fused initial cost measures.

The remainder of the paper is organized as follows: In Section 2, we review existing literature that relates to similarity measures and similarity measure fusion. In Section 3, we present our novel adaptive fusion strategy. The experimental results are presented in Section 4. We discuss our results and possible directions for future work in Section 5, and draw conclusions in Section 6.

2. Related work

Most of the recent stereo research focuses on increasing the accuracy of depth estimation by enhancing the performance of aggregation methods [38,21,22]. At the same time, significant improvements in performance have also be obtained by using enhanced similarity measures [23–25] or by fusing multiple similarity measures [10–12].

Fusion of multiple stereo similarities has been used in several stereo algorithms. Wegner and Stankiewicz [12] took the multiplication of any two similarity measures and obtain quality enhancements for their view synthesis algorithm. However, multiplication of two similarity measures may fail to increase the accuracy if at least one of the measures fail to find correct disparity. Klaus et al. [10] fused the gradient and intensity similarities using weighted-summation and obtained significant increase in accuracy. The weight is found experimentally and is not generalizable to different similarity measures (it uses a similarity measure-dependent parameter). Mei et al. [11] proposed fusing Census transform and color measures by using exponential functions and weighted averaging, which led to substantially better performances. Different from the above-mentioned algorithms Stentoumis et al. [13] fused three similarity measures with constant weights using exponential functions. However, for all of these algorithms, the parameters of their fusion strategy are found experimentally, they are static for all pixels, and are not generalizable to all kinds of similarities. In contrast to this prior work, our adaptive fusion strategy can fuse any number of similarity measures and does not require any similarity measure-dependent parameters. Moreover, it fuses similarities adaptively to exploit the benefits of different similarities in different regions of the image.

To fuse different similarity measures, it is important to observe which of the fused similarities perform better than the others at a particular location of the image. Stereo confidences are used to measure the confidence of stereo matching and to filter out the wrong estimations. The matching is considered to be more accurate as its confidence increases. An extensive evaluation of stereo confidences are presented in [15]. The confidence measures is frequently used at the refinement step of stereo matching to refine the wrong matches [14,16–18]. Stereo confidences have further been tested in applications producing stixel-world representations [28]. In this work, we incorporate top performing stereo confidence to assess the performances of various stereo similarities at different locations of the image. This allows to compute adaptive weights for the fusion of any number of stereo similarity measures.

Stereo matching is not always applied between two regular camera views. One of the most challenging stereo matching problem is cross-modal stereo matching where the matched images differ in terms of their data structure such as stereo matching between IR and RGB images [40,41]. Simple stereo similarity measures such as intensity, color, and gradient values of pixels in the image provide reasonable performances as long as the stereo

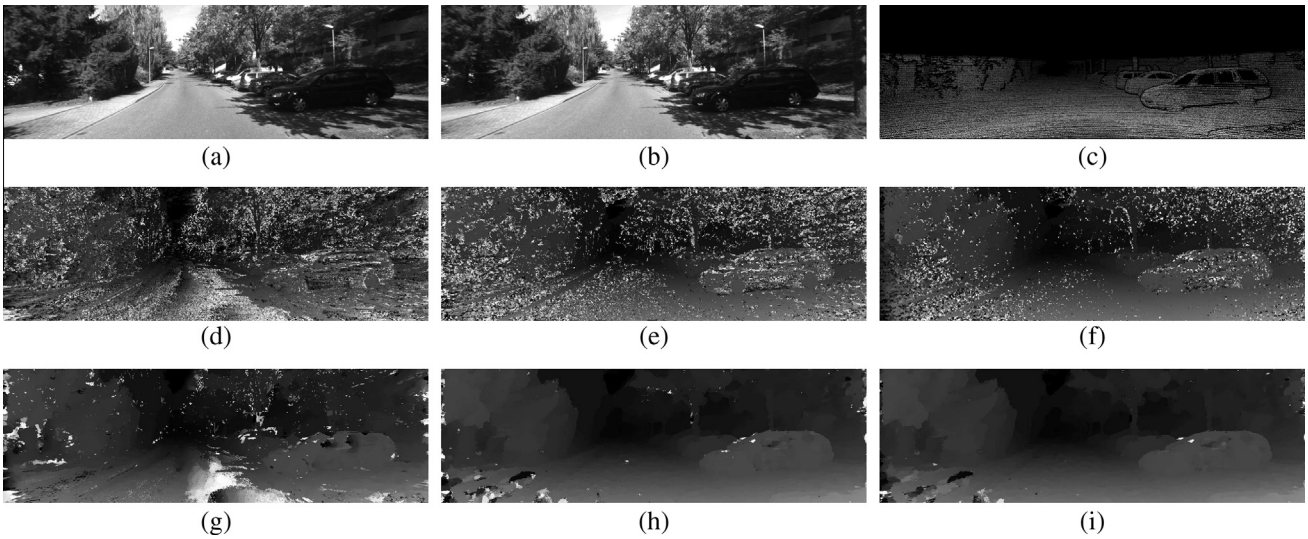

Fig. 1. Stereo results on example from the KITTI data set: (a) left image, (b) right image, (c) ground truth, initial similarity results: (d) Absolute intensity difference (AD), (e) Census, (f) Proposed (AD + Census), resulting disparities: (g) AD, (h) Census, (g) Proposed fusion strategy (AD + Census). The accuracy of the disparity maps changes substantially with the use of improved similarity measures.

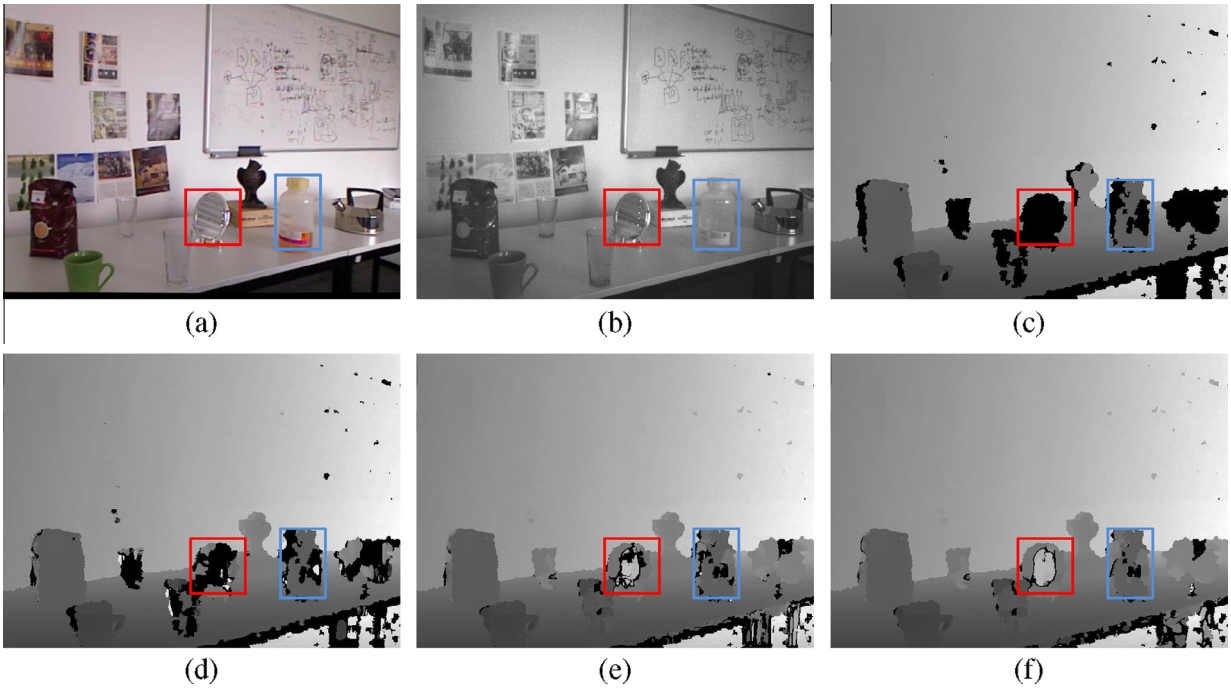

Fig. 2. Kinect cross-modal stereo results: (a) Color, (b) IR, (c) Raw depth, (d) AD, (e) Sobel, and (f) Proposed (AD + Sobel) fusion strategy. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

images are noise-free. However in the presence of noise and data structure difference (such as cross-modal stereo), the performance of such measures decreases significantly. Normalized cross-correlation (NCC) [26] is an intensity and patch-based similarity measure that is robust against Gaussian noise in the two stereo images. Zabih and Woodfill [23] introduced the rank and census transform measures in order to increase robustness against radiometric difference between stereo pairs. Egnal [27] used mutual information (MI) as a stereo similarity measure, which leads to a matching algorithm that is robust under radiometric differences. Their study showed that both MI and NCC provide reasonably good performances. Hirschmuller and Scharstein [20] provided an extensive evaluation study of the aforementioned similarity measures. The results of this evaluation show that the performance of different measures highly varies with the image structure, which provides a motivation for our adaptive similarity measures. In this work, we show that the performance of cross-modal stereo algorithms may be improved by the use of adaptive fusion algorithms.

3. Stereo matching with similarity fusion

A similarity measure may perform better than other measures in a particular location in a stereo image depending on the kind of image structure present at that location (such as at depth discontinuities, regular texture, and nearly homogeneous image regions), whilst performing worse in other regions. For instance, filter-based measures tend to blur object boundaries whereas other measures do not [19]. As a result, the performance of stereo matching may be improved by adaptively combining stereo similarity measures with different characteristics according to their (estimated) performance at a particular location. Below, we present an adaptive fusion strategy that combines multiple stereo similarities based on their performance in different image regions.

An overview of our adaptive fusion algorithm is shown in Fig. 3. Our algorithm consists of two main stages. First, we calculate pixel-wise matching scores as in most other stereo matching algorithms. This stage provides the initial cost measures for each

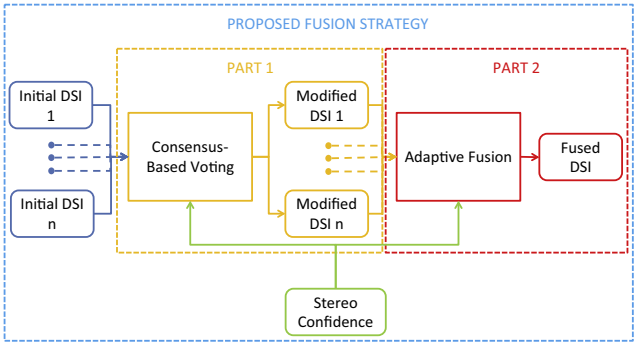

Fig. 3. The flowchart of the proposed adaptive fusion algorithm.

disparity per pixel, the so-called disparity space image (DSI). The initial matching is done using each measure from a set of similarity measures that are explained in detail in Appendix A. Second, we compute the confidence of the matching at every image location for each of the similarity measures using the confidence measures presented by [15]. We use the confidence to compute weights for all stereo similarity measures at all image locations. Our final stereo similarity measure is given by a locally weighted combination of the stereo similarities. We discuss both stages of our algorithm in more detail below.

The first part of our fusion strategy consists of building a modified initial DSI for each similarity measure that is based on our confidence-based voting. To obtain more robust matching scores and exploit spatial correlation without using aggregation over the cost space, we construct a consensus set, $H(x,y,d)$, that is built for each pixel, (x,y). Our fusion strategy incorporates a stereo confidence, $S_i(x,y)$, to build the consensus and to weight the costs for the fusion where i denote a similarity. Let (x_n,y_n) be the pixels around a neighbourhood $N(x,y)$ with size h_w of a pixel (x,y) and let d_n be the winner disparity of (x_n,y_n). $H(x,y,d)$ is defined as:

$$\mu(d_n, d) = \begin{cases} 1, & d_n = d \\ 0, & \text{otherwise}, \end{cases} \quad (1)$$

$$H(x, y, d) = \sum_{\forall i} \sum_{(x_n, y_n) \in N(x,y)} S_i(x_n, y_n) \mu_i(d_n, d).$$

Each consensus set is built around a center pixel including the pixel itself. Each pixel in the consensus set votes for the consensus disparity of the set such that the weights of the votes are proportional to the confidences of the pixels. The consensus disparity, d^*, is selected by taking the disparity that received most votes:

$$d^* = \underset{d}{\operatorname{argmax}}(H(x, y, d)). \quad (2)$$

The fusion is applied to the DSIs of the similarities to obtain a single (fused) DSI for the aggregation step of the matching. Hence, the consensus disparity cannot be used directly in the fusion step of the stereo similarities. Moreover, the DSI of the center pixel in each consensus set does not necessarily support its consensus disparity. Therefore, to use the consensus result in the fusion step of the similarities, the initial DSIs of the center pixels should be modified according to their consensus disparity. To modify the initial DSIs of each center pixel such that it supports the consensus disparity without doing computationally expensive aggregation over the cost space, we assign the DSI of the pixel in the consensus set that favors the consensus disparity with the highest confidence for each similarities:

$$n^* = \underset{n}{\operatorname{argmax}}(S_i(x_n, y_n) \mu_i(d_n, d^*)), \quad (3)$$

$$C_i^*(x, y, d) = C_i(x_{n^*}, y_{n^*}, d). \quad (4)$$

In the second stage, the DSIs of all of the similarity measures $C_i^*(x, y, d)$ are aggregated adaptively using weighted average with respect to their confidence:

$$w_i = \frac{S_i(x, y)}{\sum_{\forall i} S_i(x, y)},$$

$$C_f(x, y, d) = \sum_{\forall i} w_i C_i^*(x, y, d), \quad (5)$$

where $C_f(x, y, d)$ is the fused DSI measure for the pixel at location (x, y).

It is important to note that our algorithm can be used to fuse any number of similarity measures without introducing any similarity measure-dependent hyperparameters.

4. Experiments

We evaluate the performance of our algorithm by performing several experiments on benchmark Middlebury [3] and KITTI [32] data sets. The Middlebury stereo dataset has images of size 400×375 approximately, whereas KITTI dataset images have sizes around 1200×370. We evaluate the performance of the stereo algorithms by measuring the percentage of disparity estimates that has a difference of more than one compared to the ground truth disparity. We also evaluate the number of errors in specific parts of images such as non-occluded parts and locations close to disparity discontinuities. Our algorithm has only one hyperparameter, which is the consensus window size, h_w.

The performance of our algorithm depends on the performance of the stereo confidence measure used. In Section 4.1, we investigate the effect of different stereo confidences on the performance of our fusion strategy; we select the best-performing confidence for the remaining experiments. Specifically, we performed experiments using four different stereo confidence measures. To test the effect of h_w and to select the optimum value, we performed experiments with our fusion strategy using different consensus window sizes in Section 4.2. In Section 4.3, we tested our fusion strategy with all stereo similarities and compare its accuracy with (non-adaptive) fusion strategies presented in prior work. Intensity is the cheapest stereo similarity measure to extract. Almost all works on fusion aim to fuse intensity with a more complex similarity measure to achieve higher accuracies. In Section 4.4, we also fuse intensity in a similar way with other similarity measures and compare our strategy's performance with that of other strategies. To further compare the performance of our algorithm with the fusion strategies presented in prior work, we evaluate our algorithm in terms of initial matching cost improvements by using constant window aggregation (Section 4.5) and global energy minimization (Section 4.7). For all of the above-mentioned experiments, we used the Middlebury data set [3]. The errors are measured for all pixels (all), pixels that are not occluded (non-occlusion), and pixels that are close to disparity discontinuities and not occluded (dis-occlusion). These regions are shown in Fig. 4. To test our fusion strategy on a larger, more challenging dataset, we also performed experiments on the KITTI data set, comparing the performance of our algorithm to that of a state-of-the-art local stereo matching algorithm (Section 4.8).

4.1. Performance of stereo confidence

We performed experiments to find out which confidence measure provides the best results in our fusion strategy. In this experiment, we set $h_w = 3$. We experimented with the LRD, PKRN, MLM and LC confidence measures that are described in Appendix A on the Middlebury data set. We fused all of the eleven stereo similarities that are presented in Appendix A. The results of our experiment are shown in Fig. 5. The results show that LRD outperforms other confidences for almost all images and regions. The difference is most noticeable in the Cones dataset where LRD substantially outperforms other confidences in discontinuous regions. In our further experiments, we used LRD as the confidence measure for our fusion strategy.

4.2. Experiments with different consensus window sizes

The only hyperparameter of our fusion strategy is the consensus size, h_w. Although it is a similarity measure-independent parameter, it may still effect the final fusion results. Therefore, we tested the performance of our algorithm using various consensus window sizes. Fig. 6 shows the error as a function of the value of h_w. As h_w increase, more neighboring pixels vote in the consensus and the accuracy increases. However as Fig. 6b shows, the performance near discontinuous regions decreases as h_w increases. Although it is possible to achieve higher accuracies with larger consensus window sizes, in this paper, we therefore opt to use a consensus region of 3×3 pixels in the rest of our experiments.

4.3. Fusion of all similarities

Although it may be computationally impractical to compute all of the similarities, we did perform an experiment fusing eleven similarities in order to see the overall performance of our fusion strategy. We tested our strategy against straightforward fusion methods. The methods that we compared with are:

- **Most** utilizes the initial DSI of the most confident similarity measure for each location.
- **Avg** is the fusion strategy that takes the average of all of the similarity measure's costs.
- **Mult** [12] is based on the multiplication of cost values of each similarity measures.
- **Conf** denotes the adaptive aggregation of initial DSI using the weights that are obtained from the LRD confidence of the similarity measures at each pixel. This is equivalent to the second part of our fusion strategy without the first part.

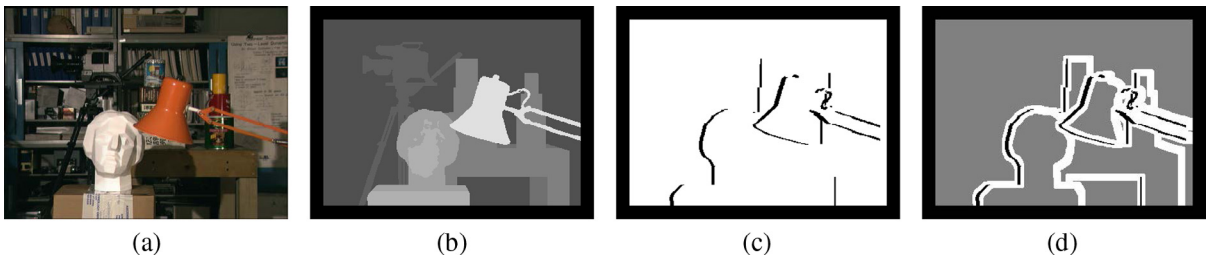

Fig. 4. (a) Color image, (b) ground truth disparity image, (c) non-occlusion pixels, and (d) dis-occlusion pixels, the pixels where the errors are evaluated are indicated in white. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

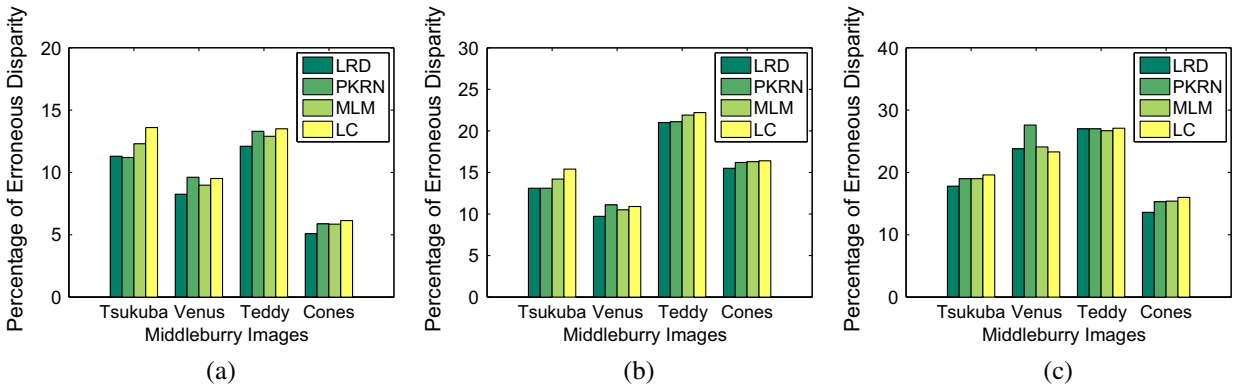

Fig. 5. Percentage of erroneous disparity values of four different stereo confidence measures in the fusion of eleven similarity measures with our fusion strategy in: (a) non-occluded, (b) all, and (c) discontinuous image regions.

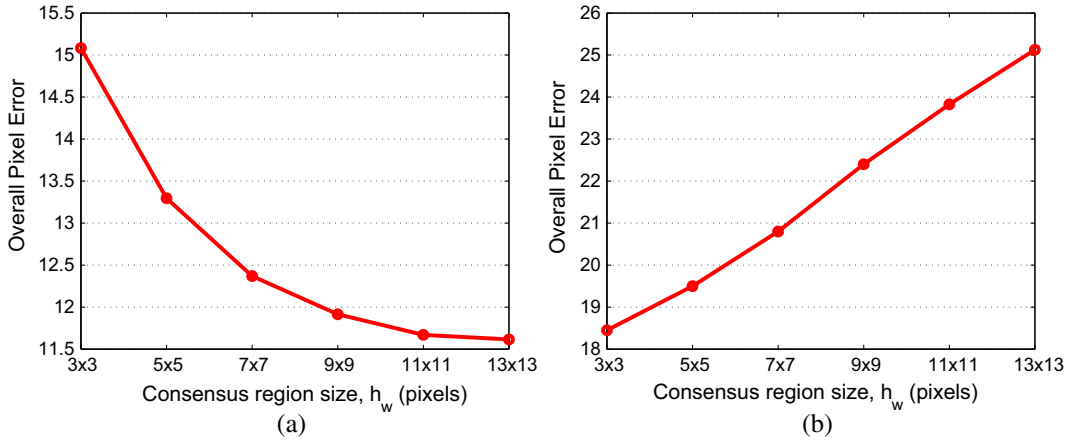

Fig. 6. The performance of our algorithm with respect to different consensus region sizes, h_w, for: (a) all pixels (all), (b) pixels at dis-occlusion regions (disc).

- **Voting** uses our consensus strategy but without confidences. The confidence in the disparity voting is always equal to 1. Furthermore, the cost functions of similarities are averaged rather than confidence-weighted aggregation.
- **Voting + Conf** is based on our consensus strategy and adaptive aggregation of the similarities using LRD confidence measure.

Table 1 shows the performance of eleven different similarity measures on the Middlebury image pairs. In general, the best performing similarity measure is the census transform. The census transform is also one of the most computationally expensive similarity measures. AD is computationally the cheapest similarity measure, and it still outperforms LoG and MI. Pfeiffer et al. [28] has shown that MI does not perform well without global smoothing. This result is confirmed by our results: MI performs poorly in the initial stereo matching stage. BT and Mean perform better than MI but not significantly better than intensity. NCC and ZNCC appear to be the best alternatives to the census transform.

Table 2 presents the results of combining all similarity measures with different fusion strategies. The best performer in all of the data set image pairs and in all image regions is our proposed strategy, Voting + Conf. Voting using our consensus algorithm also performs significantly better than all of the individual similarity measures. Even with the initial DSI, the Conf strategy performs better than the Avg strategy in discontinuous image regions, which illustrates the importance of using confidences in fusion. Furthermore, using Voting without confidences leads to worse accuracy, which also emphasize the effect of confidence on fusion. Directly choosing the highest confidence is not a good strategy for the fusion because it provides worse accuracy compared to most other

Table 1
Percentage of erroneous disparity values of individual similarity measures.

	Tsukuba			Venus			Teddy			Cones		
	nonocc	all	disc	nonocc	all	disc	nonocc	all	disc	nonocc	all	disc
AD	21.5	23.2	21.6	27.2	28.4	30.7	35.7	42.2	41.3	37.4	42.2	40.5
Rank	29.8	31.2	34.0	23.8	30.8	33.6	36.1	42.5	43.9	23.0	31.5	36.0
Census	17.1	18.8	**22.6**	**12.6**	**14.0**	**25.0**	**15.0**	**23.6**	**28.7**	**7.1**	**17.2**	**17.6**
NCC	**16.5**	**18.2**	26.6	13.4	14.8	30.4	16.7	25.2	33.7	9.7	19.9	25.9
ZNCC	18.6	20.3	27.2	13.8	15.3	32.3	18.6	27.0	36.4	10.5	20.6	26.3
Sobel	23.4	25.0	32.2	27.7	28.9	36.5	40.7	46.7	48.3	28.6	36.4	41.9
LoG	38.4	39.7	43.8	41.5	42.4	49.9	54.0	38.6	60.7	40.7	47.2	53.5
Mean	20.4	22.2	37.2	22.6	23.9	46.6	28.8	36.0	45.6	19.5	28.4	41.6
BT	20.5	22.3	21.3	26.9	28.1	28.2	36.2	42.6	42.2	40.2	46.6	42.0
MI	33.3	34.6	35.1	40.4	41.3	42.5	52.8	57.6	60.2	49.1	54.6	57.1

The bold values represent the lowest errors.

Table 2
Percentage of erroneous disparity values of proposed algorithm with eleven different similarity measures and comparison with various fusion strategies.

	Tsukuba			Venus			Teddy			Cones		
	nonocc	all	disc	nonocc	all	disc	nonocc	all	disc	nonocc	all	disc
Most	14.4	16.3	21.1	12.5	13.9	26.3	15.9	24.6	28.6	7.79	18.2	18.8
Avg	13.9	15.7	21.3	12.1	13.5	27.3	14.0	22.9	27.3	6.64	17.2	17.2
Mult [12]	27.9	29.4	31.2	25.0	26.3	36.2	29.4	36.7	38.3	19.1	28.2	30.0
Conf	13.6	15.5	20.6	12.1	13.5	25.7	14.8	23.6	27.3	6.79	17.3	17.2
Voting	11.6	13.5	18.2	9.6	11.1	**23.3**	12.4	21.4	25.1	5.4	15.8	14.8
Voting + Conf	**11.3**	**13.1**	**17.8**	**8.2**	**9.7**	23.8	**12.1**	**21.0**	**25.1**	**5.1**	**15.5**	**13.6**

The bold values represent the lowest errors.

strategies (with the exception of Mult [12]). Mult is the worst performer in our experiments, presumably, because errors in any of the similarity measures can significantly affect the final result.

4.4. Fusion of intensity (AD) similarity with other similarities

Intensity is computationally the cheapest stereo similarity measure for stereo matching. Therefore, various prior studies aim to fuse intensity with more complex similarities to increase the accuracy. In this experiment, we fuse all of the ten stereo similarities with intensity and compare our results with other fusion strategies.

Table 3 presents the results for the fusion of different similarity measures with the intensity similarity measure. The accuracy of the fused measures are higher than their individual accuracies except for the NCC measure. This is mainly because LRD does not perform as good with NCC as it performs with other measures. The best performer of this experiment is AD + Census. The individual accuracy of MI is one of the worst compared to other measures, however, its performance significantly increased when it is fused with intensity.

Fig. 7 shows the results of our adaptive fusion algorithm when used to fuse AD and census similarities on Middlebury data set. The improvements that are achieved by fusion are indicated in red. For all images, the results of fusion are substantially better than the individual results of each of the fused similarities.

4.5. Constant window aggregation

To explore the performance of our algorithm with respect to other fusion strategies, we perform two experiments with AD–Census and AD–Sobel features in constant window aggregation-based stereo matching. AD and Census features are fused using exponential functions by Mei et al. [11] in order to obtain higher accuracy using the function:

$$C_{AC}(x,y,d) = 2 - e^{-\frac{C_{AD}(x,y,d)}{\sigma_{AD}^2}} - e^{-\frac{C_{Census}(x,y,d)}{\sigma_{Census}^2}}, \quad (6)$$

where σ_{AD} and σ_{Census} are set to 10 and 30, respectively. Additionally, Klaus et al. [10] fused Gradient and AD features linearly (weighted average) via:

$$C_{WA}(x,y,d) = (1-\alpha)C_{AD}(x,y,d) + \alpha C_{SB}(x,y,d). \quad (7)$$

The optimal α for $C_{WA}(x,y,d)$ is not explicitly given by Klaus et al. [10]. We tried different α values and experimentally found that setting α equal to 0.9 gives the best results. We also use the multipli-

Table 3
Percentage of erroneous disparity values that are obtained from the fusion of all similarities.

	Tsukuba			Venus			Teddy			Cones		
	nonocc	all	disc	nonocc	all	disc	nonocc	all	disc	nonocc	all	disc
AD + Rank	17.8	19.5	21.1	18.9	20.2	24.0	21.2	29.3	29.3	9.6	19.7	18.7
AD + Census	**11.4**	**13.2**	**15.8**	**7.8**	**9.3**	**20.5**	**11.6**	**20.6**	**23.5**	**4.9**	**15.4**	**12.7**
AD + NCC	13.8	15.5	16.7	18.0	19.2	26.6	22.3	30.0	32.7	20.0	28.7	26.2
AD + ZNCC	12.7	14.5	17.9	8.6	10.1	25.3	13.0	21.9	27.6	6.3	16.7	16.0
AD + Sobel	12.4	14.2	18.2	16.3	17.7	24.5	22.2	30.2	30.9	10.1	20.1	19.8
AD + LoG	14.9	16.7	18.6	19.5	20.8	28.7	25.3	32.9	33.7	17.1	26.3	25.6
AD + Mean	12.7	14.5	22.2	15.6	16.9	35.7	18.3	26.6	31.8	10.6	20.5	23.2
AD + BT	14.3	16.0	15.9	18.5	19.8	24.1	28.0	35.3	34.4	27.6	35.5	29.9
AD + MI	14.1	15.8	18.0	17.2	18.5	24.6	27.0	34.4	34.2	24.0	32.3	28.7

The bold values represent the lowest errors.

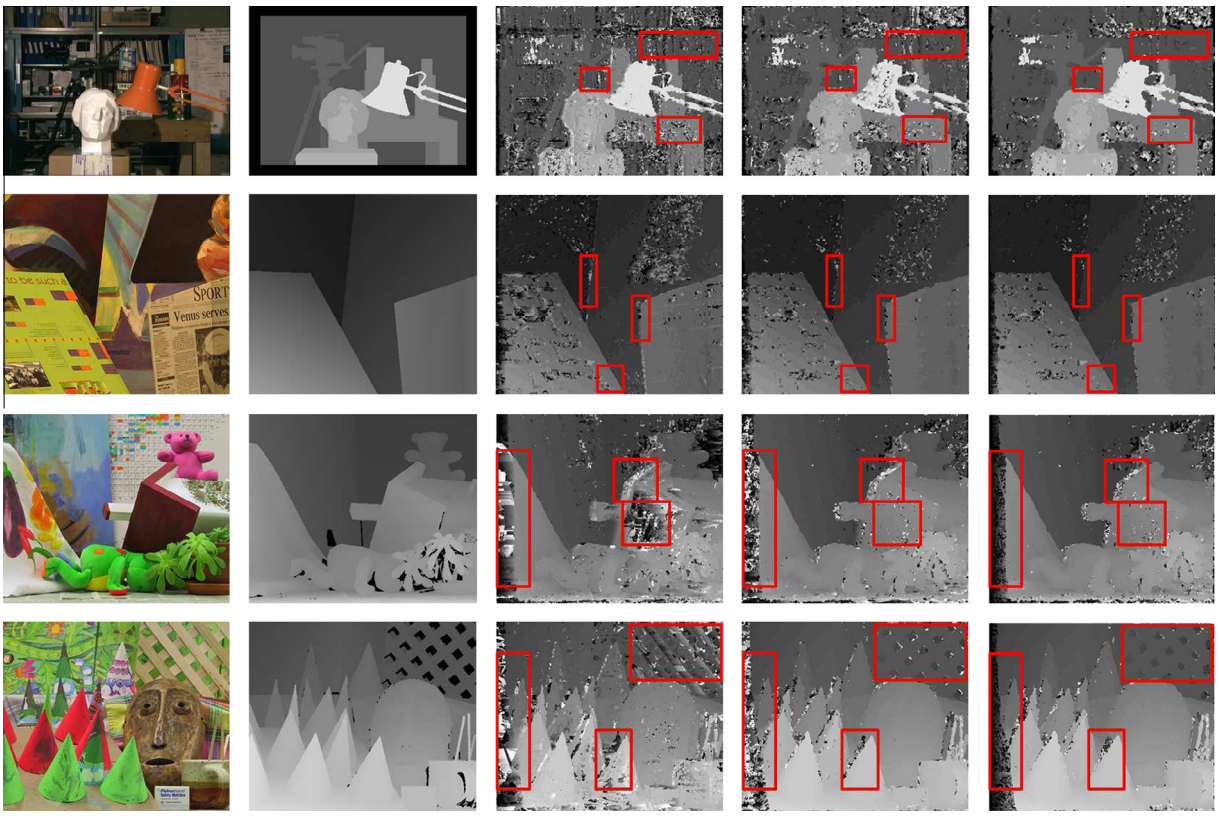

Fig. 7. From left to right: Color images, ground truth disparities, AD similarity, Census similarity, Proposed algorithm results (AD + Census) respectively. Some of the notable differences are marked in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

cation of initial DSI of two features (MULT) in accordance with Wegner and Stankiewicz [12] as:

$$C_{WAC}(x,y,d) = C_{AD}(x,y,d) \times C_{Census}(x,y,d), \tag{8}$$

$$C_{WWA}(x,y,d) = C_{AD}(x,y,d) \times C_{SB}(x,y,d). \tag{9}$$

Stentoumis et al. [13] fused three similarities, AD, Census and Gradient using exponential functions similar to Mei et al. [11]:

$$C_{ACG}(x,y,d) = 3 - e^{-\frac{C_{AD}(x,y,d)}{\lambda_{AD}^2}} - e^{-\frac{C_{Census}(x,y,d)}{\lambda_{Census}^2}} - e^{-\frac{C_{SB}(x,y,d)}{\lambda_{SB}^2}}, \tag{10}$$

where $\lambda_{AD}^2, \lambda_{Census}^2$, and λ_{SB}^2 are set to 5, 45 and 18, respectively.

The results of our experiments for different-sized aggregation windows are presented in Fig. 8. The results are averaged over all of the four data set images and errors are evaluated over all pixels in the images. In both of our experiments, AD is the least accurate similarity measure after aggregation. In the fusion of AD with Census experiment in Fig. 8a, our algorithm performs the best, in particular, for smaller aggregation windows. Mei et al. [11] performs on par with the census similarity for smaller-sized windows. Yet, census outperforms Mei et al. [11] when the aggregation window size increases. In the fusion of AD with Sobel similarity experiment, our proposed method is the best performer for all sizes of the aggregation window, as shown in Fig. 8b. Sobel similarity and the method by Klaus et al. [10] performs similarly in terms of accuracy. Yet, for smaller sized windows, the method of Klaus et al. [10] performs slightly better than Sobel and significantly better than the algorithm based on AD similarities. For the fusion of three similarities, the method by Stentoumis et al. [13] performs on par with the census transform whereas our fusion strategy performs the best especially for small aggregation window sizes. In all of the experiments, the accuracy is saturated to a constant value when the aggregation window size approach to its largest size (13×13). The reason for this is as the aggregation size increases, the disparities near discontinuities in the image are smoothed. In all of our experiments, the multiplication of different initial DSIs (Mult) performs better than AD. However it does not perform better than the other measures or than any of the other fusion strategies.

4.6. Execution time

In general, computational speed is an important parameter of the matching algorithms. In order to decrease the execution time of fusion, we fuse similarity measures before the aggregation step. Therefore, we do not perform aggregation for each of the similarity measures. Table 4 presents the execution times that are required for different fusion strategies with sequential coding. Because of their simplicity, the methods by Klaus et al. [10], Mei et al. [11], and Stentoumis et al. [13], as well as the Mult [12] method are faster than our fusion strategy. The most computationally expensive parts of our fusion strategy are the extraction of confidence map (0.16 s) and constructing the consensus (0.12 s), which is highly parallelizable. However, the execution time of our algorithm is still lower than the straightforward constant window aggregation for window sizes larger than 7×7 as depicted in Table 5.

Let N be the total number of pixels, H be the size of the consensus, I be the number of similarities to be fused, and D be the total number of disparities, Our algorithm complexity is $O(NI(H+D))$. Fusion strategies with static weights [11,10,12,13] has the complexity of $O(N(H+D)$. Therefore, our fusion strategy is $O(I)$ worse in terms of complexity than other algorithms. However, since our calculations are mostly independent, our fusion strategy is highly parallelizable.

4.7. Effect of fusion on global energy minimization

Global energy minimization (GEM) algorithms such as graph-cuts are widely used in stereo matching [33–36]. The main benefit

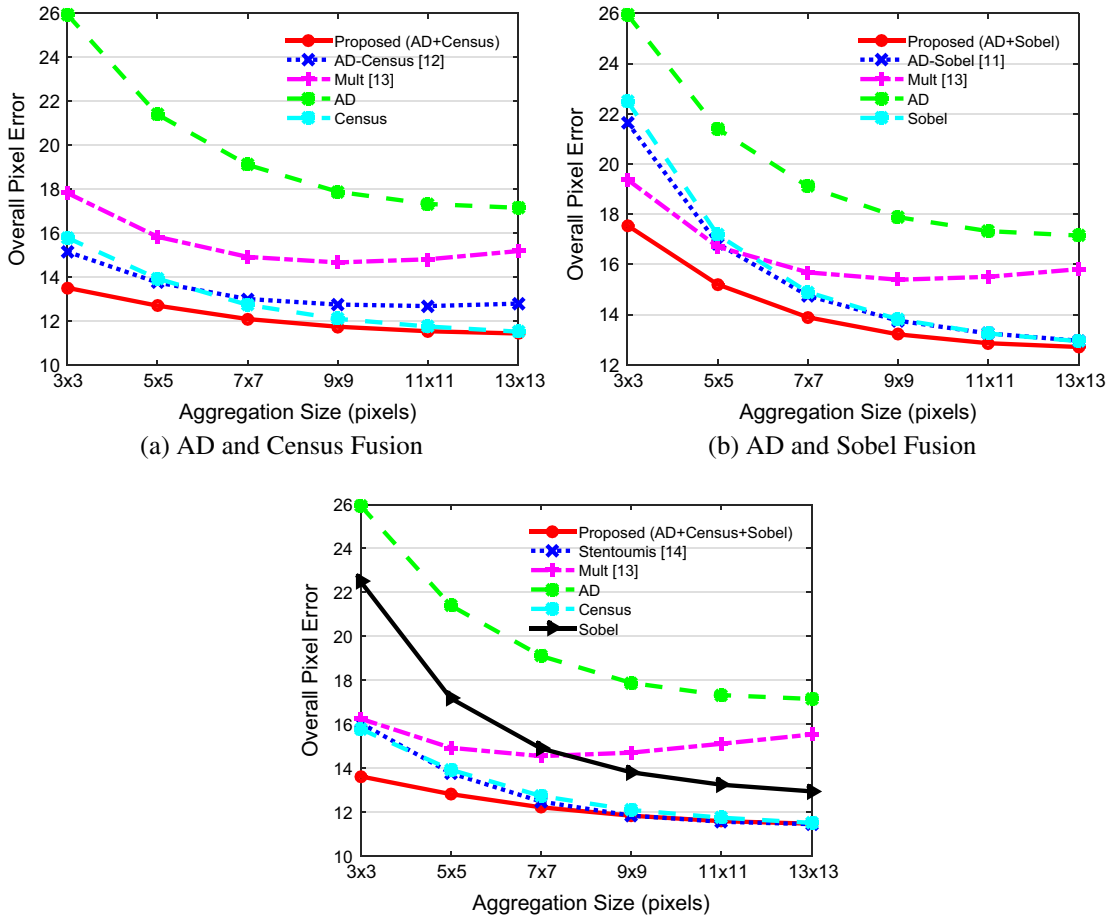

Fig. 8. The proposed algorithm performance with respect to (a) AD and Census similarity measures, Mei et al. [11] and Mult [12], (b) AD, Sobel similarity measures, Klaus et al. [10] and Mult [12], and (c) Color, Census, Sobel similarity measures, Stentoumis et al. [13] and Mult [12]. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 4
The execution times of different fusion strategies on a computer with Intel Xeon E3-1270 processor.

	AD	Sobel	Census	Klaus et al. [10]	Mei et al. [11]	Stentoumis et al. [13]	Mult [12]	Proposed
Time (s)	0.17	0.18	1.05	0.01	0.01	0.01	0.005	0.3

Table 5
Time required to apply aggregation on a computer with Intel Xeon E3-1270 processor.

	3×3	5×5	7×7	9×9	11×11	13×13
Time (s)	0.13	0.27	0.5	0.79	1.15	1.57

of using GEM is that the pairwise interactions between disparity estimates at nearby image locations smooth the final disparity estimates. As a result, stereo algorithms with GEM generally construct better disparity maps.

We tested one of the most commonly used GEM algorithms, graph-cuts (GC), with our fusion strategy. In our experiments, we used the same parameters and configurations for all of our tests as suggested in [36]. The left borders of the reference images are the regions which cannot be matched effectively because they do not exist in the target image. Therefore, the matching costs on the left borders of all images are penalized in order to prevent biased smoothing for different similarities. Table 6 shows the quantitative results of single similarity measures for fusion and fusion strategies. As before, census is the best performer among the individual similarity measures. The methods by Mei et al. [11] and Klaus et al. [10] do not achieve better results than their single best-performing similarity measures (census and Sobel) in any of the experiments. This result illustrates the potential of incorporating (adaptive) fusion strategies in stereo matching. Since multiplying similarity measures amplifies errors, Mult [12] performs the worst in all of our experiments. Our strategy achieves the best results on the Teddy and Cones image pairs, and it performs on par with other measures on the Tsukuba and Venus image pairs. These results indicate the importance of having accurate initial disparity estimations for stereo algorithms with global energy minimization. For the fusion of three similarities, our algorithm performs the best for all of the images except only for one image on its discontinuity locations. As more similarities are fused with static weights, it becomes non-trivial to set the optimal weights experimentally. Hence, our adaptive strategy becomes more effective as the number of stereo similarity measures increases.

Table 6
Percentage of erroneous disparity values of GC algorithm that is applied to: Single Similarities (S. Sim.), fusion of AD and Census (A.C.), fusion of AD and Sobel (A.S.), and fusion of AD, Census, and Sobel (A.C.S). The best results for each test group is underlined and the best results of the overall test are marked as bold.

		Tsukuba			Venus			Teddy			Cones		
		nonocc	all	disc	nonocc	all	disc	nonocc	all	disc	nonocc	all	disc
S. Sim.	AD	6.4	8.2	23.0	3.3	4.5	23.8	33.9	40.7	45.9	14.6	23.1	27.1
	Sobel	**3.2**	**5.1**	16.1	1.7	2.7	20.7	12.3	21.1	30.4	9.0	17.5	22.4
	Census	4.6	6.4	14.7	**1.4**	**2.3**	16.4	10.7	19.5	28.7	7.5	16.0	18.0
A.C.	Mei et al. [11]	16.5	17.7	46.7	14.3	15.4	36.5	41.6	47.4	63.9	27.1	33.5	42.2
	Mult [12]	10.8	12.5	33.9	7.5	8.7	29.9	41.8	47.7	55.6	16.4	24.2	29.5
	Proposed	4.5	6.3	**14.3**	**1.4**	2.4	**16.2**	**10.1**	**19.0**	**26.8**	**7.2**	**15.8**	**16.9**
A.S.	Klaus et al. [10]	**3.2**	**5.1**	16.1	1.8	2.8	20.7	13.6	22.6	32.1	8.9	17.4	22.0
	Mult [12]	11.0	12.7	33.9	8.7	9.7	33.6	41.9	47.7	53.8	17.3	24.9	31.1
	Proposed	3.3	5.2	15.5	1.8	2.9	21.3	12.0	20.7	30.2	8.2	16.8	19.9
A.C.S	Stentoumis et al. [13]	5.5	7.5	15.3	2.7	3.7	19.1	12.9	21.5	26.8	7.8	16.4	18.5
	Mult [12]	14.5	16.4	20.0	15.7	17.1	25.7	19.1	27.3	28.1	9.9	19.9	20.3
	Proposed	4.4	6.3	14.9	1.6	2.5	17.4	10.7	19.5	27.7	7.4	16.0	17.5

Table 7
The mean and standard deviations of erroneous disparity percentage of [38] algorithm that is applied to: Single Similarities (S. Sim.), fusion of AD and Census (A.C.), fusion of AD and Sobel (A.S.), and fusion of AS, Census, and Sobel, and averaged over KITTI dataset. The best results for each test group is underlined and the best results of the overall test are marked as bold.

		nonocc	occ
S. Sim.	AD	31.8 ± 13.6	33.3 ± 13.4
	Sobel	27.4 ± 7.5	28.9 ± 7.4
	Census	11.8 ±5.9	13.7 ±6
A.C.	Mei et al. [11]	13.8 ± 7.3	15.8 ± 7.3
	Mult [12]	22.6 ± 9.7	24.4 ± 9.6
	Proposed	**11.2 ±5.6**	13.2 ±5.6
A.S.	Klaus et al. [10]	26.3 ±7.5	27.8 ±7.4
	Mult [12]	28.7 ± 9.3	30.3 ± 9.2
	Proposed	24.4 ±7.2	26.0 ±7.1
A.C.S.	Stentoumis et al. [13]	17.3 ± 9	19.1 ± 9
	Mult [12]	25.4 ± 7.9	27.1 ± 7.9
	Proposed	**11.2 ± 5.7**	**13.1 ± 5.7**

4.8. Results on the KITTI data set

The KITTI data set is a recently released benchmark that contains real-world images. KITTI contains 194 test images with ground truth depth captured by a Volodyne laser scanner. The ground truth is available for both occluded and non-occluded surfaces. The error is calculated by finding the percentage of the pixels that has a difference between estimated and ground-truth of more than three pixels.

One of the most popular local stereo algorithms in the literature is adaptive-support weight of Yoon and Kweon [9]. The algorithm [37] aggregates initial cost measures adaptively to obtain accurate depth estimations that preserves the depth discontinuities. As a further improvement, Hosni et al. [38] incorporated edge-preserving bilateral filtering [39] over the initial cost space as aggregation step to preserve the sharpness of disparities at discontinuity locations. In order to evaluate the performance of our adaptive fusion strategy in state-of-the-art local stereo matching algorithms, we integrated our fusion algorithm with edge-preserving bilateral filtering [39,38] and tested it on the KITTI data set. The mean and standard deviation of the errors over all images are presented in Table 7 for both non-occluded and occluded image regions. Our algorithm performs the best overall and in each of different fusion classes. Since census significantly outperforms the other measures, the improvement with census is not as large as with AD and Sobel. Our fusion strategy outperforms the best performer of the fused similarity measures in all cases. Fig. 9 shows an example of the result of our fusion algorithm on the KITTI data set. Supporting to our previous results with global energy minimization, our algorithm clearly outperforms the method of Stentoumis et al. [13]. The method of Mei et al. [11] also outperforms Stentoumis et al. [13], which shows that using static weights to fuse multiple similarities get less efficient as the number of similarities increases.

4.9. Cross-modal stereo between infra-red and RGB views

The Microsoft Kinect depth sensor was introduced in 2010 as a human–computer interaction device that can provide high resolution depth maps of a scene in real-time. Kinect can recover the

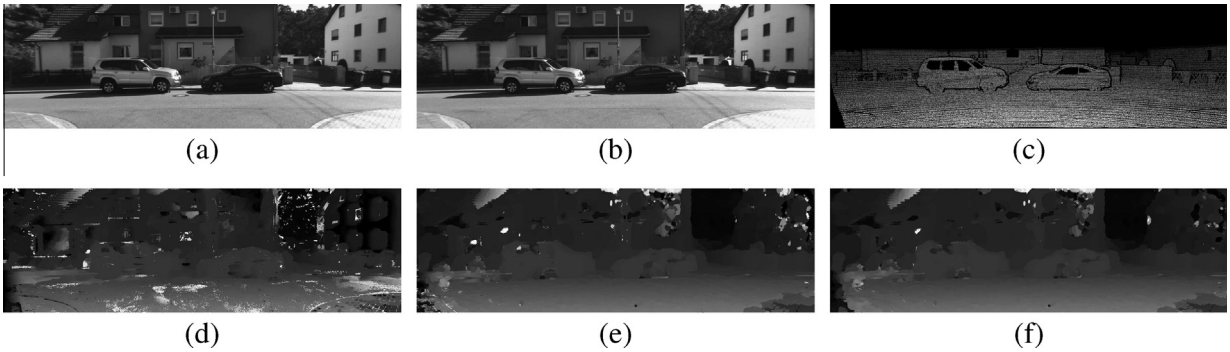

Fig. 9. Stereo results from KITTI dataset: (a) left image, (b) right image, (c) ground truth, resulting disparities: (g) AD, (h) Census, and (g) Proposed (AD + Census).

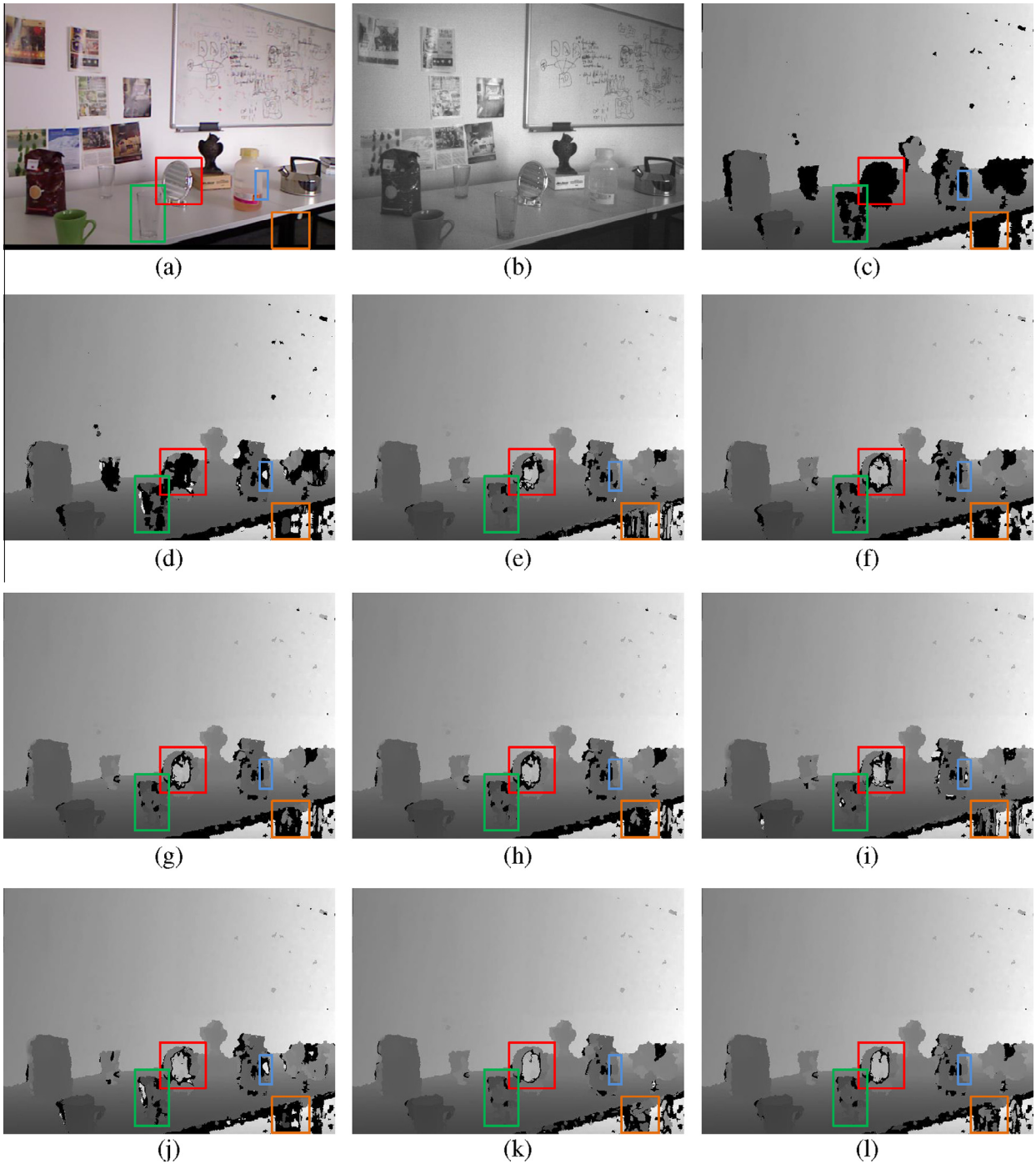

Fig. 10. Kinect cross-modal stereo results: (a) Color, (b) IR, (c) Raw depth, (d) AD, (e) Sobel, (f) Census, (g) Klaus et al. [10], (h) Mei et al. [11], (i) Mult (AD + Sobel) [12], (j) Mult (AD + Census), (k) Proposed(AD + Sobel), (l) Proposed (AD + Census). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

depth of many kinds of surfaces but it cannot measure depth on transparent objects. In order to obtain sparse depth estimations on these objects, Chiu et al. [40,41] proposed to use cross-modal stereo between the infra-red (IR) and RGB cameras of the Kinect. Since the IR and RGB views are not structurally the same, the main challenge in cross-modal stereo is to find as many reliable matches as possible. Therefore, the development of robust similarity measures is essential in using cross-modal stereo.

We tested our fusion strategy for cross-modal stereo and compared it with different fusion strategies and individual similarity measures. Fig. 10 presents the results of the tested measures. The black regions are the regions where the stereo match has not enough confidence. The regions with color indicate locations where there are significant differences between different algorithms. Our algorithm produces the densest correct matches in most of the indicated locations. This result shows the importance of robust fusion strategies in challenging stereo problems. Although our fusion strategy achieves improved results compared to single similarities and other fusion strategies, there are still errors in the measurements. These errors are likely the result of the fact that our fusion strategy also relies on the performance of its similarities. When none of the similarities achieves correct esti-

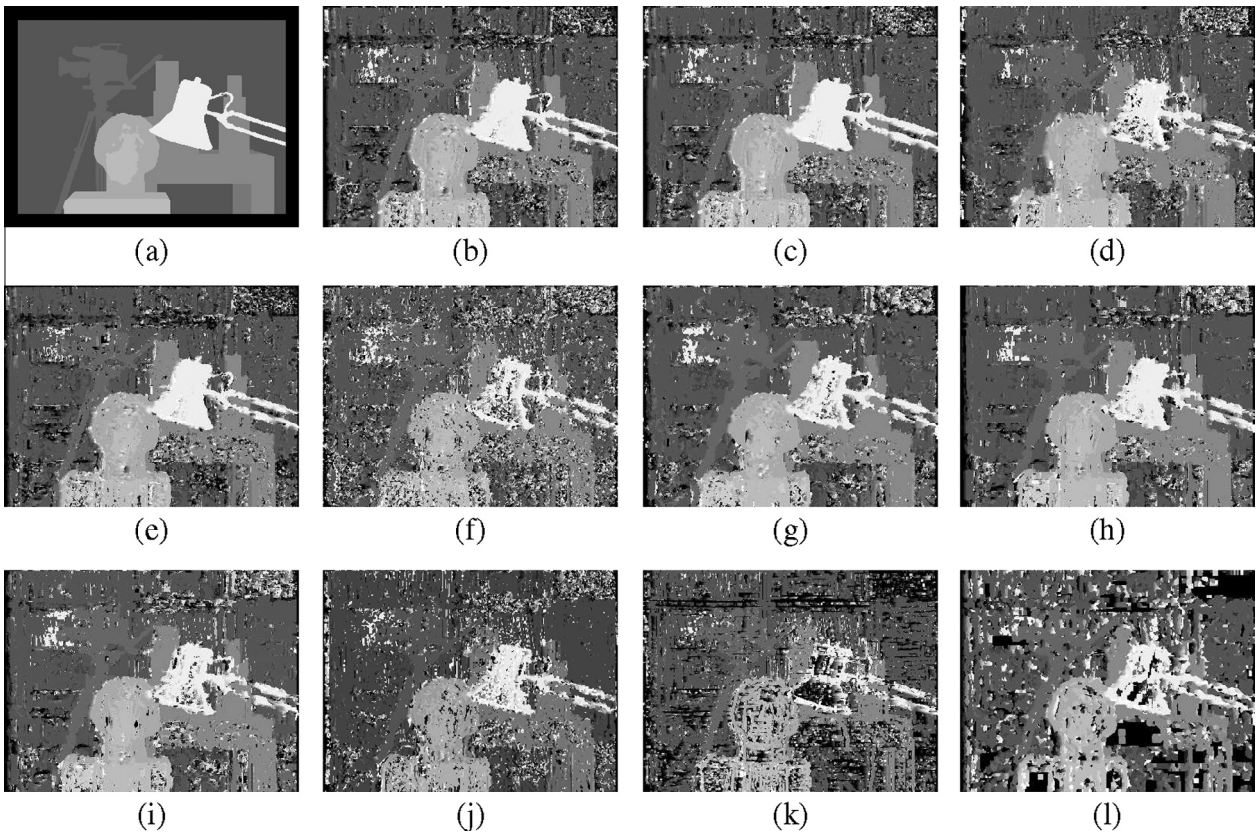

Fig. 11. Performances of stereo similarity measures on Tsukuba image in Middlebury stereo dataset: (a) Ground truth, (b) AD, (c) RGB, (d) Mean, (e) BT, (f) Rank, (g) Census, (h) NCC, (i) ZNCC, (j) Sobel, (k) LoG, and (l) MI.

mations on a particular image region, such as a specular surface, our fusion strategy cannot improve the accuracy: any fusion algorithm is limited by the performance of its inputs.

5. Discussion

Our experimental results illustrate the potential of adaptive stereo similarity fusion on the accuracy of stereo matching. In this section, we further elaborate on the potential merits of our fusion strategy, and we discuss possible improvements as well as possible limitations of our approach.

5.1. Effect of confidence

The most common way of fusing similarity measures is to weight each measure and combine them using manually defined functions, such as via direct aggregation [10] or using exponentials [11]. Our experiments show that using confidence-guided weights for the fusion of similarity measures may improve stereo matching results substantially compared to constant-weight fusion algorithms. This happens because each similarity measure may perform differently at different locations of the image. The incorporation of confidences can be further applied to cost aggregation step of stereo matching algorithms where adaptive window sizes can be used with the guidance of stereo confidences rather than constant window sizes through the image.

5.2. Effect of consensus

Stereo similarity measure algorithms in the literature use pixel-wise weighted sums of initial cost measures. However, to fuse different similarity measures, it is important to exploit spatial correlation between neighboring pixels since pixel-wise measures tend to be noisy. Additionally, confidence measures are not very precise on the noisy initial DSI. Therefore, the accuracy of fusion substantially increased when we incorporated a consensus strategy. Our experiments also show that as consensus window size increases, it violates disparity discontinuities and the performance on discontinuous image regions decreases. To address this problem, discontinuity-preserving windows can be investigated for building the consensus.

5.3. Effect of improvement on the initial DSI

In this work, we also show that improvements in the initial cost measurements can substantially affect the final result of the stereo algorithm. Additionally, improvements in the initial cost measurements can further increase the performance of stereo algorithms on challenging stereo problems such as cross-modal stereo matching.

5.4. Possible limitations

Our algorithm improves initial cost measures substantially and this improves the final result. However the errors caused by the remaining steps such as aggregation and refinement may damage the improvements that are achieved at the initial step. For example, smoothing the image substantially without preserving the disparity discontinues at the refinement step may eliminate the improvements of our fusion strategy in discontinuous image regions. This is one of the reasons why the fusion algorithms do not always obtain the top accuracy with the global energy minimization algorithm in our experiments.

Fusing many similarity measures is one of the capabilities of our fusion strategy. However, our experiments show that fusing simi-

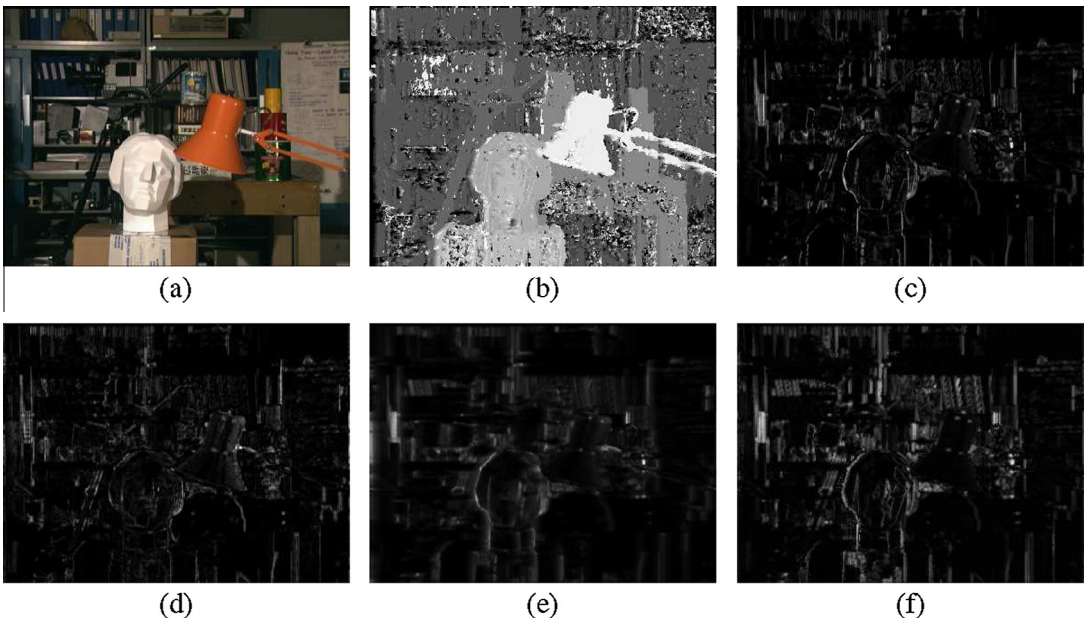

Fig. 12. The performances of stereo confidences. Brighter regions show confident estimations whereas darker regions are less confident estimations: (a) color, (b) disparity (c) LRD, (d) PKRN, (e) MLM, (f) LC. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

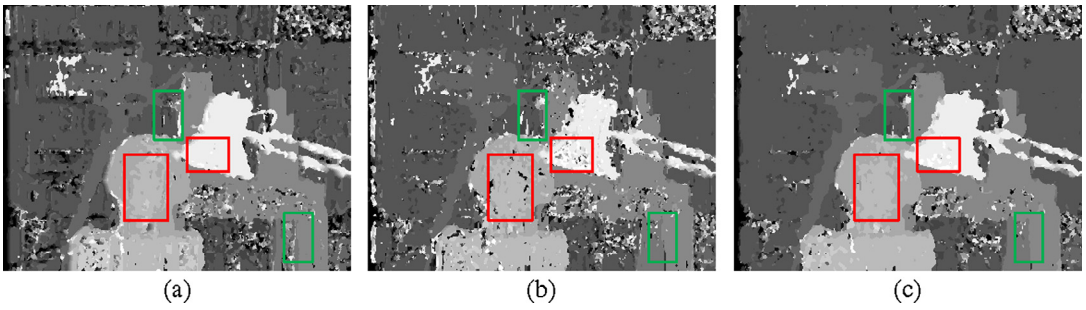

Fig. 13. The improvements that are achieved by fusion of intensity and sobel similarity measures: (a) intensity, (b) Sobel, (c) Proposed fusion. The red region shows where intensity is better than sobel. On the contrary, green region shows where sobel is better than intensity. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

larity measures that perform not as good as others does not improve the fusion results. Using stereo confidences with better performance may improve the fusion performance and circumvent this problem.

5.5. Possible future directions

Similar problems as in stereo matching arise in medical image registration where the aim of the registration is to find the mapping (displacement) of one image to the other. Cost functions such as NCC and MI are commonly used in such image registration problems. As future work, we aim to generalize our adaptive fusion algorithm to medical image registration problems.

6. Conclusion

In this paper, we presented a novel adaptive fusion algorithm for stereo similarity measures, which uses stereo confidences to determine the fusion weights. To the best of our knowledge, our stereo fusion algorithm is the first that does not require any similarity measure-dependent parameters and that can be applied to fuse any number of similarity measures. The results of our experiments show that substantial accuracy increases may be obtained compared to stereo matching algorithms based on individual similarity performances or on non-adaptive fusion strategies. Specifically, the results show that stereo confidences can be used as the basis for computing adaptive per-pixel weights for stereo similarity metric fusion. Additionally, we showed that exploiting spatial correlation in a local region by means of a consensus neighborhood may increase the fusion accuracy. Our approach is also effective in challenging stereo problems such as cross-modal stereo matching, where individual stereo similarity measures generally fail to find sufficient good matches.

Appendix A. Stereo similarity measures

The first step of the stereo matching algorithms is measuring matching costs of pixels using a similarity measure. We implemented eleven different similarity measures based on [20]. The values parameters for each of the similarity measures are set according to their implementations in the literature. Fig. 11 represents the disparity estimations from initial matching for different similarity measures on Tsukuba dataset.

In this section, we describe the similarity (Appendix A) and confidence measures (Appendix B) that will be used in our experiments.

AD measures the absolute intensity difference between the reference (left) image $I(x,y)$ and the target (right) image $I'(x-d,y)$:

$$C_{AD}(x,y,d) = |I(x,y) - I'(x-d,y)|. \quad (.1)$$

RGB measures the absolute difference of all color channels R,G,B and aggregates them to have the final matching score:

$$C_{RGB}(x,y,d) = \sum_{i=R,G,B} |I_i(x,y) - I'_i(x-d,y)|. \quad (.2)$$

Mean is similar to **AD** but the images are filtered with a mean filter of size N before the matching:

$$I_{Mean}(x,y) = I(x,y) - \frac{1}{|N|}\sum_{i=-N/2}^{N/2} I(x-i,y-i),$$
$$C_{Mean}(x,y,d) = |I_{Mean}(x,y) - I'_{Mean}(x-d,y)|. \quad (.3)$$

BT is the Birchfield–Tomasi similarity measure [29] which is a sampling-intensive absolute difference measure. Unavoidable image sampling may introduce errors in stereo matching since the correct match may not be shifted in discrete disparity intervals. BT suppresses such errors by using interpolation of neighbouring pixels:

$$I^-(x,y) = (I(x-1,y) + I(x,y))/2,$$
$$I^+(x,y) = (I(x+1,y) + I(x,y))/2,$$
$$I^{min}(x,y) = \min(I^-(x,y), I(x,y), I^+(x,y)),$$
$$I^{max}(x,y) = \max(I^-(x,y), I(x,y), I^+(x,y)),$$
$$A(x,y,d) = \max(0, I(x,y) - I'^{max}(x-d,y), I'^{min}(x-d,y) - I(x,y)),$$
$$B(x,y,d) = \max(0, I'(x-d,y) - I^{max}(x,y), I^{min}(x,y) - I'(x-d,y)),$$
$$C_{BT}(x,y,d) = \min(A(x,y,d), B(x,y,d)). \quad (.4)$$

Rank [23] transform is a non-parametric image transform that models the structure of the neighbourhood of pixels by exploiting the intensity variation. Eq. (.5) represents the rank transform $RT(x,y)$ of a pixel (x,y) inside a local neighbourhood $N(x,y)$ of size 7×7 and initial DSI, $C_{RT}(x,y,d)$ as:

$$RT(x,y) = |(x',y') \in N(x,y) | I(x',y') < I(x,y)|,$$
$$C_{RT}(x,y,d) = |RT(x,y) - RT'(x-d,y)|. \quad (.5)$$

Census [23] transform models the structure of 7×7 neighbourhood of pixels that is denoted by k, as represented in Eq. (.6). Census is one of the most robust similarity measure against radiometric differences between stereo pairs [20] and it is calculated as:

$$CT(x,y)[k] = \begin{cases} 1, & \text{iff } I(x_k,y_k) > I(x,y) \\ 0, & \text{otherwise,} \end{cases}$$
$$\mu_c(x,y,d)[k] = \begin{cases} 1, & \text{iff } CT(x,y)[k] = CT'(x-d,y)[k] \\ 0, & \text{otherwise,} \end{cases}$$
$$C_{CT}(x,y,d) = \sum_{\forall k} \mu_c(x,y,d)[k]. \quad (.6)$$

The initial DSI of Census transformed images calculated by using Hamming distance [23].

NCC is an intensity and patch based matching method that is especially robust against Gaussian noise between the matched patches. For simplified notation, let I_p and I_{p-d} denote the pixels at (x,y) and $x-d,y$ respectively. Eq. (.7) presents the initial DSI calculation using NCC:

$$C_{NCC}(p,d) = \frac{\sum_{p'\in N_p} I_{p'} I'_{p'-d}}{\sqrt{\sum_{p'\in N_p} I^2_{p'} \sum_{p'\in N_p} I'^2_{p'-d}}}. \quad (.7)$$

ZNCC is similar to NCC whereas it provides more robustness against gain and offset variation between matched image patches [30]:

$$C_{ZNCC}(p,d) = \frac{\sum_{p'\in N_p} (I_{p'} - \bar{I}_p)(I'_{p'-d} - \bar{I}'_p)}{\sqrt{\sum_{p'\in N_p} (I_{p'} - \bar{I}_p)^2 \sum_{p'\in N_p} (I'_{p'-d} - \bar{I}'_p)^2}}. \quad (.8)$$

For calculating NCC and ZNCC, we choose a patch size, $|N_p|$ of 5×5.

Sobel can suppress the noise in the intensity images. Let $I_s(x,y)$ denote sobel filter of size 3×3 response of image I at pixel (x,y):

$$C_{SB}(x,y,d) = |I_s(x,y) - I'_s(x-d,y)|. \quad (.9)$$

LoG can suppress the noise and provide robustness against offset in intensities. Similar to [20], we incorporated LoG kernel with size 5×5 and standard deviation of 1:

$$L(x,y) = \frac{-1}{\pi\sigma^4}\left(1 - \frac{x^2+y^2}{2\sigma^2}\right) e^{-\frac{x^2+y^2}{2\sigma^2}},$$
$$I_{LoG}(x,y) = I(x,y) \otimes L(x,y),$$
$$C_{LoG}(x,y,d) = |I_{LoG}(x,y) - I'_{LoG}(x-d,y)|. \quad (.10)$$

Mutual Information (MI). If we assume the intensity values of pixels as random variables (RV), X, that has probability density function P, we can find the correlation between the two distributions in stereo images. Eq. (.11) depicts the entropy, $H(X)$, of the distribution of X. MI is used to find how similar the two distributions are between two image patches in stereo images and calculated as depicted in Eq. (.12).

$$H(X) = -\sum_x P_X(x)(\log(P_X(x))), \quad (.11)$$
$$MI(X,Y) = H(X) + H(Y) - H(X,Y), \quad (.12)$$
$$C_{MI}(x,y,d) = -MI(X,Y), \quad (.13)$$

where $H(X,Y)$ is the joint entropy of the two distributions, X and Y. In Eq. (.13), we negate the mutual information in order to obtain a cost measure rather than a correlation. Therefore, as the mutual information decrease, better stereo correspondences are obtained.

AD, RGB, Mean, BT, Rank, Census, Sobel and LoG are calculated pixel-wise, however NCC, ZNCC and MI are calculated over a neighbourhood. In order to compensate this difference, the pixel-wise costs are aggregated over 3×3 windows. All of the costs are normalized to have values in $[0,1]$ before confidence estimation and fusion in order to prevent range difference in between.

Appendix B. Stereo Confidence Measures

To filter the wrong matches in stereo matching, it is important to measure the confidence of matching at each pixel. The confidence of a match can be measured using the DSI. Hu and Mordohai [15] provided a survey of stereo confidences by incorporating 18 different stereo confidences. Pfeiffer et al. [28] picked four best performing confidence measures based on [15] in their experiments. In our work, we also choose four best performing confidence measures based on [15,28] and incorporate them in our fusion algorithm. Fig. 12 shows the four confidence estimation results for the initial matching on Tsukuba dataset (see Fig. 13).

Left–Right Difference (LRD) is one of the top performing confidence measures because of its cross control over the left and right matching scores. Let d_1, c_1, and c_2 be the winner disparity, minimum, and second minimum costs respectively, LRD confidence, $S_{LRD}(x,y)$, can be calculated as:

$$S_{LRD}(x,y) = \frac{c_2 - c_1}{\left| c_1 - \min_{d'}(c'(x-d_1,y,d')) \right| + \epsilon}, \quad (.14)$$

where $c'(x-d_1,y,d')$ denotes target to reference cost and d' represents the disparities from target to reference. As the difference between c_1 and $\min_{d'}(c'(x-d_1,y,d'))$ decrease, the confidence

increases. We would expect that they would be equal if there is no error and no occlusion between the two images. Therefore, we add a small positive value ϵ to the denominator to ensure we do not obtain a zero denominator.

Naive Peak Ratio (PKRN) is one of the simple yet effective stereo confidence measure that uses only the reference matching scores. S_{PKRN} is defined as:

$$S_{PKRN} = \frac{c_2}{c_1 + \epsilon}. \quad (.15)$$

S_{PKRN} observes the ratio of the minimum and second minimum costs. As long as this ratio is high, the best match is unique compared to other possible matches and vice versa.

Maximum Likelihood Measure (MLM) assumes that the matching cost values has a normal distribution so its pdf can be calculated. S_{MLM} can be calculated as:

$$S_{MLM} = \frac{e^{-\frac{c_1}{2\sigma_{MLM}^2}}}{\sum_d e^{-\frac{c(d)}{2\sigma_{MLM}^2}}}, \quad (.16)$$

where $c(d)$ describes the cost value for each disparity d. S_{MLM} also considers only the reference cost values.

Local Curve (LC) [31] is incorporated by Pfeiffer et al. [28] and provided the best accuracy in their Stixel World experiments. LC considers the difference of the costs of the candidate disparities around the winner disparity with minimum cost. Let c_+ and c_- be the costs of the adjacent disparities around d_1, S_{LC} is calculated as:

$$S_{LC} = \frac{\max(c_+, c_-) - c_1}{\gamma}, \quad (.17)$$

where γ is a parameter to separate the distribution nicely and assigned to 480 in [28].

References

[1] M. Pollefeys, R. Koch, M. Vergauwen, L. Van Gool, Automated reconstruction of 3D scenes from sequences of images, ISPRS J. Photogramm. Rem. Sens. 55 (2000) 251–267.
[2] M. Schmeing, X. Jiang, Depth image based rendering, Pattern Recogn. Mach. Intell. Biometrics (2011) 279–310.
[3] D. Scharstein, R. Szelinski, A taxonomy and evaluation of dense two-frame stereo correspondence algorithms, IJCV 47 (2003) 7–42.
[4] Y.C. Hsieh, D.M. McKeown Jr., F.P. Perlant, Performance evaluation of scene registration and stereo matching for artographic feature extraction, PAMI (14) (1994) 214–238.
[5] K. Schauwecker, R. Klette, A. Zell, A new feature detector and stereo matching method for accurate high-performance sparse stereo matching, Intell. Robots Syst. (IROS) (2012) 5171–5176.
[6] G. Saygili, L. van der Maaten, E.A. Hendriks, Improving segment based stereo matching using SURF key points, in: 19th IEEE International Conference on Image Processing, 2012, pp. 5171–5176.
[7] P.F. Felzenszwalb, D.P. Huttenlocher, Efficient belief propagation for early vision, IJCV 70 (2006) 41–54.
[8] V. Kolmogorov, R. Zabih, Computing visual correspondence with occlusions using graph cuts, in: Proceedings Eighth IEEE International Conference on Computer Vision (ICCV), vol. 2, 2001, pp. 508–515.
[9] P.F. Felzenszwalb, D.P. Huttenlocher, Adaptive support-weight approach for correspondence search, IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) 28 (2006) 650–656.
[10] A. Klaus, M. Sormann, K. Karner, Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure, in: IEEE International Conference on Pattern Recognition (ICPR), 2006, pp. 15–18.
[11] X. Mei, X. Sun, M. Zhou, S. Jiao, H. Wang, X. Zhang, On building an accurate stereo matching system on graphics hardware, in: ICCV Workshops, 2011, pp. 467–474.
[12] K. Wegner, O. Stankiewicz, Similarity measures for depth estimation, in: 3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), 2009, pp. 1–4.
[13] C. Stentoumis, L. Grammatikopoulos, I. Kalisperakis, G. Karras, On accurate dense stereo-matching using a local adaptive multi-cost approach, ISPRS J. Photogramm. Rem. Sens. (2014).
[14] P. Mordohai, The self-aware matching measure for stereo, ICCV (2009) 1841–1848.
[15] X. Hu, P. Mordohai, A quantitative evaluation of confidence measures for stereo vision, PAMI 34 (2012) 2121–2133.
[16] K.J. Yoon, I.S. Kweon, Distinctive similarity measure for stereo matching under point ambiguity, CVIU 112 (2008) 173–183.
[17] A. Milella, R. Siegwart, Stereo-based ego-motion estimation using pixel tracking and iterative closest point, in: IEEE International Conference on Computer Vision Systems, (ICVS), 2006.
[18] P. Steingrube, S.K. Gehrig, U. Franke, Performance evaluation of stereo algorithms for automotive applications, in: Computer Vision Systems, Springer, 2009, pp. 285–294.
[19] H. Hirschmuller, D. Scharstein, Evaluation of cost functions for stereo matching, CVPR (2007) 1–8.
[20] H. Hirschmuller, D. Scharstein, Evaluation of stereo matching costs on images with radiometric differences, PAMI 31 (2009) 1582–1599.
[21] Q. Yang, A non-local cost aggregation method for stereo matching, CVPR (2012) 1402–1409.
[22] K. Zhang, J. Lu, G. Lafruit, Cross-based local stereo matching using orthogonal integral images, IEEE Trans. Circ. Syst. Video Technol. 19 (2009) 1073–1079.
[23] R. Zabih, J. Woodfill, Non-parametric local transforms for computing visual correspondence, ECCV (1994) 151–158.
[24] H. Hirschmuller, Accurate and efficient stereo processing by semi-global matching and mutual information, CVPR (2) (2005) 807–814.
[25] K. Zhang, J. Lu, G. Lafruit, R. Lauwereins, L. Van Gool, Robust stereo matching with fast normalized cross-correlation over shape-adaptive regions, in: IEEE International Conference on Image Processing (ICIP), 2009, pp. 2357–2360.
[26] N. Einecke, J. Eggert, A two-stage correlation method for stereoscopic depth estimation, in: International Conference on Digital Image Computing: Techniques and Applications (DICTA), 2010, pp. 227–234.
[27] G. Egnal, Mutual Information as a Stereo Correspondence Measure, Technical Reports (CIS), 2000 (113).
[28] D. Pfeiffer, S. Gehrig, N. Schneider, Exploiting the power of stereo confidences, CVPR (2013) 297–304.
[29] S. Birchfield, C. Tomasi, Depth discontinuities by pixel-to-pixel stereo, IJCV 35 (1999) 269–293.
[30] J. Banks, P. Corke, Quantitative evaluation of matching methods and validity measures for stereo vision, Int. J. Robot. Res. 20 (2001) 512–532.
[31] A. Wedel, A. Meißner, C. Rabe, U. Franke, D. Cremers, Detection and segmentation of independently moving objects from dense scene flow, EMMCVPR (2009) 14–27.
[32] A. Geiger, P. Lenz, R. Urtasun, Are we ready for autonomous driving? in: The KITTI Vision Benchmark Suite CVPR, 2012.
[33] V. Kolmogorov, R. Zabin, What energy functions can be minimized via graph cuts?, PAMI 26 (2004) 147–159
[34] Y. Boykov, O. Veksler, R. Zabih, Fast approximate energy minimization via graph cuts, PAMI 23 (2001) 1222–1239.
[35] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, C. Rother, A comparative study of energy minimization methods for Markov random fields, ECCV (2006) 16–29.
[36] Y. Boykov, O. Veksler, R. Zabih, Fast approximate energy minimization via graph cuts, PAMI 23 (2001) 1222–1239.
[37] K. Yoon, I. Kweon, Adaptive support-weight approach for correspondence search, TPAMI 28 (2006) 650–656.
[38] A. Hosni, C. Rhemann, M. Bleyer, C. Rother, M. Gelautz, Fast cost-volume filtering for visual correspondence and beyond, TPAMI 35 (2013) 504–511.
[39] K. He, J. Sun, X. Tang, Guided image filtering, ECCV (2010) 1–14.
[40] W. Chiu, U. Blanke, M. Fritz, Improving the Kinect by cross-modal stereo, BMVC 1 (2011).
[41] W. Chiu, U. Blanke, M. Fritz, I spy with my little eye: learning optimal filters for cross-modal stereo under projected patterns, in: ICCV Workshops, 2011, pp. 1209–1214.