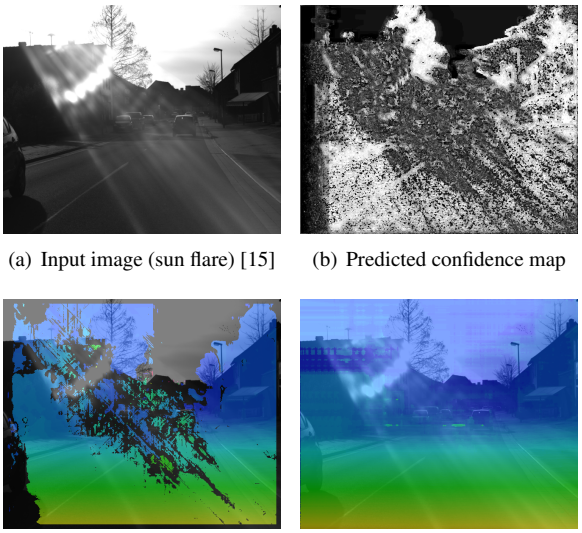
Leveraging Stereo Matching with Learning-based Confidence Measures

Min-Gyu Park and Kuk-Jin Yoon
Computer Vision Laboratory, GIST, South Korea
{mpark,kjyoon}@gist.ac.kr

Abstract

We propose a new approach to associate supervised learning-based confidence prediction with the stereo matching problem. First of all, we analyze the characteristics of various confidence measures in the regression forest framework to select effective confidence measures using training data. We then train regression forests again to predict the correctness (confidence) of a match by using selected confidence measures. In addition, we present a confidence-based matching cost modulation scheme based on the predicted correctness for improving the robustness and accuracy of various stereo matching algorithms. We apply the proposed scheme to the semi-global matching algorithm to make it robust under unexpected difficulties that can occur in outdoor environments. We verify the proposed confidence measure selection and cost modulation methods through extensive experimentation with various aspects using KITTI and challenging outdoor datasets.



(a) Input image (sun flare) [15] (b) Predicted confidence map

(c) Original SGM [8] (d) SGM with our method

Figure 1. Stereo matching results in a challenging environment. (b) shows the predicted confidence map. (c) and (d) show colored disparity maps overlaid on the input image.

1. Introduction

Stereo matching has long been an important topic in computer vision, and its difficulties such as pixel indistinctiveness [14] and occlusions [3] are thoroughly examined in the literature. Based on those researches, several papers [1, 7, 10, 12, 18, 25] confirm the feasibility of detecting mismatched pixels not only to improve the quality of disparity maps [1, 12] but also to leverage a mid-level scene representation (stixel) [18]. Hence, the problem of detecting mismatched pixels becomes more important as the degree of ill-conditioning increases [7] because current solutions usually fail to find correct answers. For example, Fig. 1(a) shows a challenging stereo image captured in an uncontrolled outdoor environment [15]. The sun flare phenomenon severely degrades the quality of the disparity map computed by the semi-global matching (SGM) algorithm [8].

Regarding the detection of mismatched pixels, various confidence measures have been studied and surveyed [2, 3, 10]. A confidence measure is a function of matching costs, disparity values, or image intensities, and it should assign high values to correct matches and low values to false matches so that false matches can be determined by examining the confidence value of a pixel. One commonly used measure is the left-right consistency (LRC) measure [10], assuming that the consistently matched pixels are correct matches. This assumption is a necessary condition because correct matches satisfy the left-right consistency. However, this assumption is not a sufficient condition because consistently matched pixels are not always correct. Thus, the LRC measure does not detect all the incorrect matches, and the detection performance depends on the quality of initial disparity maps and a predefined threshold value.

On the other hand, several papers address the error detection problem from a learning perspective [1, 7] since learning-based approaches have distinct advantages. First, multiple confidence measures can be used jointly to construct a feature vector that demonstrates a better performance than individual confidence measures [1, 7]. Secondly, learning-based approaches can identify which of the

input variables (*i.e*., confidence measures) are important to make the prediction or detection [7, 13] of unreliable pixels. For example, Haeusler *et al*. [7] analyzed the importance of various confidence measures for predicting the correctness of matches using Gini and permutation importance measures.

In this paper, we consider how far a learning-based confidence prediction approach can leverage stereo matching for the practical use in general outdoor environments. The contributions of this paper are twofold. First, we analyze the characteristics of various confidence measures by estimating the permutation importance of each measure in order to select effective confidence measures. Second, we incorporate the predicted confidence value of a pixel into stereo matching algorithms by employing the confidence value in modulating the initial matching cost. Because the matching cost computation is typically the first step for stereo matching, the matching cost modulation step improves the performance of the following steps. We analyze the effect of the cost modulation scheme by applying it to a widely used algorithm for outdoor environments: the semi-global matching method [8][1]. As is shown in Fig. 1(d), the SGM method with proposed measure selection and matching cost modulation exhibits robust results even in a challenging environment.

2. Related Work

Surveys regarding confidence measures are available: Egnal and Wildes [3], Egnal *et al*., [2], and Hu and Modorhai [10]. Most confidence measures are designed to capture well-known difficulties in stereo matching such as occlusions, textureless regions, and depth discontinuities. We review representative confidence measures as well as stereo algorithms leveraged by the confidence information.

Manduchi and Tomasi [14] defined a pixel distinctiveness that is capable of detecting pixels in textureless regions as well as in regions with repeated textures, prior to the stereo matching. Yoon and Kweon [25] extended this concept to the distinctive similarity measure, in which computed matching costs are used for the semi-dense disparity map computation. Sara [20] proposed a confidently stable matching criterion, and this criterion is used for stratified stereo matching [12]. Sabater *et al*. [19] introduced a statistical approach that eliminates unreliably matched pixels. The self-aware matching measure, presented in [17], creates more distinctive matching costs. Gherardi [6] proposed a cost modulation approach that employs the difference of initially computed disparity values with neighboring pixels as a confidence measure. These works primarily focus on improving the quality of the initial or aggregated matching costs.

[1]Note that our method can be easily applied to various stereo methods.

On the other hand, confidence measures are also frequently used in the post-processing step, mainly to detect occluded pixels. Hirschmuller [8] extended the LRC measure by separating the occluded pixels from mismatched pixels based on a threshold value. Min and Sohn [16] proposed an asymmetric criterion for detecting occluded pixels using a single disparity map. Garcia *et al*. [4] defined the credibility map for depth map upsampling in order to assign a low confidence to pixels near the depth discontinuity.

From a machine learning perspective, stereo matching errors can be learned from training data, and the mismatched pixels can be determined using the learned classifiers [1, 7, 11]. Kong and Tao [11] suggested the use of nonparametric techniques to classify matches into three categories: correct, incorrect due to foreground fattening, and the other. Two recent papers [1, 7] applied random decision forests to learn stereo matching errors. They combined diverse confidence measures into a feature vector and demonstrated a superior sparsification performance for Middlebury [22] and KITTI [5] datasets, respectively. In addition, Spyropoulos and Modorhai [1] exhibited that confident pixels can be used as ground control points, and used confidence information further to adjust the smoothing strength in the Markov random field framework. Machine learning techniques are also used to learn the parameters of conditional random fields [21] and to select image features for accurate stereo matching [13].

3. Proposed Method

In this section, we explain the proposed confidence measure selection strategy based on the permutation importance [24] of the random forest. We use random forests because of their robustness against outliers, nonparametric properties, and their ability to rank the importance of input variables [23]. Hence, we suggest upper and lower bounds for the predicted confidence values according to the expected performance of the trained forest. Finally, we propose a matching cost modulation scheme, with the aid of predicted confidence values, that can be easily applied to various stereo algorithms.

3.1. Regression forest-based measure selection

We associate a feature vector $\mathbf{f}$ to a confidence value Q in $[0, 1]$ in the regression forest framework. During training, we build regression forests twice; one is on all confidence measures for selecting influential and powerful measures, and the other is on the selected confidence measures from the first regression forest as shown in Fig. 2. Once the training is complete, we estimate the probability density function $p(Q|\mathbf{f})$ using an ensemble of tree outputs [23] as

$$p(Q|\mathbf{f}) = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} p_t(Q|\mathbf{f}), \tag{1}$$

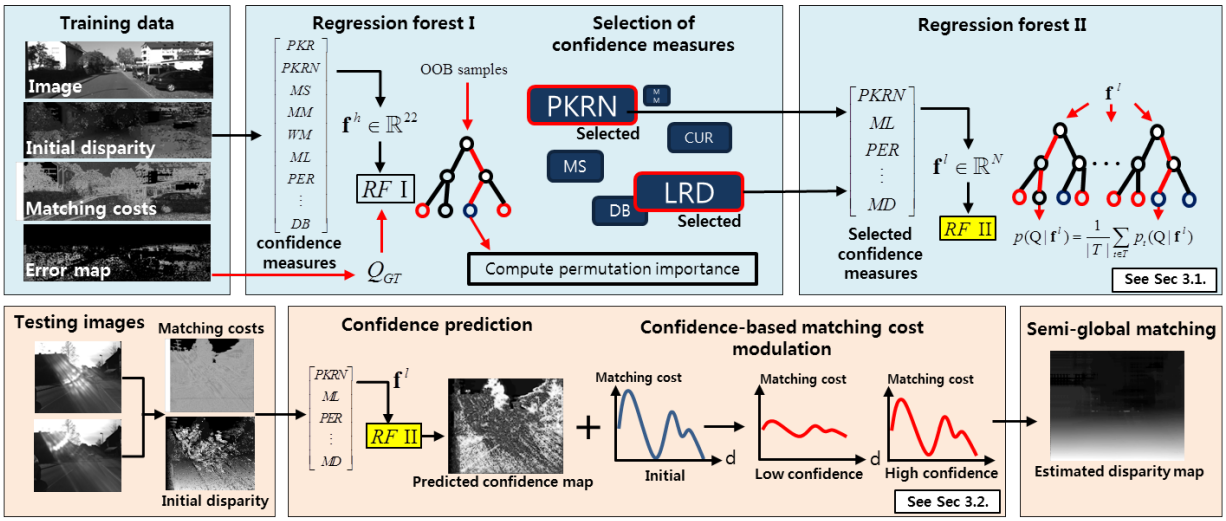


Figure 2. The overall framework with the proposed confidence measure selection and matching cost modulation.

where $\mathbf{f}$ is an input feature vector for testing, $\mathcal{T}$ is a set of tree indices, and $p_t(Q|\mathbf{f})$ is an individual tree posterior that is computed by running down the trained tree.

Feature vector: We define a high-dimensional feature vector for a pixel using a number of confidence measures:

$$\mathbf{f}^h = [f_1, f_2, f_3, ..., f_{21}, f_{22}]^\top, \tag{2}$$

in which f_i is the scalar value computed by the i^{th} confidence measure. For the sake of notational simplicity, we omit the index of a pixel. f_1 to f_{10} are calculated from confidence measures that utilize the matching costs: the peak ratio, naive peak ratio, matching score, maximum margin, winner margin, maximum likelihood, perturbation, negative entropy, left-right difference, and local curvature measures, which are all explained in [2, 3, 7, 10]. f_{11} to f_{20} are calculated from confidence measures that use an initial disparity map estimated by the winner-takes-all strategy. Here, the variances of the disparity values in a local window for four different scales are used for f_{11} to f_{14}, the distance to discontinuity is used for f_{15}, the median deviations of disparity values in four different scales are used for f_{16} to f_{19}, and the left-right consistency measure is used for f_{20}. For f_{21}, we use the magnitude of the image gradients. Lastly, the distance to the border measure [1] is used for f_{22} because pixels in the leftmost columns do not have correspondences when the left image is the reference image.[2] We exclude the distinctiveness [14], self-aware matching [17], and the distinctive similarity measures [25] because they are computationally demanding when used to construct additional matching costs from the reference image.

Selection of confidence measures: After constructing regression forests for $\mathbf{f}^h$ with all confidence measures, we make a set of important confidence measures using the permutation importance accuracy measure [23, 24]. This helps not only to understand the importance of each confidence measure but also to design better prediction models. When OOB_t denotes the set of out-of-bag sample indices for tree t, which are not used for constructing the tree t, and $|\text{OOB}_t|$ is the cardinality of OOB_t, the variable (*i.e.*, confidence measure) importance for tree t is measured by

$$\begin{aligned}\text{VI}_t(f_j) &= \frac{1}{|\text{OOB}_t|}\sum_{i\in\text{OOB}_t}(Q_i - \hat{Q}_i)^2 \\ &\quad - \frac{1}{|\text{OOB}_t|}\sum_{i\in\text{OOB}_t}(Q_i - \hat{Q}_{i,\pi_j})^2.\end{aligned} \tag{3}$$

Here, the first term in Eq. (3) is the squared error between the predicted confidence $\hat{Q}_i$ and the ground truth confidence Q_i. Similarly, the latter term is the squared error between Q_i and $\hat{Q}_{i,\pi_j}$ that is the predicted confidence of the sample generated by permuting the j^{th} variable of sample i randomly with the j^{th} variable of the other out-of-bag sample. Therefore, Eq. (3) is the difference between the prediction errors before and after permuting the j^{th} variable.[3] $\text{VI}_t(f_j) = 0$ implies that the j^{th} variable does not help to distinguish correct and false matches at all.

The importance of the j^{th} variable is aggregated for all trees and normalized using the number of trees or the standard deviation of the prediction errors. Then, we list confidence measures in descending order according to their variable importance values. Finally, we select the most important N confidence measures and define a lower dimensional feature vector $\mathbf{f}^l$ as

$$\mathbf{f}^l = [\hat{f}_1, \hat{f}_2, ..., \hat{f}_N]^\top, \tag{4}$$

where $\hat{f}_i$ indicates the value of the i^{th} important confidence measure. We then again construct the regression forest on

[2] Each confidence measure is explained in the supplementary material.

[3] For example, when $j = 1$, Eq. (3) evaluates the importance of the peak ratio measure.

the training data using $\mathbf{f}^l$. Here, the trained regression forest using $\mathbf{f}^l$ improves the prediction efficiency and accuracy as follows. First, we just need to evaluate a smaller number of confidence measures to construct feature vectors and to predict the correctness of a match, and a less number of trees can be used to construct the regression forest. Second, since unreliable confidence measures are excluded, it is possible to design a better prediction model.

Confidence rescaling: The predicted confidence value, computed from the regression forest, can be directly used for stereo matching algorithms. However, we manipulate the predicted confidence values as

$$\hat{Q}(\mathbf{p}) = Q(\mathbf{p})\overline{Q} + (1 - Q(\mathbf{p}))\underline{Q}, \tag{5}$$

where $Q(\mathbf{p})$ is a raw predicted confidence value for pixel $\mathbf{p}$ from the regression forest. We define the interval $[\underline{Q}, \overline{Q}]$ to be the range for the manipulated confidence value. The upper and lower bounds, $\overline{Q}$ and $\underline{Q}$, are defined by employing outlier probabilities of positive and negative predictions of learned forests as $\overline{Q} = 1 - \mathrm{FP}/(\mathrm{FP} + \mathrm{TP})$ and $\underline{Q} = \mathrm{FN}/(\mathrm{FN} + \mathrm{TN})$. Here, TP, FP, TN, and FN are the numbers of true positive, false positive, true negative, and false negative samples that are evaluated by using out-of-bag samples in the training step. Then, if the predicted confidence value is 0, the predicted value is increased to the outlier probability of negative predictions.

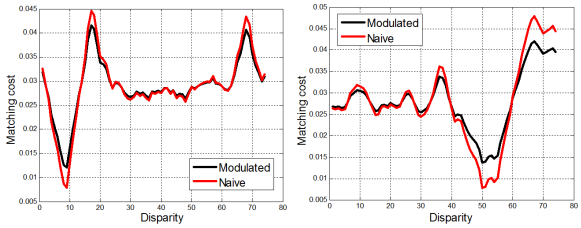
Moreover, as the expected performance of the regression forest decreases, the gap between the two bounds also decreases by definition, making correct and incorrect matches have similar manipulated confidence values. On the other hand, as the expected performance increases, the gap between the two bounds also increases, allowing correct and incorrect matches to have distinct manipulated confidence values. This simple modification makes the proposed confidence prediction less sensitive to the quality of the learned forests or classifiers.

3.2. Confidence-based matching cost modulation

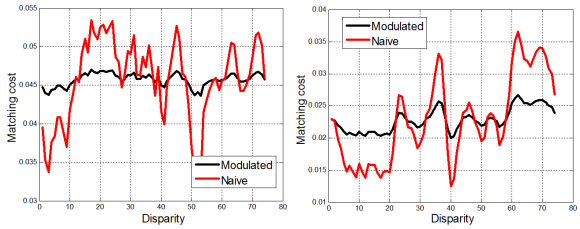
We incorporate the predicted confidence value into stereo matching by exploiting the confidence value for the modulation of matching costs, mainly because the confidence value represents the reliability of matching costs. When $C(\mathbf{p}, d)$ denotes the matching cost of pixel $\mathbf{p}$ for a disparity d, it is modulated as

$$\hat{C}(\mathbf{p}, d) = \hat{Q}(\mathbf{p})C(\mathbf{p}, d) + (1 - \hat{Q}(\mathbf{p}))\sum_{k \in \mathcal{D}} \frac{C(\mathbf{p}, k)}{|\mathcal{D}|}. \tag{6}$$

Here, the latter term is the mean of the matching costs computed for all possible disparity hypotheses $\mathcal{D}$. Therefore, as the probability of correctness increases, the data term increasingly depends on the original matching costs. Otherwise, the data term becomes the mean value of matching costs. We use the mean value instead of a uniform distribution, such as $1/|\mathcal{D}|$, in order to retain the sum of the original



(a) Modulated matching costs for highly confident pixels



(b) Modulated matching costs for unreliable pixels

Figure 3. Modulated matching costs. Matching costs of unreliable pixels in (b), which are not likely to give correct solutions, are flattened depending on the predicted confidence values whereas highly confident pixels have similar costs to its original matching costs.

matching costs. Figure 3 shows the result of the matching cost modulation, in which matching costs of confident pixels appear similar to the initial matching costs whereas unreliable pixels are flattened. Therefore, unreliable pixels can be easily dominated by more confident pixels in the optimization step. It should be noted that the proposed cost modulation can be used for any stereo algorithm.

Robust outdoor stereo matching: We apply the proposed confidence-based cost modulation scheme to the semi-global matching (SGM) algorithm [8]. The SGM is popular for outdoor environments because of its efficiency and accuracy. In essence, the stereo matching problem is considered as the energy minimization problem as

$$\begin{aligned} E(D) \quad &= \sum_{\mathbf{p}} \hat{C}(\mathbf{p}, d_{\mathbf{p}}) + \sum_{\mathbf{q} \in N_{\mathbf{p}}} P_1 T[|d_{\mathbf{p}} - d_{\mathbf{q}}| = 1] \\ &+ \sum_{\mathbf{q} \in N_{\mathbf{p}}} P_2 T[|d_{\mathbf{p}} - d_{\mathbf{q}}| > 1], \end{aligned} \tag{7}$$

in which the first term represents the pixel-wise matching cost modulated by the proposed scheme, the second term gives the penalty P_1 for the pixels having small disparity differences with neighboring pixels, and the third term gives the large penalty P_2 (in general, $P_2 > P_1$) for the pixels having disparity differences larger than 1 with neighboring pixels. We compute per-pixel matching costs using the truncated census-based Hamming distance and the truncated gradient difference as

$$\begin{aligned} C(\mathbf{p}, d) \quad &= \tfrac{\alpha}{|N_{\mathbf{p}}|} \min\Big(\sum_{\mathbf{q} \in N_{\mathbf{p}}} XOR(B^L(\mathbf{p}, \mathbf{q}), B^R(\mathbf{p}_d, \mathbf{q}_d)), \tau_c\Big) \\ &+ \tfrac{1-\alpha}{255} \min(|\nabla I^L(\mathbf{p}) - \nabla I^R(\mathbf{p}_d)|, \tau_g), \end{aligned} \tag{8}$$

where $\mathbf{p}_d$ is the pixel in the right image corresponding to $\mathbf{p}$ in the left image when the disparity of $\mathbf{p}$ is d, τ_g is the truncation value for the gradient difference, and τ_c is truncation value for the census-based cost. We use two robust measures that are less prone to error under the change of global illumination, which frequently happens in outdoor environments. $B(\mathbf{p}, \mathbf{q})$ is a binary transform function. It returns 1 if the intensity of $\mathbf{p}$ is larger than $\mathbf{q}$ and 0 otherwise. The superscript L and R indicate the left and right image, respectively. Here, we normalize the census-based cost by the number of neighboring pixels and the input image by the factor of 255 to handle the relative scale difference between two terms.

Afterward, we modulate the matching costs by utilizing the confidence value of a pixel and minimize the energy function by aggregating matching cost to sixteen directions in a recursive manner. In contrast to the original paper [8], we do not apply post-processing algorithms, such as left-right consistency checking and hole filling, in order to clearly observe the improvement from the cost modulation. In addition, we do not specially handle occluded pixels since occluded pixels are likely to have flattened modulated matching costs and, therefore, the quality degradation due to occluded pixels is not significant.

4. Experimental Results

We evaluate various confidence prediction methods including learning-based approaches [1, 7] and the proposed method with various datasets [5, 15, 22]. In addition, we evaluate the relative improvement of two stereo matching algorithms, SGM [8] and fast cost volume filtering [9] methods, by adopting the proposed modulation scheme. We also show disparity maps computed from challenging outdoor environments [15] containing sun flares, rain blur, and various weather conditions in order to verify the robustness of the proposed method.

Parameter settings: To compute pixel-wise matching costs, we used a 5×5 local window for census-based costs. The truncation values τ_c and τ_g were set to 11 and 9, respectively, which take the role of basic M-estimators to limit the influence of mismatches during an aggregation or optimization step [22]. Empirically, we recommend to set the range of a truncation value between 9 and 15. α is set to 0.6. In addition, we aggregated initial matching costs with a 5×5 box-filter to increase the accuracy of the initial disparity map. To construct the regression forests, we set the number of trees to 30. For the penalty terms in the SGM, we set $P_1 = 0.008$ and $P_2 = 0.126$, which have different scales when compared with the original paper because we normalized the matching costs. For the cost volume filtering approach, we used a 19×19 local window and set the regularization parameter ϵ to 0.01^2 as recommended in the paper [9]. For computing confidence measures in different scales such as median deviation values, we used four different sizes of local windows, 5×5, 7×7, 9×9, and 11×11. For training, we used eight stereo pairs, 43, 71, 82, 87, 94, 120, 122, and 180$^{\text{th}}$ frames, from the KITTI training data as in [7]. Hence, trained forests are used for testing of both KITTI and challenging datasets [15] because the latter does not provide ground truth information. In addition, we used the half of Middlebury 2005 and 2006 datasets for training, which are Aloe, Art, Baby1, Baby3, Bowling1, Cloth1, Cloth3, Dolls, Lampshade1, Laundry, Reindeer, Rocks2, and Wood2 datasets, and used four standard datasets, Tsukuba, Venus, Teddy, and Cones, for testing.

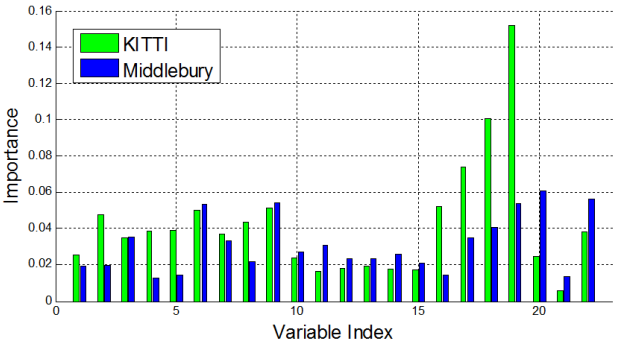
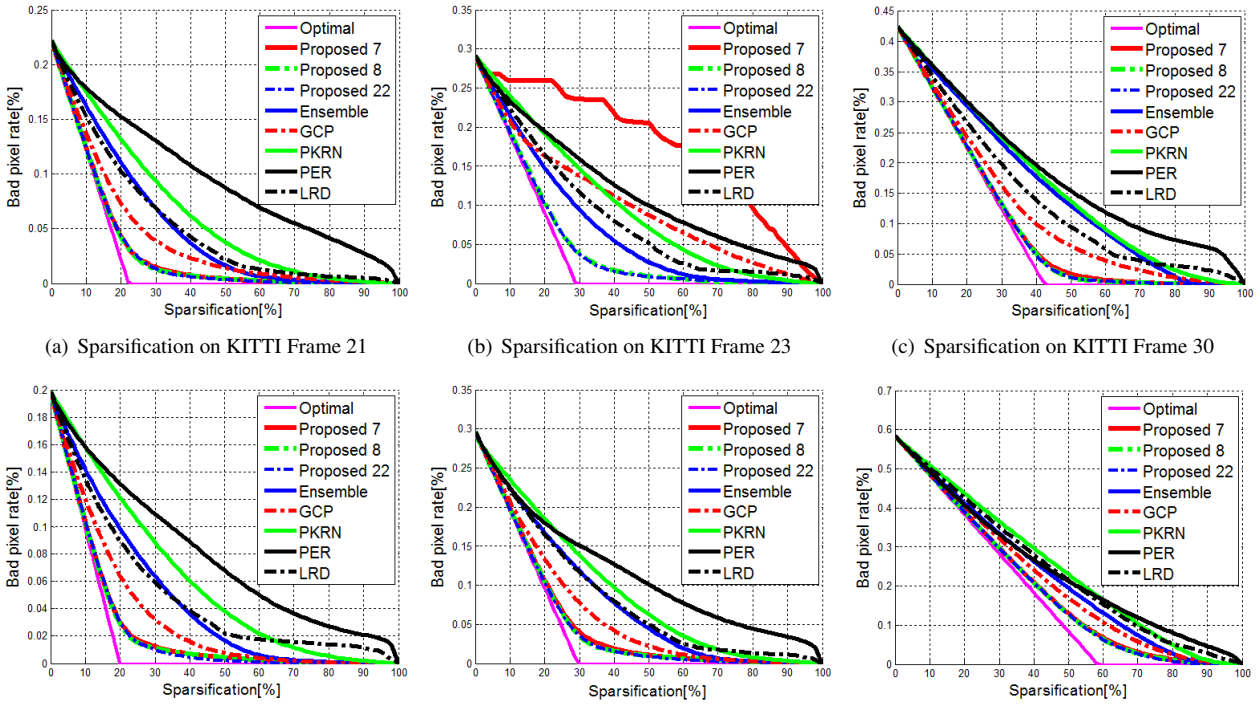


Figure 4. The importance of variables (*i.e.*, confidence measures) for KITTI and Middlebury datasets. Variable indices 1 to 10 are computed from matching costs, 11 to 20 are from the initial disparity map, 21 is from the image, and 22 is from the prior knowledge.

4.1. Learning-based confidence analysis

We analyzed the characteristics of various confidence measures as well as the detection performance of them. Firstly, we compare the importance of variables in Fig. 4 to examine which variables play the most significant roles in determining the correctness of matches. We empirically set the number of selected features N to 8 because the selected measures performed similarly with 22 confidence measures as long as the number of selected measures is greater than or equal to 8. However, the use of fewer numbers than 8 demonstrated poor performance for some test frames (7% of all frames when 7 confidence measures are selected), as can be seen in Fig. 5(b). We presume that this is due to the limited amount of training data (8 images are used for training out of 194 images) or overfitting, but this phenomenon did not happen when using more than 7 measures.

For the KITTI dataset, selected 8 confidence measures are the median disparity deviation values in four different scales ($f_{19}, f_{18}, f_{17}, f_{16}$), the left-right difference [10], the maximum likelihood measure, the naive peak ratio measure, and the negative entropy measure. For the Middlebury dataset, the left-right consistency measure, the distance to the border, the left-right difference, the median disparity deviation values in three different scales (f_{19}, f_{18}, f_{17}), the



(a) Sparsification on KITTI Frame 21 (b) Sparsification on KITTI Frame 23 (c) Sparsification on KITTI Frame 30

(d) Sparsification on KITTI Frame 50 (e) Sparsification on KITTI Frame 121 (f) Sparsification on KITTI Frame 123

Figure 5. Comparison of sparsification curves for selected images. We drew sparsification curves for naive peak ratio (PRKN) [10], left right difference (LRD) [10], and perturbation (PER) [7] measures that show superior performance among individual confidence measures. The sparsification curve for an ideal confidence map is described as optimal. Proposed 7, 8, 22, where 7, 8, and 22 are the numbers of selected confidence measures from 22 confidence measures, show similar tendency in spite of different numbers of confidence measures except frame 23. Here, GCP [1] and Ensemble [7] use 8- and 7-dimensional feature vectors, respectively, but they show poorer performance than ours.

maximum likelihood measure, and the matching score measure are selected as the best 8 measures. The image gradient exhibited the least importance, meaning the correctness of a match and the magnitude of an image gradient are correlated weakly. Hence, the difference of selected confidence measures implies that the detection performance of each confidence measure can vary for different datasets [4]. In contrast to [7], the disparity variance showed a small importance value. In the KITTI dataset, the left-right consistency showed a poor performance, because, many occluded pixels do not have ground truth information.

To analyze the detection performance of various confidence measures, we use the sparsification curve and its area under curve (AUC) value as in [7, 10]. The sparsification curve draws the change of bad pixel rates while removing least confident pixels from the disparity map, therefore, it is possible to observe the tendency of prediction errors. Sparsification curves for selected frames are shown in Fig. 5, which confirm the superiority of the proposed approaches. This is because previous feature vectors [1, 7] are designed for a specific similarity measure or a stereo algorithm. Therefore, prediction performance can vary depending on used similarity measures, parameters, and characteristics of the dataset. In addition, Fig. 6 describes the AUC values using six different methods for all KITTI training frames. Here, the gap between an AUC value and the optimal value describes the detection performance of the method, and the gap tends to increase as the optimal AUC value increases. It means that the detection problem becomes more difficult as the quality of estimated disparity maps decreases. Nevertheless, the proposed methods show consistently superior detection performance than others.

We also checked the running time for the prediction step in Tab. 1. As the number of trees in the regression forest increases, the prediction time tends to increase linearly. However, there was not a significant difference in prediction performance as long as the number of trees is larger than 10. In addition to the prediction time, the feature vector construction requires about 7 ms for the matching cost-based measures and about 5 ms for the disparity map-based confidence metrics. Therefore, for the 8-dimensional feature vector, feature construction time for the KITTI dataset takes about 50 ms (using C++ with an i7 3.5GHz single core CPU) in addition to the time required for the prediction.

[4] It will be also valuable to see the difference of selected features for various similarity measures and parameters.

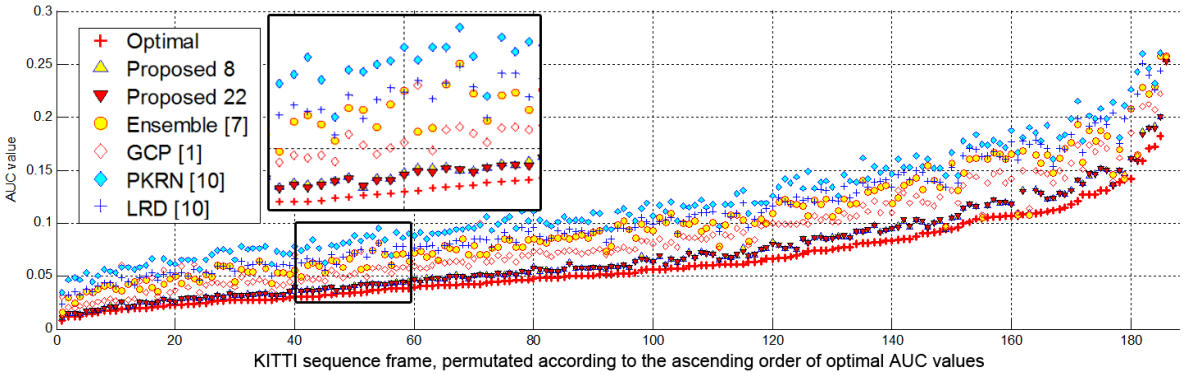


Figure 6. AUC values in the ascending order according to optimal AUC values. For clear comparison, we selected six methods that show superior performance than others. Ensemble [7] and GCP [1] are based on a learning technique, and PKRN and LRD [10] are based on matching cost information.

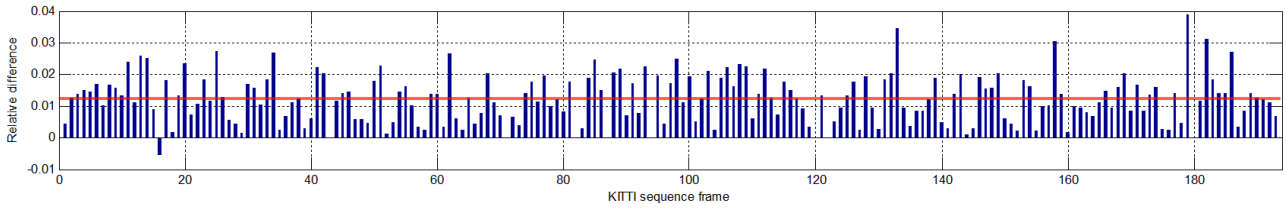


Figure 7. Relative performance improvement compared with the SGM algorithm: each bar indicates the performance improvement for the modified SGM method, which results from the confidence-based matching cost modulation step (only difference between the algorithms), and horizontal line indicates average improvement in terms of the bad pixel rate.

Table 1. The prediction time w.r.t. the number of trees.

# of trees	10	15	20	25	30	50
Time (ms)	358	562	801	1050	1262	2213

4.2. Stereo performance improvement analysis

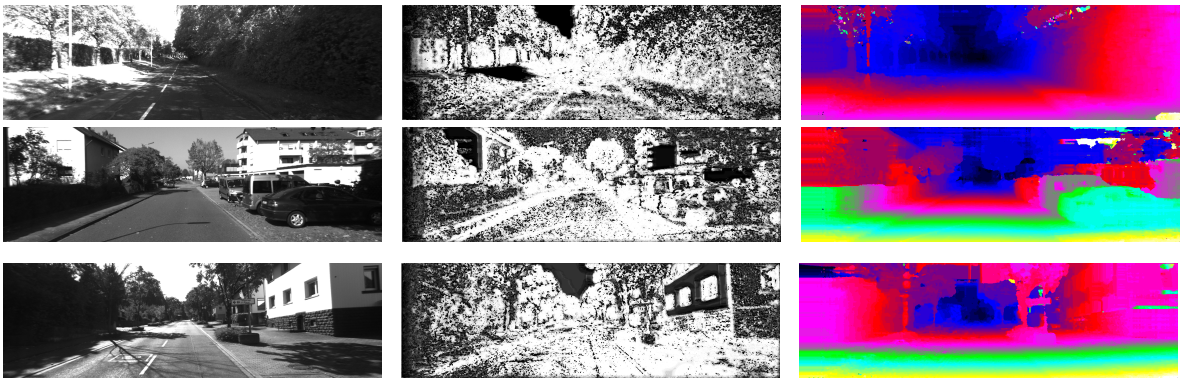
We evaluate the relative performance improvement of stereo algorithms, SGM [8] and fast cost volume filtering [9], leveraged by our cost modulation scheme. We used the KITTI dataset for evaluating the SGM method; the modified SGM method–which uses modulated matching costs–improved the disparity map's accuracy by 1.22% for the 186 images on average. The overall improvement is shown in Fig. 7, in which we see the proposed approach consistently improves the SGM. The average initial error was 10.61% (in our implementation) including occluded regions and the modified version showed 9.38% of errors without post-processing. Figure 8 shows computed confidence maps and disparity maps for a few images. Moreover, we evaluated the fast cost volume filtering algorithm to show that the proposed method is not limited to a particular algorithm or a dataset. The modified fast cost volume filtering approach reduced bad pixel rates from 5.05% to 4.38% in non-occluded regions for four standard datasets without post-processing. Here, we used same parameters with the SGM setting. The proposed method improves the local algorithm moderately compared to the SGM, which is because that the Middlebury dataset contains a smaller number of unreliable pixels than the KITTI dataset.

4.3. Stereo evaluation in challenging environments

The application of a confidence-based matching cost modulation to a stereo algorithm not only improves the stereo matching's accuracy but also enhances robustness under a variety of difficulties presented in outdoor environments. Therefore, we evaluated the proposed method using challenging datasets [15] where 11 selected video sequences are posted on the website [5] with the SGM results. The challenging datasets were captured by different cameras and settings–the image resolution is 656×541, and the baseline is about 30 cm which is smaller than that of the KITTI dataset, though we used the KITTI training dataset to construct the regression forest while ignoring these differences due to the absence of ground truth data.

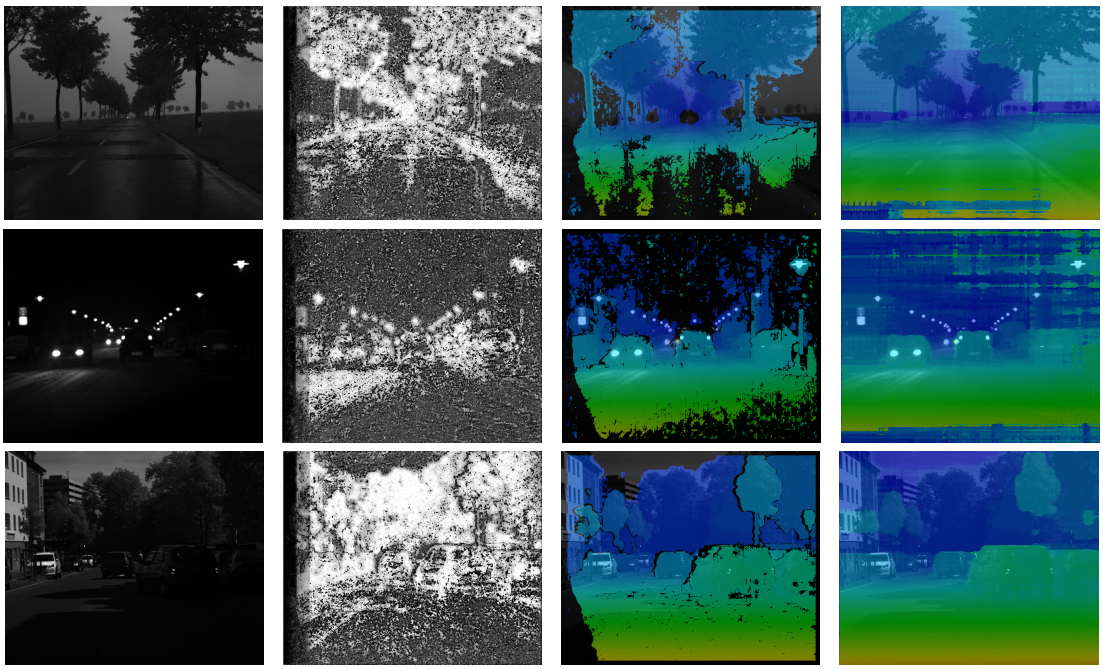
The selected images were captured under a substantial variety of different weather conditions, motions, and depth layers, as shown in Fig. 9(a). The input image in the first row is captured under rain blur where pixels are blurred differently for left and right images. The image in the second row is captured on a snowy night where the light sources are limited to street lamps and car headlights. The third row shows the reflecting car dataset, which contains a couple of cars in the scene with highly reflective surfaces. Figure

[5] http://hci.iwr.uni-heidelberg.de//Benchmarks/document/Challenging_Data_for_Stereo_and_Optical_Flow/



(a) Input images (b) Predicted confidence maps (c) SGM with our methods

Figure 8. Selected results for the KITTI dataset.



(a) Left input images (b) Predicted confidence (c) SGM results [8] (d) SGM with our methods

Figure 9. Results of the proposed method under various challenging circumstances [15]. Images overlaid with disparity maps are encoded by the toolbox [15]: top row–blur due to raindrops on the windshield, second row–snowy night, third row–reflecting cars in the shadow. More results are given in the supplementary material.

9(b) shows the predicted confidence maps of three images, in which the intensity value is proportional to the pixel's confidence level. Figure 9(c) and 9(d) show the SGM results and the proposed results for these input images, respectively. Compared to the SGM, the proposed method demonstrates robust results despite aforementioned difficulties. The main reason for this improvement is that challenging datasets contain a large amount of unreliable pixels that violate underlying assumptions of binocular stereo matching such as brightness constancy. The experimental results verify that the detection of unreliable pixels becomes particularly important for these kinds of input images.

5. Conclusion

We established the relationship between learning-based confidence measures and stereo matching algorithms. First, we demonstrated the selection of powerful confidence measures based on the permutation importance in the regression forest framework. Second, we presented a generalized approach for improving the accuracy and robustness of stereo algorithms with the confidence-based matching cost modulation scheme. The stereo algorithms leveraged by the proposed methods exhibited accurate and robust results in public datasets as well as challenging outdoor environments. We share our code at http://cvl.gist.ac.kr/project.

Acknowledgement

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT & Future Planning (NRF-2012R1A1A1010871), Global Frontier Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT & Future Plannig (NRF-2013M3A6A3075453), and ICT R&D program of MSIP/IITP [B0101-15-0552, Development of Predictive Visual Intelligence Technology].

References

[1] N. K. A. Spyropoulos and P. Mordohai. Learning to detect ground control points for improving the accuracy of stereo matching. In *CVPR*, 2014.

[2] G. Egnal, M. Mintz, and R. P. Wildes. A stereo confidence metric using single view imagery with comparison to five alternative approaches. *Image and Vision Computing*, 22(12):943 – 957, 2004.

[3] G. Egnal and R. P. Wildes. Detecting binocular half-occlusions: Empirical comparisons of five approaches. *PAMI*, 24(8):1127–1133, 2002.

[4] F. Garcia, B. Mirbach, B. E. Ottersten, F. Grandidier, and A. Cuesta. Pixel weighted average strategy for depth sensor data fusion. In *ICIP*, pages 2805–2808, 2010.

[5] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013.

[6] R. Gherardi. Confidence-based cost modulation for stereo matching. In *ICPR*, pages 1–4, Dec 2008.

[7] R. Haeusler, R. Nair, and D. Kondermann. Ensemble learning for confidence measures in stereo vision. In *CVPR*, pages 305–312, 2013.

[8] H. Hirschmuller. Stereo processing by semiglobal matching and mutual information. *PAMI*, 30(2):328–341, 2008.

[9] A. Hosni, C. Rhemann, M. Bleyer, C. Rother, and M. Gelautz. Fast cost-volume filtering for visual correspondence and beyond. *PAMI*, 35(2):504 – 511, 2013.

[10] X. Hu and P. Mordohai. A quantitative evaluation of confidence measures for stereo vision. *PAMI*, 34(11):2121–2133, 2012.

[11] D. Kong and H. Tao. A method for learning matching errors in stereo computation. In *BMVC*, 2004.

[12] J. Kostková. Stratified dense matching for stereopsis in complex scenes. In *BMVC*, pages 339–348, 2003.

[13] M. S. Lew, T. S. Huang, and K. Wong. Learning and feature selection in stereo matching. *PAMI*, 16(9):869–881, 1994.

[14] R. Manduchi and C. Tomasi. Distinctiveness maps for image matching. In *ICIAP*, pages 26–31, 1999.

[15] S. Meister, B. Jähne, and D. Kondermann. Outdoor stereo camera system for the generation of real-world benchmark data sets. *Optical Engineering*, 51(02):021107–1–021107–6, 2012.

[16] D. B. Min and K. Sohn. An asymmetric post-processing for correspondence problem. *Sig. Proc.: Image Comm.*, 25(2):130–142, 2010.

[17] P. Mordohai. The self-aware matching measure for stereo. In *ICCV*, pages 1841–1848, 2009.

[18] D. Pfeiffer, S. Gehrig, and N. Schneider. Exploiting the power of stereo confidences. In *CVPR*, pages 297–304, 2013.

[19] N. Sabater, A. Almansa, and J.-M. Morel. Meaningful matches in stereovision. *PAMI*, 34(5):930–942, 2012.

[20] R. Sara. Finding the largest unambiguous component of stereo matching. In *ECCV*, pages 900–914, 2002.

[21] D. Scharstein. Learning conditional random fields for stereo. In *CVPR*, 2007.

[22] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 47:7–42, 2002.

[23] L. B. Statistics and L. Breiman. Random forests. In *Machine Learning*, pages 5–32, 2001.

[24] G. Ulrike. Variable importance assessment in regression: Linear regression versus random forest. *The American Statistician*, 63(4):308–319, 2009.

[25] K.-J. Yoon and I.-S. Kweon. Distinctive similarity measure for stereo matching under point ambiguity. *CVIU*, 112(2):173–183, 2008.