Evaluation of Stereo Confidence Indoors and Outdoors

Xiaoyan Hu
Stevens Institute of Technology
Hoboken, New Jersey, USA
xhu2@stevens.edu

Philippos Mordohai
Stevens Institute of Technology
Hoboken, New Jersey, USA
mordohai@cs.stevens.edu

Abstract

We present an extensive evaluation of 13 confidence metrics for stereo matching that compares the most widely used metrics as well as four novel techniques proposed here. We begin by categorizing the methods according to which aspects of stereo computation they take into account and, then, assess their strengths and weaknesses. The evaluation is conducted on indoor and outdoor datasets with ground truth and measures the capability of each confidence metric to rank depth estimates according to their likelihood for being correct, to detect occluded pixels and to generate low-error depth maps by selecting among multiple hypotheses for each pixel. We believe that such an evaluation is missing from the rapidly maturing stereo literature and that our findings will be helpful to researchers in binocular and multi-view stereo.

1. Introduction

While several confidence metrics for stereo matching have been proposed in the literature and benchmarks with ground truth have been available for years, the criteria for selecting a confidence metric and the relative merits of different metrics have not been investigated thoroughly. We study these issues using indoor and outdoor stereo imagery with ground truth [18, 20]. Our goal is to categorize the different methods and to shed light on their performance according to specific criteria described below.

We focus on methods that estimate the confidence of disparity assignments in a winner-take-all (WTA) setting, without considering neighboring pixels or global information. Cost or similarity values for each disparity hypothesis are computed and a disparity map is generated by selecting the hypothesis with the minimum cost, or maximum similarity, for each pixel. The cost values for all hypotheses are used as input to 13 confidence methods that assign confidence values to the selected disparities. We require that these confidence values have the following properties:

- Be high for correct disparities and low for errors. If matched pixels were ranked in order of decreasing confidence, all errors should be ranked last. The ranking should also be correct for pixels of special interest, such as those near discontinuities.
- Be able to detect occluded pixels.
- Be useful for selecting the true disparity among hypotheses generated by different matching strategies.

We have evaluated the degree that each method satisfies the above criteria using a set of experiments on matching volumes computed using both cost and similarity functions aggregated in square windows of various sizes. We have performed these tests on binocular stereo images in the rectified canonical configuration [18] and on multi-baseline imagery collected outdoors [20].

Since stereo matching is known to be prone to errors, the capability to predict where these errors occur is desirable. A WTA framework is appropriate for our evaluation because, in general, confidence for a particular match cannot be estimated using global optimization methods, such as Markov Random Fields. This is due to the impracticality of estimating the minimum energy labeling that assigns a particular disparity to a given pixel and repeating the computation over all disparity values for all pixels. Confidence estimation, on the other hand, is possible when dynamic programming is used for optimization. Gong and Yang [10] defined the reliability of a disparity assignment (match) for a pixel as the cost difference between the best path that does not pass through the match and the best path that passes through it. This method can be viewed as a generalization of the Peak Ratio metric described in Section 3.

In summary, the contributions of this paper are:

- A classification of several confidence measures.
- A set of criteria for evaluating them.
- New confidence metrics (NLM, $PKRN$, $WMNN$ and LRD) that often perform better than conventional methods.
- Quantitative and qualitative comparisons of 13 confidence methods on indoor and outdoor imagery.

2. Related Work

Our work complements surveys on binocular [18, 3] and multi-view [19] stereo, as well as on specific aspects of stereo, such as cost functions [12], cost aggregation [9, 21] and color spaces [1]. These efforts, aided by standardized benchmarks [18, 19], have led to significant progress and more principled design of stereo algorithms. The renewed interest in confidence estimates for multi-view stereo [15, 2, 4] and the lack of recent, comprehensive surveys on confidence estimation are the motivations for our work. For clarity, we present the confidence metrics in Section 3. Here, we only discuss methods that combine multiple confidence estimation techniques.

Approaches that combine multiple confidence metrics [5, 11, 13] aim at rejecting erroneous matches to obtain error-free quasi-dense depth maps. Different indicators of matching errors, such as left-right consistency, flatness of the cost curve and the matching cost itself are combined using heuristic rules to detect mismatches. Results show that these methods work reasonably well, but provide little information on the contribution of each element and the suitability of different techniques to specific types of imagery.

Arguably, the most significant and most recent comparison of stereo confidence metrics was carried out by Egnal et al. [6]. Five metrics, four of which are included here, were evaluated against single-view stereo on predicting matching errors on three stereo pairs with ground truth.

A relevant problem to the detection of mismatches is the detection of occluded pixels. An evaluation of four methods for occlusion detection was performed earlier by Egnal and Wildes [7]. Ideally, occlusion detection should not be performed locally since occlusion is a result of long range interaction between remote surfaces. We are more interested in the fundamental question whether occlusion can be detected locally based on confidence. Following the common assumption that occluded pixels can be identified due to high matching cost, we evaluate matching cost as an indicator of occlusion and compare it with the other confidence metrics.

3. Confidence Metrics

Before describing the confidence metrics, let us introduce the notation used throughout. The experiments were carried out using the sum of absolute color differences in RGB (SAD), which is a cost metric, and zero-mean normalized cross-correlation (NCC), which measures similarity. For NCC, the mean is computed separately per RGB channel, but a single variance is estimated for the $3 \times N \times N$ vector obtained by stacking all the elements in the window after the mean RGB values have been removed. This reduces sensitivity to image regions with small variance in any one channel. Square $N \times N$ windows are used in all cost computations. For uniformity, NCC is converted to a cost function by replacing it with 1-NCC. SAD values are normalized by the number of pixels in the window.

The cost value (SAD or 1-NCC) assigned to a disparity hypothesis d for a pixel (x, y) is denoted by $c(x, y, d)$ or $c(d)$, if pixel coordinates are unambiguous. The minimum cost for a pixel is denoted by c_1 and the corresponding disparity value by d_1; $c_1 = c(d_1) = \min c(d)$. We also define c_2 to denote the *second smallest value* of the cost that occurs at disparity d_2, as well as c_{2m} at disparity d_{2m} to denote the *second smallest local minimum*. The default reference image for a binocular pair is the left one. If the right image is used as reference, $c_R(x_R, y, d_R)$ denotes the cost function, with $d_R = -d$.

The disparity map for the reference image is denoted by $D(x, y)$ and is obtained by simply selecting the disparity with the minimum cost for each pixel. A confidence map is denoted by $C_{METH}(x, y)$, where $METH$ corresponds to the confidence method being used.

3.1. Categorization of Confidence Metrics

We can now introduce the confidence metrics grouped according to the aspects of cost they consider.

1. Matching Cost The matching cost is used as a confidence measure.
The **Matching Score Metric (*MSM*)**: is the simplest confidence metric [6] and serves as a baseline in our experiments.

$$C_{MSM} = -c_1 \tag{1}$$

2. Local properties of the cost curve The shape of the cost curve around the minimum (the sharpness or flatness of the valley) is an indication of certainty in the match.
Curvature (*CUR*): has been evaluated in [6] and is widely used in the literature. It is defined as:

$$C_{CUR} = -2c(d_1) + c(d_1 - 1) + c(d_1 + 1) \tag{2}$$

If $d_1 - 1$ or $d_1 + 1$ are outside the disparity range, the available neighbor of the minimum is used twice.

3. Local minima of the cost curve The presence of other strong candidates is an indication of uncertainty, while their absence of certainty.
Peak Ratio (*PKR*): Among several equivalent formulations [6, 11], we have implemented PKR as:

$$C_{PKR} = \frac{c_{2m}}{c_1} \tag{3}$$

We have also implemented a naive version ***PKRN***, which does not require the numerator to be a local minimum.

$PKRN$ can be viewed as a combination of PKR and CUR that assigns low confidence to matches with flat minima or strong competitors.

$$C_{PKR} = \frac{c_2}{c_1} \tag{4}$$

The margin between c_1 and c_2 is also an indication of confidence. After a nonlinear transformation for visualization purposes, we define the **Nonlinear Margin (*NLM*)** as:

$$C_{NLM} = e^{\frac{c_2 - c_1}{2\sigma^2_{NLM}}} \tag{5}$$

4. The entire cost curve These methods convert the cost curve to a probability distribution function over disparity.

The **Probabilistic Metric (*PRB*)** operates on a *similarity* function, normalizing the values to sum to unity. It is only used on NCC here.

$$C_{PRB} = \frac{NCC(d_1)}{\sum_d NCC(d)} \tag{6}$$

The **Maximum Likelihood Metric (*MLM*)** is based on [14], in which SSD was used as the cost function. We generalize the approach to other cost functions and obtain a probability density function for disparity given cost by assuming that the cost follows a normal distribution and that the disparity prior is uniform. After normalization, C_{MLM} is defined as follows.

$$C_{MLM} = \frac{e^{-\frac{c_1}{2\sigma^2_{MLM}}}}{\sum_d e^{-\frac{c(d)}{2\sigma^2_{MLM}}}} \tag{7}$$

MLM assumes that the matching cost can attain the ideal value of 0. Merrell et al. [15] proposed a variant, termed here **Attainable Maximum Likelihood (*AML*)**, that models the cost for a particular pixel using a Gaussian distribution centered at the minimum cost value that is actually achieved for that pixel (c_1 in our notation).

$$C_{AML} = \frac{e^{-\frac{(c_1 - c_1)^2}{2\sigma^2_{AML}}}}{\sum_d e^{-\frac{(c(d) - c_1)^2}{2\sigma^2_{AML}}}} \tag{8}$$

(The numerator is always 1, but is shown here for clarity.)

The **Negative Entropy Metric (*NEM*)** was proposed by Scharstein and Szeliski [17]. Cost values are converted to a *pdf*, the negative entropy of which is used as a measure of confidence.

$$\begin{aligned} p(d) &= \frac{e^{-c_1}}{\sum_d e^{-c(d)}} \\ C_{NEM} &= -\sum_d p(d) \log p(d) \end{aligned} \tag{9}$$

The **Winner Margin (*WMN*)** was also proposed in [17]. It is a hybrid method that normalizes the difference between the two smallest local minima by the sum of the cost curve. The intuition is that we would like the global minimum to be clearly preferable to the second best alternative and also the total cost to be large indicating that not many disparities are acceptable.

$$C_{WMN} = \frac{c_{2m} - c_1}{\sum_d c(d)} \tag{10}$$

As for PKR, we define a naive alternative (**WMNN**) that does not require the second candidate to be a local minimum.

$$C_{WMNN} = \frac{c_2 - c_1}{\sum_d c(d)} \tag{11}$$

5. Consistency between the left and right disparity maps These methods examine whether the disparity map for the right image is consistent with that of the left image. Note that while both disparity maps can be produced by traversing the left cost volume $c(x, y, d)$, we will use $c_{_R}(x_{_R}, y, d_{_R})$ for clarity.

Left Right Consistency (*LRC*) has been widely used as a binary test for the correctness of matches. Egnal et al. [6] defined LRC as the absolute difference between the selected disparity for a pixel in the left image ($d_1 = argmin_d\{c(x, y, d)\}$) and the disparity $D_R(x - d_1, y) = argmin_{d_{_R}}\{c_{_R}(x - d_1, y, d_{_R})\}$ assigned to the corresponding pixel in the right image.

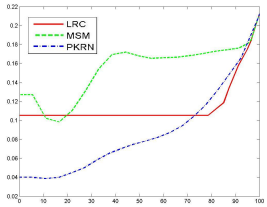
$$C_{LRC}(x, y) = -|d_1 - D_R(x - d_1, y)| \tag{12}$$

We negate the absolute difference so that larger values of C_{LRC} correspond to higher confidence. LRC produces quantized integer values for the confidence and sub-pixel implementations are of dubious value.

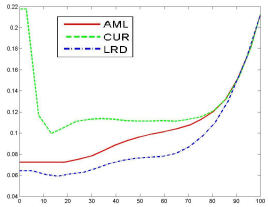
Left Right Difference (*LRD*) is a new metric proposed here that favors large margin between the two smallest minima of the cost and also consistency of the minimum costs across the two images.

$$C_{LRD}(x, y) = \frac{c_2 - c_1}{|c_1 - min\{c_{_R}(x - d_1, y, d_{_R})\}|} \tag{13}$$

The intuition is that truly corresponding windows should result in similar cost values and small values of the denominator. This formulation provides safeguards against two failure modes. If the margin $c_2 - c_1$ is large, but the pixel has been mismatched the denominator will be large and confidence will be low. If the margin is small, the match is likely to be ambiguous. In this case, a small denominator indicates that a correspondence between two similar pixels has been established.

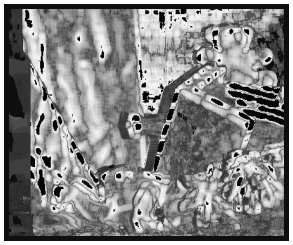


(a) MSM, LRC and $PKRN$



(b)AML, CUR and LRD

(c) PKR confidence map



(d) LRD confidence map

Figure 1. Top: ROCs of error rate over disparity map density for Teddy using SAD in 9×9 windows. Bottom: confidence maps using PKR and LRD. Bright pixels indicate high confidence. (Confidence maps are scaled nonlinearly for visualization.)

4. Experiments on Indoor Images

In this section, we present our evaluation methodology and results on the Middlebury benchmark data [18]. We evaluate the ability of the methods of Section 3.1: to predict the correctness of matches for non-occluded pixels and pixels at discontinuities; to detect occluded pixels; and to select the correct disparities among multiple options for the same pixel. All experiments were performed on cost volumes computed in square windows ranging from 1×1 to 15×15 for SAD and 3×3 to 15×15 for NCC (converted to cost by taking 1-NCC). Confidence values were computed using all methods described in Section 3.1. The value of σ, which is used to model noise, is set to 0.5 for all methods that require it.

Additional results can be found at http://personal.stevens.edu/~xhu2/Papers/CVPR2010/CVPR2010.html.

4.1. Detection of Correct Matches

To assess the capability of a confidence metric to predict whether a disparity is correct, we rank all disparity assignments in decreasing order of confidence and compute the error rate in disparity maps with increasing density. Specifically, for each cost volume and each confidence metric, we select the top 5% of the matches according to confidence and measure the error rate, defined as the percentage of pixels with disparity errors large than one [18], then repeat for the top 10% and so on. Ties are resolved by including all matches with equal confidence. (For example, the first sample using LRC includes all matches with $C_{LRC} = 0$ which could be more than 70% of the total.) This produces ROC curves of error rate as a function of disparity map density [10, 16]. The area under the curve (AUC) measures the ability of a confidence metric to predict correct matches.

Method	SAD	AUC	NCC	AUC
MSM	15	0.097	9	0.162
CUR	9	0.126	11	0.129
PKR	9	0.113	7	0.120
PKRN	11	**0.086**	11	0.097
NLM	9	0.108	11	0.095
PRB			9	0.131
MLM	15	0.098	9	0.106
AML	9	0.102	9	0.105
NEM	11	0.188	9	0.157
WMN	9	0.124	7	0.127
WMNN	11	0.097	11	0.096
LRC	15	0.112	9	0.115
LRD	9	0.089	11	**0.075**

Table 1. Minimum AUC for each confidence metric on Teddy. The second and fourth column show the window size used to obtain the minimum AUC. The error rates at 100% density are approximately 0.21 for SAD and 0.18-0.22 for NCC depending on the window size. All methods perform better than random.

	SAD		NCC		All	
Method	Av	Min	Av	Min	Av	Min
MSM	12	5	8	7	11	7
CUR	10	11	13	11	10	11
PKR	8	8	5	8	6	8
PKRN	3	**1**	**1**	2	2	2
NLM	6	9	7	4	7	5
PRB			12	12		12
MLM	2	2	4	6	3	3
AML	4	7	3	5	4	6
NEM	11	12	11	13	12	13
WMN	7	10	9	10	8	10
WMNN	5	3	6	3	5	4
LRC	9	6	10	9	9	9
LRD	**1**	4	2	**1**	**1**	**1**

Table 2. Confidence metrics ranked according to average and minimum AUC over all window sizes for SAD and NCC separately and best overall performance.

Ideally, all correct matches should be selected before all errors, resulting in the smallest possible AUC for a given disparity map. Random selection of matches produces a flat ROC with an AUC equal to the error rate of the disparity map at full density. *The last point of all ROCs for the same disparity map has a y-coordinate equal to the error rate at full density*. Some of the metrics perform worse than random chance in some of the experiments. Figure 1 shows some examples of ROCs and confidence maps for Teddy. We have computed the AUC for all combinations of window size, cost function (SAD and NCC) and confidence metric. The lowest AUC and the window size with which it was

obtained for each metric for Teddy is shown in Table 1.

These experiments are summarized in Table 2, which contains the rank of each method on the Middlebury dataset according to: the *minimum* AUC achieved for each cost function and the *average* AUC over all window sizes for a given cost function. The AUCs for each cost function and window size, are averaged over the four stereo pairs before the minimum is selected. That is, the minimum AUC reported has been obtained by applying the confidence method on all images with fixed parameters. The purpose of using the average AUC as a criterion is to evaluate the sensitivity of the confidence metrics to the selection of the underlying cost function and window size. Performance should be stable even for suboptimal choices, since this is often the case when one encounters unfamiliar scenes.

4.2. Evaluation at Discontinuities

We have also performed similar experiments on non-occluded pixels near discontinuities using the provided ground truth [18]. In this case, only pixels labeled as discontinuities are taken into account in the computation of the ROCs. There are a few differences with the evaluation on all non-occluded pixels that should be pointed out: NCC results in lower AUC for all methods; several methods perform worse than random chance for some of the cost functions; the improvement over random chance is smaller than for all non-occluded pixels. The best performing methods over all non-occluded pixels cover approximately 44% of the AUC obtained by random chance. The same figure rises to 60% near discontinuities. Tsukuba is the hardest dataset: the overall best result covers 89.5% of the AUC obtained by random chance near discontinuities. Details can be found on the web page listed above.

4.3. Occlusion Detection

One of our requirements for a confidence metric is to assign low confidence to occluded pixels. We evaluated occlusion detection by counting the number of occluded pixels included in each disparity map as more matches are added in order of decreasing confidence. Better performance is indicated by smaller area under this curve. Most methods fail compared to random chance on Tsukuba. Only MSM and CUR succeed for more than 30% of the other input disparity maps. Even then, the AUC is at best 80% of random chance. Performance is also not good on Venus. One can speculate that this could be due to the very small fraction of occluded pixels in the older images, but more analysis is required to confirm this hypothesis.

Our results confirm the conventional wisdom that MSM and LRC/LRD are well suited for this task. They also show that performance here is more unstable than in the previous experiments. Details can be found on the web page listed above.

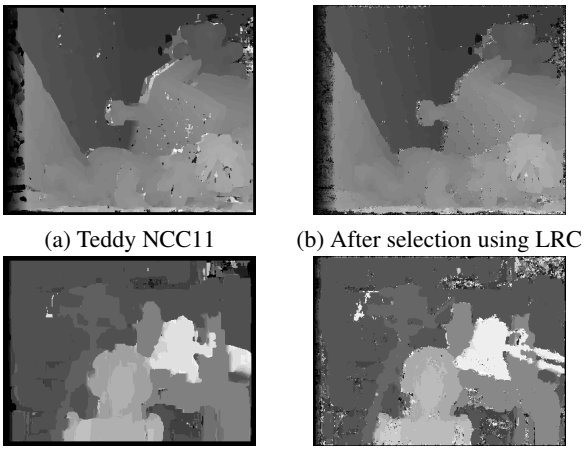
Method	Data	Cost	Optimal	Input	Actual
LRD	Cones	SAD	4.1	15.8	10.2
LRD	Cones	NCC	4.6	12.0	9.6
LRD	Teddy	SAD	8.1	20.9	15.8
LRD	Venus	SAD	3.6	11.2	8.6
AML	Cones	NCC	4.6	12.0	11.1
AML	Teddy	SAD	8.1	20.9	16.7
AML	Venus	SAD	3.6	11.2	8.9
NLM	Cones	SAD	4.1	15.8	11.8
NLM	Teddy	SAD	8.1	20.9	17.1
NLM	Venus	SAD	3.6	11.2	9.2
CUR	Cones	SAD	4.1	15.8	12.4
CUR	Teddy	SAD	8.1	20.9	17.3
CUR	Venus	SAD	3.6	11.2	8.9
WMN	Cones	NCC	4.6	12.0	10.8
WMN	Cones	SAD	4.1	15.8	12.8
WMN	Teddy	SAD	8.1	20.9	17.6
WMN	Teddy	NCC	8.0	17.7	17.4
WMN	Venus	SAD	3.6	11.2	9.0
LRC	Cones	NCC	4.6	12.0	10.6
LRC	Teddy	NCC	8.0	17.7	15.5
PKR	Cones	NCC	4.6	12.0	9.8
PKRN	Cones	NCC	4.6	12.0	9.8
MSM	Cones	NCC	4.6	12.0	10.9
WMNN	Cones	NCC	4.6	12.0	11.7

Table 3. Results of disparity selection using confidence. Only methods that make improvements are listed. The three rightmost columns report the error rates: after optimal selection, of the best input disparity map and the one obtained by selection. See text for details.

4.4. Disparity Selection

The final experiment on the Middlebury data aims at selecting disparities from multiple disparity maps according to confidence. The intuition is that different window sizes are more effective for different types of pixels. If WTA stereo algorithms were able to select the right window for each pixel, they would perform significantly better than they currently do. To test whether the confidence metrics are useful in the selection process, we compute disparity maps using window sizes ranging from 1×1-15×15 using SAD and 3×3-15×15 using NCC.

These computations provide 8 datasets: one each for SAD and NCC, for each stereo pair. Each dataset comprises 8 (7 for NCC) disparity maps. A confidence method is applied on a dataset, e.g. Cones-SAD, to estimate the confidence of all 8 disparity estimates for each pixel. Then, the disparity estimate with the highest confidence value is selected for that pixel without considering any neighborhood information. (More sophisticated strategies are likely to be more effective, but our goal is to evaluate confidence in isolation.) We have also tried the same experiment using ranks instead of raw confidence values with similar results.



(a) Teddy NCC11 (b) After selection using LRC

(c) Tsukuba NCC15 (d) After selection using PKRN

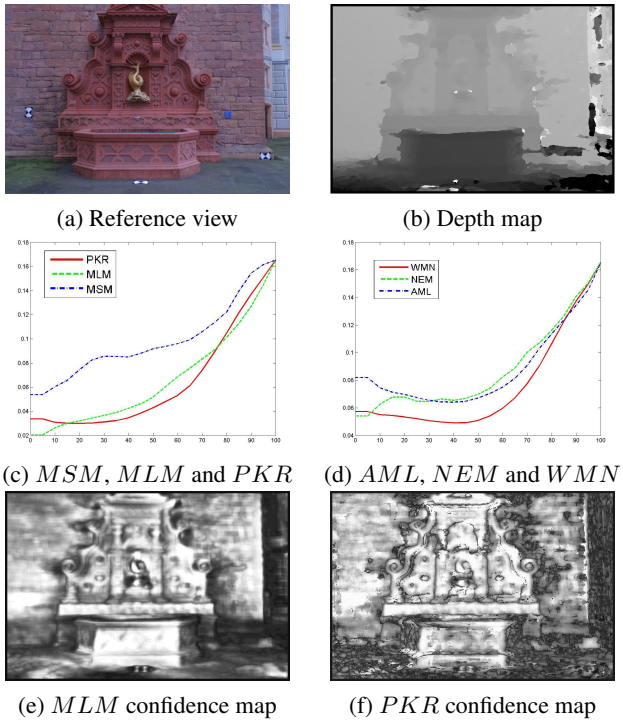
Figure 2. Disparity selection. Left column: the input disparity map with minimum error rate. Right column: results of selection according to confidence. The error rate increases for Tsukuba, but the appearance is significantly improved.

There are two landmarks with which we compare the error rate of the obtained disparity maps: the error rate of the optimal selection and the minimum error rate of the input disparity maps. The former is the error rate obtained if we could somehow make the optimal choice among all the disparity estimates for each pixel – an error occurs only if none of the input disparity maps is correct for that pixel. The minimum error rate of the inputs is an indicator of whether the combination is beneficial, or whether standard stereo using a single window size would have been more effective. No method was able to make an improvement for Tsukuba, while many of the methods fail for all datasets. In Table 3, we report only the methods that resulted in an improvement over the best input disparity map, along with the error rate of the optimal selection, the minimum input error rate and the error rate measured after the selection process. In some cases, selection is able to reduce the error rate by about a third. See Fig. 2 for examples of disparity selection.

5. Experiments on Outdoor Images

In this section, we present results on the *fountain-P11* dataset, courtesy of Strecha et al. [20]. The dataset consists of 11 images and is one of the few publicly available outdoor datasets with ground truth. We compute depth maps for the central image using all ten other images as matching targets. We implemented the plane sweeping algorithm [8] and performed multi-baseline matching using one sweeping direction (fronto-parallel). We downsampled the images to 615×410 and swept a plane in 1000 steps along the optical axis of the reference camera. SAD and NCC are computed using the same window sizes as in Section 4.

The result of the plane sweeping algorithm is a volume that contains the cost value for each of the 1000 depth candidates for each pixel. Therefore, it is a cost volume of the same form as those of the previous section. Depths are assigned to pixels as the candidates with minimum cost. All confidence methods can be computed in a straightforward manner, except for LRC and LRD that require the computation of at least one more depth map and rendering of depth estimated from this depth map to the reference view.



(a) Reference view (b) Depth map

(c) MSM, MLM and PKR (d) AML, NEM and WMN

(e) MLM confidence map (f) PKR confidence map

Figure 3. The *fountain-P11* dataset. (a): one of the input images. (b): the depth map using SAD in 11×11 windows. (c) and (d): ROCs for *fountain-P11* for SAD in 11×11 windows. (e) and (f): confidence maps using MLM and PKR. Bright pixels correspond to higher confidence.

5.1. Detection of Correct Matches

Since we are interested in more than just the error at full density, we do not use the online evaluation tool of [20]. Instead, we generated ground truth depth maps by rendering the provided 3D model for the *fountain-P11* dataset. The variances of the lidar measurements are also not available, so we use the average distance from the ground truth as the error metric, as well as the percentage of bad pixels, defined as the percentage of pixels with error above a certain threshold, set here to 1% of the depth range of the scene. Tables 4 and 5 show the minimum AUC for each method under SAD and NCC according to the two error metrics.

Depth selection, along the lines of Section 4.4, can be implemented in a straightforward way for the multi-view data, but the results are uninformative. This is because the vast majority of the pixels are on smooth surfaces and the overall error rate decreases with increasing window size.

Method	SAD AUC	Rank	NCC AUC	Rank
MSM	0.091	7	0.043	5
CUR	0.117	10	0.070	10
PKR	0.059	**1**	0.035	**1**
PKRN	0.083	6	0.045	6
NLM	0.101	8	0.070	9
PRB			0.062	7
MLM	0.061	2	0.038	3
AML	0.079	4	0.038	4
NEM	0.082	5	0.094	11
WMN	0.070	3	0.035	2
WMNN	0.108	9	0.069	8

Table 4. AUC and rank for all confidence methods applied to SAD and NCC cost volumes for the *fountain-P11* dataset using **average distance from ground truth** as the error metric. The best result has been selected for each method. In general, the best results are obtained using large windows due to the smoothness of the scene.

Method	SAD AUC	Rank	NCC AUC	Rank
MSM	0.032	6	0.010	5
CUR	0.056	10	0.021	9
PKR	0.018	**1**	0.008	**1**
PKRN	0.033	7	0.011	6
NLM	0.041	8	0.022	10
PRB			0.018	7
MLM	0.021	2	0.008	2
AML	0.028	5	0.008	4
NEM	0.025	4	0.033	11
WMN	0.022	3	0.008	3
WMNN	0.046	9	0.021	8

Table 5. AUC and rank for all confidence methods on *fountain-P11* using **fraction of bad pixels** as the error metric.

Therefore, selecting depth estimates computed using small windows usually leads to an increase in error.

6. Conclusions

The most significant conclusions from our experiments are the following:

- Most confidence metrics meet the requirements of Section 1 and they are often able to outperform the baseline method (MSM), except on occlusion detection. According to our results, LRD, $PKRN$ and MLM are better on the Middlebury data, while PKR, MLM and WMN work well for *fountain-P11*.
- Often the minimum AUC is not achieved for the window size with minimum total error. Small variations in window size can trade off between lower error rate or higher predictability of correctness. The differences are small, but the choice depends on the application requirements. For multi-view stereo predictability may be preferable to lower error at full density.
- Methods that consider the entire cost curve assign abnormally high confidence to pixels with very small numbers of valid disparity choices. This is not an issue for the outdoor data, but affects binocular results.
- $PKRN$ and $WMNN$ outperform PRK and WMN on the Middlebury data because, in some sense, they combine the criteria on the local neighborhood and on the presence of competing hypotheses. This relationship is reversed in the multi-baseline setting in which the steps in depth do not correspond to single disparity steps on the images. Uniform sampling in depth at large distances results in small motions of the matching window on the target images and, thus, flat cost curves. Generating depth hypotheses that correspond to equal steps in *all* target images simultaneously is far from straightforward.
- Five of the methods are generally successful in the disparity selection task. A common failure mode for the others is bias for small or large window sizes. The former results in salt and pepper noise and the latter in "foreground fattening".

There are also several informative findings on the performance of individual methods.

The **Matching Score Metric (*MSM*)** performs much better on SAD than NCC. It is less stable than most other methods and fluctuates as the window size varies. MSM does not perform well for small windows or near discontinuities, but it is the *best method for occlusion detection.*

Curvature (*CUR*) tends to rank some errors very highly because it assigns high confidence to pixels near discontinuities due to the accompanying discontinuity in the cost curve. As a result, it also performs poorly near discontinuities. CUR performs worse than expected given its popularity, but it is one of the few methods that are successful in disparity selection. It is a bad choice for the outdoor data, due to the uneven spacing of the depth candidates in terms of disparity on the target images.

The **Peak Ratio (*PKR*)** is the *best method on the outdoor data.* It does not do particularly well on the indoor experiments, in which it is worse than $PKRN$.

The **Naive Peak Ratio (*PKRN*)** is one of the top methods on the Middlebury data, especially near discontinuities. It is not effective in disparity selection, however, due to bias for small windows that leads to salt and pepper noise.

The **Nonlinear Margin (*NLM*)** is reliable, but not outstanding. It is effective, however, for disparity selection.

The **Probabilistic Metric (*PRB*)** shows that some form of nonlinearity is apparently necessary, as it fares worse than the other methods that consider the entire cost curve (MLM and AML).

The **Maximum Likelihood Metric (*MLM*)** is the *best method near discontinuities; second best outdoors and third best for non-occluded pixels on the Middlebury data*. It generates confidence maps with the sharpest boundaries, but it is biased towards small windows during selection.

The **Attainable Maximum Likelihood (*AML*)** performs worse, in general, than MLM on all experiments. Unlike MLM, AML is successful in disparity selection. This is due to the removal of the bias towards smaller windows by subtracting the minimum attained cost during the conversion from cost to pdf.

The **Negative Entropy Metric (*NEM*)**: does not perform well, as noted also in [17]. It consistently ranks last with the exception of the outdoor dataset using SAD.

The **Winner Margin (*WMN*)** is usually worse than PKR, but still top-3 outdoors. It is worse than $WMNN$ indoors, but better outdoors (see PKR and $PKRN$). It is effective for disparity selection.

The **Naive Winner Margin (*WMNN*)** is worse than $PKRN$. It is better indoors and worse outdoors than WMN, but fails at disparity selection due to bias for small windows.

LRC and LRD were only evaluated indoors because outdoor extensions would require the computation of at least one additional depth map, which would favor these methods. (In a binocular configuration, the left and right cost volume can be generated from each other by re-arranging their contents.)

The **Left Right Consistency (*LRC*)** achieves average performance due to quantization. More than 50% of the matches in almost all experiments are consistent, resulting in a very large set of matches that appear to have equal confidence. It is effective in occlusion detection.

The **Left Right Difference (*LRD*)** is the *best overall method indoors.* It also performs very well near discontinuities and in occlusion detection and is the *best method in disparity selection.*

Disparity selection and extension to a true depth map fusion approach are the most interesting directions of future work. It appears that combinations of some of the metrics within a learning approach should lead to significant progress, but the design of appropriate training and testing conditions is far from trivial.

References

[1] M. Bleyer, S. Chambon, U. Poppe, and M. Gelautz. Evaluation of different methods for using colour information in global stereo matching approaches. In *Int. Society for Photogrammetry and Remote Sensing*, pages 63–68, 2008.

[2] D. Bradley, T. Boubekeur, and W. Heidrich. Accurate multi-view reconstruction using robust binocular stereo and surface meshing. In *CVPR*, 2008.

[3] M. Brown, D. Burschka, and G. Hager. Advances in computational stereo. *PAMI*, 25(8):993–1008, 2003.

[4] N. Campbell, G. Vogiatzis, C. Hernandez, and R. Cipolla. Using multiple hypotheses to improve depth-maps for multi-view stereo. In *ECCV*, pages 766–779, 2008.

[5] C. Dima and S. Lacroix. Using multiple disparity hypotheses for improved indoor stereo. In *ICRA*, volume 4, pages 3347–3353, 2002.

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

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

[8] D. Gallup, J.-M. Frahm, P. Mordohai, Q. Yang, and M. Pollefeys. Real-time plane-sweeping stereo with multiple sweeping directions. In *CVPR*, 2007.

[9] M. Gong, R. Yang, L. Wang, and M. Gong. A performance study on different cost aggregation approaches used in real-time stereo matching. *IJCV*, 75(2):283–296, 2007.

[10] M. Gong and Y. Yang. Fast unambiguous stereo matching using reliability-based dynamic programming. *PAMI*, 27(6):998–1003, 2005.

[11] H. Hirschmuller, P. Innocent, and J. Garibaldi. Real-time correlation-based stereo vision with reduced border errors. *IJCV*, 47(1-3):229–246, 2002.

[12] H. Hirschmüller and D. Scharstein. Evaluation of stereo matching costs on images with radiometric differences. *PAMI*, 31(9):1582–1599, 2009.

[13] S. Lefebvre, S. Ambellouis, and F. Cabestaing. A colour correlation-based stereo matching using 1d windows. In *IEEE Conf. on Signal-Image Technologies and Internet-Based System*, pages 702–710, 2007.

[14] L. Matthies. Stereo vision for planetary rovers: Stochastic modeling to near real-time implementation. *SPIE*, 1570:187–200, 1991.

[15] P. Merrell, A. Akbarzadeh, L. Wang, P. Mordohai, J.-M. Frahm, R. Yang, D. Nistér, and M. Pollefeys. Real-time visibility-based fusion of depth maps. In *ICCV*, 2007.

[16] P. Mordohai. The self-aware matching measure for stereo. In *ICCV*, 2009.

[17] D. Scharstein and R. Szeliski. Stereo matching with nonlinear diffusion. *IJCV*, 28(2):155–174, 1998.

[18] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 47(1-3):7–42, 2002.

[19] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *CVPR*, pages 519–528, 2006.

[20] C. Strecha, W. von Hansen, L. Van Gool, P. Fua, and U. Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *CVPR*, 2008.

[21] F. Tombari, S. Mattoccia, L. di Stefano, and E. Addimanda. Classification and evaluation of cost aggregation methods for stereo correspondence. In *CVPR*, 2008.