A Quantitative Evaluation of Confidence Measures for Stereo Vision

Xiaoyan Hu, *Student Member*, *IEEE*, and Philippos Mordohai, *Member*, *IEEE*

Abstract—We present an extensive evaluation of 17 confidence measures for stereo matching that compares the most widely used measures as well as several novel techniques proposed here. We begin by categorizing these methods according to which aspects of stereo cost estimation they take into account and then assess their strengths and weaknesses. The evaluation is conducted using a winner-take-all framework on binocular and multibaseline datasets with ground truth. It measures the capability of each confidence method 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. Our work was motivated by the observation that such an evaluation is missing from the rapidly maturing stereo literature and that our findings would be helpful to researchers in binocular and multiview stereo.

Index Terms—Stereo vision, 3D reconstruction, confidence, correspondence, distinctiveness

1 Introduction

While several confidence measures for stereo matching have been proposed in the literature and benchmarks with ground truth depth have been available for years, the criteria for selecting a confidence measure and the relative merits of different measures have not been investigated thoroughly. We study these issues using binocular and multibaseline stereo imagery with ground truth [1], [2], [3]. Our goal is to categorize the different methods and to shed light on their performance according to the 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 17 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.

- *The authors are with the Department of Computer Science, Stevens Institute of Technology, Castle Point on Hudson, Hoboken, NJ 07030. E-mail:{xhu2, philippos.mordohai}@stevens.edu.*

Manuscript received 14 Dec. 2010; revised 7 Oct. 2011; accepted 23 Jan. 2012; published online 30 Jan. 2012.
Recommended for acceptance by W. Förstner.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2010-12-0955.
Digital Object Identifier no. 10.1109/TPAMI.2012.46.

We have evaluated the degree to which each method satisfies the above criteria using a set of experiments on stereo matching 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 [1] and on multibaseline imagery collected indoors [2] and outdoors [3].

Since stereo matching is known to be prone to errors, the capability of predicting 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 without a cumbersome procedure for estimating marginals for each pixel [4]. Confidence estimation is more practical when dynamic programming is used for optimization; Gong and Yang [5] 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. Here, we restrict the analysis to a WTA stereo framework.

In summary, the contributions of this paper are:

- A classification of several confidence measures.
- A set of criteria for evaluating them.
- Four new confidence measures that often perform better than conventional methods.
- Quantitative and qualitative comparisons of a large number of confidence methods on binocular and multiview imagery.

A preliminary version of this work presenting some of the following results on a subset of the methods and a much smaller dataset appeared in [6].

1.1 Motivation

One of the motivations for our work is the observation that conventional matching functions, such as the Sum of Absolute Differences (SAD) or Normalized Cross Correlation (NCC),[1] do not assign the lowest cost or highest

1. NCC always refers to zero-mean NCC in this paper. See Section 3.

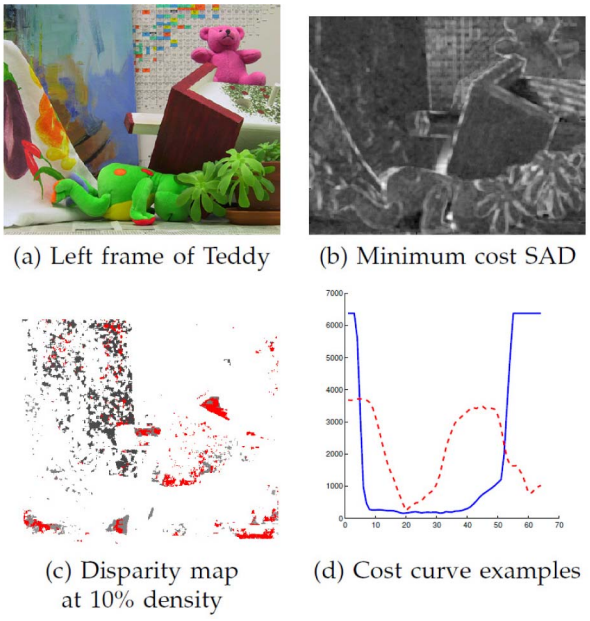


Fig. 1. (a) The left image of the Teddy stereo pair from the Middlebury Stereo Vision Page and (b) the minimum cost map computed using the Sum of Absolute Differences in 5×5 windows. High intensity corresponds to large cost. Note that pixels in uniform areas have lower cost values under SAD since the cost can approach 0, while larger values are observed at textured pixels, even though they may be correct. A sparse disparity map (red pixels indicate wrong disparity) that includes the 10 percent of the matches with the lowest cost has an error rate of 28.8 percent (c) and examples of cost curves that lead to matching errors (d).

similarity to the most unambiguous matches. Fig. 1 shows the left image of the Teddy stereo pair [7] and the minimum cost for each pixel computed using SAD. If we use cost to select matches in order to make the best possible disparity map of 10 percent density, the error rate would be 28.8 percent. On the other hand, we show that if we use the Self-Aware Matching Measure (SAMM) [8] instead, we can obtain a disparity map of the same density containing only 4.62 percent wrong matches. Selecting matches with minimum cost fails due to competing hypotheses (multiple local minima) or flat valleys around the true minimum, as shown in Fig. 1d. Matching cost is still of some value, as evidenced by the success of methods that detect ground control points or seed matches [9], [10], [11], [12], [13] based on their low cost values, but we are not the first to claim that improvements are possible. Several authors [8], [14], [15], [16], [17], [18], [19] have proposed algorithms that examine the cost curve and assign a scalar confidence value to each potential pixel match. These confidence values can be used to rank matches from most to least reliable. In the remainder of this paper, we examine these methods and compare them according to their ability to rank potential matches.

In this study, we focus on methods that operate on individual pixels by examining their cost curves. The "ideal" cost curve as a function of disparity for a pixel is shown in Fig. 2a. It has a single, distinct minimum. The cost curves in Figs. 1d and 2b are more ambiguous because they have multiple local minima or multiple adjacent disparities with similar cost making exact localization of the global minimum hard. Most confidence measures extract local or global features of the cost curve to characterize the reliability of the match corresponding to the minimum cost. In Section 3, we

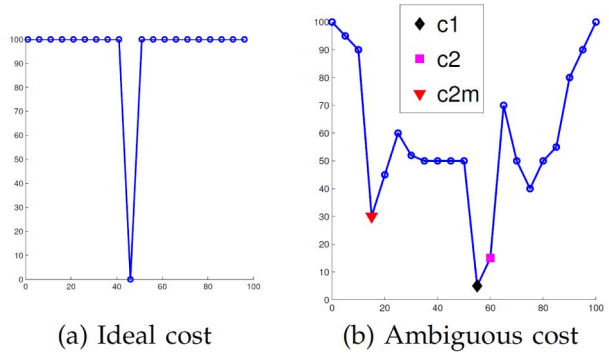


Fig. 2. The ideal cost curve for stereo has one distinct minimum. A less ideal cost curve has several local minima and/or flat regions of low cost. c_1 denotes the minimum value of a cost curve, c_2 is the second smallest value, and c_{2m} is the second smallest local minimum. c_2 and c_{2m} may coincide in some cases.

classify these methods according to the type of features they take into account.

To the best of our knowledge, this is the first survey and experimental comparison of confidence measures for stereo since the work of Egnal et al. [20]. We have included the methods evaluated by Egnal et al. as well as a large number of other methods. We perform experiments on the core depth estimation modules, namely, the binocular and multibaseline configurations, but expect our findings to have potentially larger impact on multiview systems that merge partial reconstructions in order to reconstruct large-scale scenes [21], [22], [23], [24].

2 Related Work

Our work complements surveys on binocular [1], [25] and multiview [26] stereo, as well as on specific aspects of stereo, such as cost functions [27], [28], cost aggregation [29], [30], and color spaces [31], [32]. These efforts, aided by standardized benchmarks [1], [26], have led to significant progress and more principled design of stereo algorithms. For clarity, we present the confidence measures in Section 3. Here, we only discuss related work that is not included in our evaluation.

Arguably, the most significant and most recent comparison of stereo confidence measures was carried out by Egnal et al. [20]. Five measures, four of which are included here, were evaluated on single-view stereo in predicting matching errors on three stereo pairs with ground truth. It is worth noting that the stereo algorithms struggled with two of the pairs, resulting in error rates around 80 percent. In our case, matching is considerably more effective, making a comparison of the findings of [20] and ours hard.

Approaches that combine multiple confidence measures [17], [33], [34] aim at rejecting erroneous matches to obtain error-free quasidense depth maps. Different indicators of matching errors, such as left-right consistency (LRC), 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.

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 [35]. A region-based approach was presented by Jodoin et al. [36], who first segment the images and classify segments as occluded or not according to the density of matched pixels in them. Ideally, occlusion detection should not be performed locally since occlusion is a result of long range interaction between remote surfaces. Depth and occlusion estimation are coupled and one can benefit greatly from correct estimates of the other. This relationship is expressed by the uniqueness [37] and visibility [38] constraints. Global optimization approaches [38], [39], [40] have achieved good results by jointly estimating disparity and occlusion maps. We are more interested in the fundamental question of 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 measures.

3 Confidence Measures

Before describing the confidence measures, let us introduce the notation used throughout. Square $N \times N$ windows are used in all cost computations. The experiments were carried out using the sum of absolute color differences in RGB (SAD), which is a cost measure, and zero-mean normalized cross correlation, which measures similarity:

$$SAD(x,y,d) = \sum_{i \in W} |I_L(x_i,y_i) - I_R(x_i - d, y_i)|,$$

$$NCC(x,y,d) = \frac{\sum_{i \in W}(I_L(x_i,y_i) - \mu_L)(I_R(x_i - d, y_i) - \mu_R)}{\sigma_L \sigma_R},$$

where I_L and I_R are the two images of the stereo pair, μ_L and μ_R are the means, and σ_L and σ_R are the standard deviations of all pixels in the square window in the left and right image, respectively. Means are computed separately per RGB channel, but a single standard deviation 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. For uniformity, NCC is converted to a cost function by replacing it with $1 - \mathrm{NCC}$ so that all values are nonnegative and 0 is the minimum attainable cost value. SAD values are normalized by the number of pixels in the window.

The cost value (SAD or $1 - \mathrm{NCC}$) 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* (see Fig. 2b). 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.

3.1 Categorization of Confidence Measures

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

1. Matching cost. The matching cost is used as a confidence measure.

The **Matching Score Measure (MSM)** is the simplest confidence measure [20] and serves as the *baseline* in our experiments. We use the negative of the cost so that large values correspond to higher confidence:

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

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 [20] and is widely used in the literature. It is defined as

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

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 indicates certainty. A similar idea has also been applied on invariant feature descriptors [41]. **Peak Ratio (PKR):** Among several equivalent formulations [17], [20], we have implemented PKR as

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

We have also implemented a naive version, **PKRN**, which does not require the numerator to be a local minimum (see Fig. 2). PKRN can be viewed as a combination of PKR and CUR that assigns low confidence to matches with flat minima or strong competitors:

$$C_{\mathrm{PKRN}} = \frac{c_2}{c_1}. \tag{4}$$

The margin between c_1 and c_2 is also an indication of confidence. We define the **Maximum Margin (MMN)** as

$$C_{\mathrm{MMN}} = c_2 - c_1. \tag{5}$$

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

The **Probabilistic Measure (PRB)** [16] operates on a *similarity* function by treating the value assigned to each potential disparity as a probability for the disparity. This can easily be achieved by normalizing the values to sum to unity. PRB is only used on NCC in this paper, as we do not attempt to convert cost to likelihood via some linear or nonlinear mapping:

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

The **Maximum Likelihood Measure (MLM)** is inspired by [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:

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

MLM assumes that the matching cost can attain the ideal value of 0. Merrell et al. [18] 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):

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

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

The **Negative Entropy Measure (NEM)** was proposed by Scharstein and Szeliski [15]. Cost values are converted to a pdf, the negative entropy of which is used as a measure of confidence:

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

The **Number of Inflection Points (NOI)** measures the number of minimum valleys in cost curves. In the original implementation [34], the second order derivative was used to localize the minima. Since this approach is susceptible to image noise, in our implementation each cost curve is preprocessed with a low-pass filter before the number of local minima is counted:

$$\begin{aligned} C_{\mathrm{NOI}} &= -|M|, \\ M &= \{d_i : c_s(d_i - 1) > c_s(d_i) \wedge c_s(d_i) < c_s(d_i + 1)\}, \end{aligned} \tag{10}$$

where $|M|$ is cardinality of the set of local minima of the smoothed cost curve c_s.

The **Winner Margin (WMN)** was also proposed in [15]. 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:

$$C_{\mathrm{WMN}} = \frac{c_{2_m} - c_1}{\sum_d c(d)}. \tag{11}$$

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

$$C_{\mathrm{WMNN}} = \frac{c_2 - c_1}{\sum_d c(d)}. \tag{12}$$

5. Consistency between the left and right disparity maps. These methods examine whether the disparity of 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 use $c_R(x_R, y, d_R)$ here for clarity.

Left Right Consistency has been widely used as a binary test for the correctness of matches. Egnal et al. [20] 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:

$$C_{\mathrm{LRC}}(x, y) = -|d_1 - D_R(x - d_1, y)|. \tag{13}$$

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 subpixel implementations are of dubious value.

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

$$C_{\mathrm{LRD}}(x, y) = \frac{c_2 - c_1}{|c_1 - min\{c_R(x - d_1, y, d_R)\}|}. \tag{14}$$

The intuition is that truly corresponding windows should result in similar cost values and thus 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.

6. Distinctiveness (DTS)-based confidence measures. The essence of distinctiveness-based measures is to handle point ambiguity, since even very salient image points (e.g., edges and corners) may be hard to match because of repetitive patterns. These methods incur higher computational cost because matching costs for pixels of the same image also have to be computed.

The notion of **distinctiveness** for stereo matching was introduced by Manduchi and Tomasi in [42]. Distinctive points are less likely to be falsely matched between reference and target images; therefore, point distinctiveness can be used to represent matching confidence. The distinctiveness of an image point is defined as the perceptual distance to the most similar other point in the search window in the reference image. We adopt the search window definition from [19]:

$$d_{min} - d_{max} \leq d_s \leq d_{max} - d_{min},$$

in which d_s is the search window in disparity, d_{min} and d_{max} represent minimum and maximum disparity values, respectively.

Then, the distinctiveness of a pixel is

$$C_{\mathrm{DTS}}(x, y) = \min_{d \in d_s, d \neq 0} c_{LL}(x, y, d), \tag{15}$$

in which c_{LL} is the cost volume for matching left image pixels within the same scan line in the same image. It should be noted that DTS is a single-image property since the target image does not enter the computation.

The **Distinctive Similarity Measure (DSM)** [19] utilizes the definition of distinctiveness maps of DTS, but makes use of information from both the left and right image and also considers the similarity between two potentially corresponding points. DSM is defined as follows:

$$C_{\mathrm{DSM}}(x, y) = \frac{C_{\mathrm{L\_DTS}}(x, y) \times C_{\mathrm{R\_DTS}}(x - d_1, y)}{c_1^2}. \tag{16}$$

$C_{\mathrm{L\_DTS}}$ and $C_{\mathrm{R\_DTS}}$ are the distinctiveness maps of the left and right image, respectively. This definition is different from the original paper where the denominator was just c_1

[19]. The squared denominator renders C_{DSM} a dimensionless quantity, which is more suitable for a confidence measure, and our experiments show that this modification results in better performance.

The observation that the behavior of the cost curve around the true disparity is similar to the behavior of the *self-matching cost curve*[2] around zero disparity motivated the **Self-Aware Matching Measure** [8]. In this approach, the correlation coefficient of the two cost curves is used as the similarity measure between them. Unlike DTS and DSM, point distinctiveness is not required to match pixels reliably. The definition of SAMM is

$$C_{\mathrm{SAMM}}(x,y) = \frac{\sum_d (c(x,y,d-d_1)-\mu_{LR})(c_{LL}(x,y,d)-\mu_{LL})}{\sigma_{LR}\sigma_{LL}}, \tag{17}$$

in which μ_{LR} and σ_{LR} are mean and standard deviation of the cross-matching function over the valid disparity range and μ_{LL} and σ_{LL} are defined likewise for the self-matching cost curve. Note that this is the nonsymmetric version of SAMM, as defined in [8]. In the original paper, self-matching takes place over a disparity range which is twice as large as the disparity range used for cross-matching centered at $d = 0$. However, we found in our experiments that this setting does not give the best predictions, so a smaller value which generates better results is used in the following experiments.

In summary, we have presented 17 methods, divided into six categories, that will be evaluated in the following sections. We consider PKRN, WMNN, MMN, and LRD as novel contributions of this paper. Moreover, AML and SAMM have been proposed in our previous work [8], [18], with additional coauthors. PRB is only applicable on cost curves computed using NCC, while all other methods can be applied to either SAD or NCC cost volumes.

4 Experiments on Binocular Data

In this section, we present our evaluation methodology and results on the extended Middlebury benchmark data [1], [2] that includes *31 stereo pairs* published between 2002 and 2007. We evaluate the ability of the methods of Section 3.1: to predict the correctness of matches for nonoccluded 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 $1 - \mathrm{NCC}$). Confidence values were computed using all methods described in Section 3.1. To compare all methods fairly, we tested them on a subset of the Middlebury dataset and selected the parameters that gave the best result for each confidence measure. The parameter values for our experiments are as follows: $\sigma_{MLM} = 0.3$, $\sigma_{AML} = 0.2$ for NCC, $\sigma_{AML} = 0.1$ for SAD, the self-matching disparity range for SAMM is 28, and the width of the low-pass filter for NOI is 5.

4.1 Detection of Correct Matches

To assess the capability of a confidence measure 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 measure, we select the top 5 percent of the matches according to confidence and record the error rate, defined as the percentage of pixels with disparity errors larger than one [1], then repeat for the top 10 percent and so on. Ties are resolved by including all matches with equal confidence. (For example, the first sample using LRC includes all matches with $C_{\mathrm{LRC}} = 0$, which could be more than 70 percent of all pixels.) This produces receiver operating characteristic (ROC) curves of error rate as a function of disparity map density [5], [8], which can also be thought of as cumulative error distributions. (A similar criterion has also been used for evaluating confidence of optical flow [43].) The area under the curve (AUC) measures the ability of a confidence measure to predict correct matches. We opted for the simple ROC criterion of Gong and Yang [5], instead of a similar criterion proposed by Kostliva et al. [44]. Our concern about the latter is that errors can be forgiven if they are caused by other errors, which makes correct disparity assignment impossible when one considers the uniqueness constraint. Since we are mostly dealing with noisy disparity maps, using [44] could neglect certain types of errors.

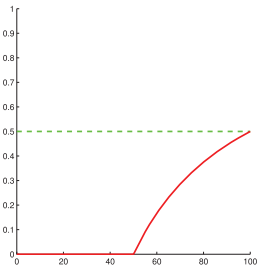


Fig. 3. Optimal AUC and AUC for random chance. The solid red curve is optimal (all correct matches selected first) while the dashed green curve represents random chance.

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, after averaging a large number of trials: $AUC_{rand} = \varepsilon$. *The last point of all ROCs for the same disparity map has a y-coordinate equal to the error rate at full density*. Fig. 3 shows an illustration of the optimal AUC and the AUC obtained by random chance. If d_m denotes the density of a quasidense disparity map that includes the matches with the highest confidence and ε the error rate at full density, the analytic form of the optimal ROC curve is

$$R_{opt}(d_m) = \begin{cases} 0 & d_m \le 1-\varepsilon \\ \dfrac{d_m - (1-\varepsilon)}{d_m} & d_m > 1-\varepsilon. \end{cases} \tag{18}$$

The disparity map can reach density $1 - \varepsilon$ before any wrong matches are included. Then, the fraction of wrong matches grows until it reaches ε. The area under this curve as a function of ε is

$$A_{opt} = \int_{1-\varepsilon}^{1} \frac{d_m - (1-\varepsilon)}{d_m} dd_m = \varepsilon + (1-\varepsilon)ln(1-\varepsilon). \tag{19}$$

Fig. 4 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),

2. The self-matching curve results from matching the reference image with a duplicate of itself.

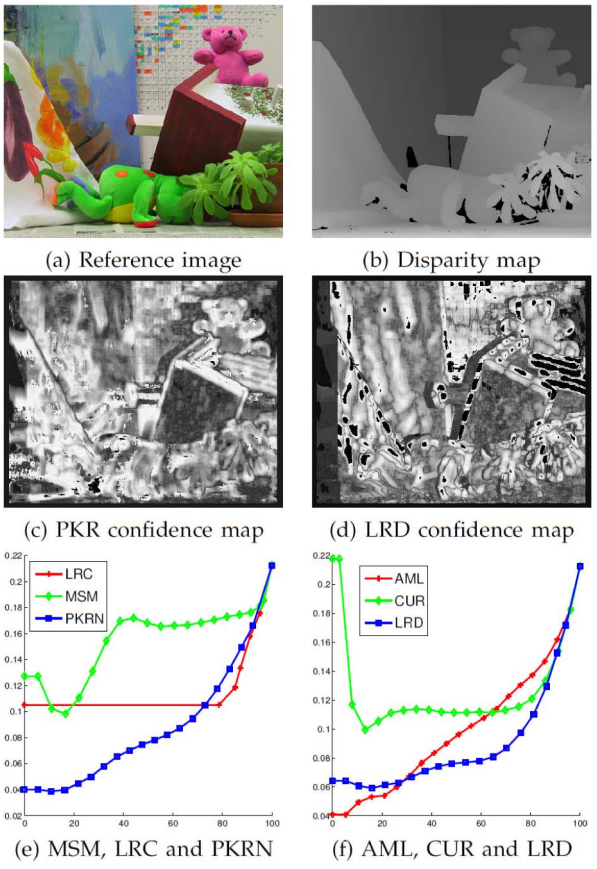


Fig. 4. Top: Reference image and ground truth disparity map for the Teddy dataset. Middle: Confidence maps using PKR and LRD. Bright pixels indicate high confidence. (Confidence maps are scaled nonlinearly for visualization.) Bottom: ROCs of error rate over disparity map density for Teddy using SAD in 9×9 windows. The minimum AUC among these curves is obtained by LRD. Note that values on the y-axis do not start from 0 and that all curves terminate at $(1.0, \varepsilon)$.

and confidence measure. The lowest AUC and the window size with which it was obtained for each measure for Teddy is shown in Table 1. The lowest AUC achieved by any method on Teddy is 0.075 by LRD and it is significantly lower than that of MSM (0.162) and random chance (0.177), but far from ideal (0.0167).

These experiments are summarized in Table 2, which contains the rank of each method on the 31 stereo pairs of the Middlebury dataset according to: the *average* AUC achieved for each cost function and the *minimum* AUC over all window sizes for a given cost function. The AUCs for a given cost function and window size are averaged over all 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.

Table 3 shows the average ratio of optimal AUC (obtained with perfect knowledge of which matches are correct) to the AUC obtained by each measure and Table 4 shows the improvement made by each method over random chance ($(AUC_{rand} - AUC)/AUC_{rand}$). We consider the former a measure of the performance of each method, with 1 being the maximum possible score. Most methods easily outperform the baseline (MSM), with LRD being the best, followed by DSM and PKRN. Table 4 reveals that several methods are not better than random chance.

TABLE 1
Minimum AUC for Each Confidence Measure on Teddy

Method	SAD	AUC	NCC	AUC
MSM	15×15	0.097	9×9	0.162
CUR	9×9	0.126	11×11	0.129
PKR	9×9	0.113	7×7	0.120
PKRN	11×11	**0.086**	11×11	0.097
MMN	9×9	0.108	11×11	0.095
PRB			9×9	0.131
MLM	15×15	0.096	9×9	0.097
AML	15×15	0.095	9×9	0.096
NEM	11×11	0.188	9×9	0.157
NOI	15×15	0.162	9×9	0.190
WMN	9×9	0.124	7×7	0.127
WMNN	11×11	0.097	11×11	0.096
LRC	15×15	0.112	9×9	0.115
LRD	9×9	0.089	11×11	**0.075**
DTS	5×5	0.204	15×15	0.133
DSM	9×9	0.099	11×11	0.085
SAMM	7×7	0.090	7×7	0.099
Random	11×11	0.209	11×11	0.177
Optimal	11×11	0.024	11×11	0.017

The second and fourth columns show the window size used to obtain the minimum AUC. The last two rows show the performance of random *selection, which is expected to be equal to the error rate at 100 percent density and the optimal AUC value obtained if selection was perfect. All methods except NOI using NCC perform better than random, but are far from being optimal.*

4.2 Evaluation at Discontinuities

We have also performed similar experiments on nonoccluded pixels near discontinuities using the provided ground truth for the four original stereo pairs [1], see Fig. 5 for examples. In this case, only pixels labeled as discontinuities are taken into account in the computation of the ROCs. These experiments are restricted to the four original stereo pairs (Tsukuba, Venus, Cones, and Teddy) which have ground truth discontinuity maps. We decided not to generate our own ground truth discontinuity maps due to

TABLE 2
Confidence Measures Ranked According to Average and Best (Minimum) AUC over All Window Sizes for SAD and NCC Separately and Best Overall Performance

	SAD		NCC		All
Method	Ave	Best	Ave	Best	Ave
MSM	14	11	12	12	13
CUR	10	10	13	13	11
PKR	13	13	5	7	10
PKRN	4	**1**	4	3	3
MMN	2	6	10	9	8
PRB			14	14	
MLM	6	5	7	6	6
AML	5	4	6	5	5
NEM	15	15	15	15	15
NOI	16	16	16	17	16
WMN	12	14	11	10	12
WMNN	7	3	8	8	7
LRC	9	9	9	11	9
LRD	**1**	2	2	2	1
DTS	11	12	17	16	14
DSM	8	7	**1**	**1**	2
SAMM	3	8	3	4	4

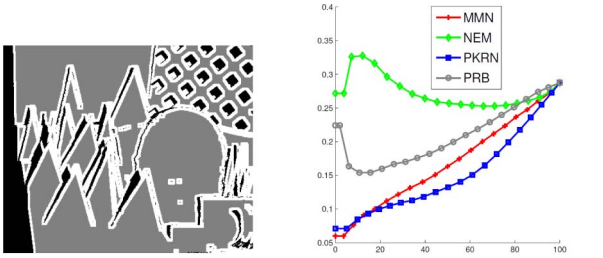
Best AUC corresponds to the minimum AUC that was obtained by each method run with fixed window size, averaged over all 31 stereo pairs. NCC always outperforms SAD according to this criterion, and would always be selected as the overall best. Hence, we omit the "best" overall column since it is identical to the one for NCC.

TABLE 3
Quantitative Results on Performance, Defined as the Average Ratio of the Optimal AUC over the AUC Obtained by Each Method over All Image-Cost Function-Window Size Combinations

Method	Rank	SAD Performance	Rank	NCC Performance	All
MSM	12	0.186	11	0.161	13
CUR	10	0.236	13	0.120	11
PKR	14	0.184	5	0.203	10
PKRN	2	0.328	4	0.219	3
MMN	7	0.303	10	0.162	8
PRB			14	0.111	
MLM	4	0.312	7	0.189	6
AML	3	0.322	6	0.191	5
NEM	15	0.160	15	0.100	15
NOI	16	0.154	16	0.093	16
WMN	13	0.184	9	0.168	12
WMNN	6	0.303	8	0.169	7
LRC	9	0.245	12	0.160	9
LRD	**1**	0.331	2	0.230	**1**
DTS	11	0.196	17	0.082	14
DSM	8	0.286	**1**	0.267	2
SAMM	5	0.308	3	0.223	4

Larger values indicate better performance, with 1 being the best possible.

the ambiguity in the definition of what a nonoccluded discontinuity is. There are a few differences with the evaluation on all nonoccluded 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 nonoccluded pixels. The methods that perform worse than random chance and the fraction of experiments in which this happens are: MSM 22 percent, CUR 10 percent, PKR 18 percent, MMN 10 percent, PRB 13 percent, NEM 60 percent, WMN 8 percent, NOI 28 percent, and DTS 48 percent. The best performing methods over all nonoccluded pixels cover approximately 25 percent of the AUC obtained by random chance. The same figure rises to 51 percent near discontinuities. Tsukuba is the hardest



(a) Discontinuity map (b) PKRN, MMN, PRB and NEM

Fig. 5. Evaluation of **Cones near discontinuities**. (a) White marks the pixels under consideration here. Occluded pixels are colored black and regular pixels are colored gray. (b) ROCs for Cones using NCC in 5×5 windows.

dataset: The overall best result covers 86.9 percent of the AUC obtained by random chance near discontinuities. The rank of each method can be seen in Table 5.

4.3 Occlusion Detection

One of our requirements for a confidence measure 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. See Fig. 6 for examples of ROCs for occlusion detection. Quantitative results can be seen in Table 6.

Our results confirm the conventional wisdom that MSM and LRC/LRD are well suited for this task, but DSM and SAMM are also very competitive. They also show that performance on occlusion detection is more unstable than in the previous experiments.

4.4 Disparity Selection

The final experiment on binocular data aims at selecting disparities from multiple disparity maps according to confidence. The intuition is that different window sizes are more effective for different pixels. If WTA stereo

TABLE 4
Quantitative Results of Overall Improvement

Method	Rank	SAD Improvement	Rank	NCC Improvement	All
MSM	11	-0.010	12	0.319	11
CUR	10	0.263	13	0.171	10
PKR	14	-0.268	7	0.481	12
PKRN	2	0.464	3	0.566	3
MMN	6	0.430	9	0.428	8
PRB			14	0.084	
MLM	4	0.437	6	0.487	6
AML	3	0.450	5	0.492	5
NEM	12	-0.195	16	-0.270	15
NOI	13	-0.235	17	-0.404	16
WMN	15	-0.282	11	0.395	13
WMNN	5	0.431	8	0.454	7
LRC	9	0.306	10	0.399	9
LRD	1	0.470	2	0.594	1
DTS	16	-0.286	15	-0.170	14
DSM	8	0.400	1	0.630	2
SAMM	7	0.406	4	0.538	4

Improvement is defined as $(AUC_{rand} - AUC)/(AUC_{rand})$. Larger values indicate larger improvement, while negative values mean that the measure performs worse than random chance.

TABLE 5
Similar Rankings as in Table 2 Considering Only Matches Near **Discontinuities**

Method	SAD Av	SAD Best	NCC Av	NCC Best	All Av	All Best
MSM	15	14	10	12	12	14
CUR	12	11	14	13	13	11
PKR	11	10	7	8	9	10
PKRN	3	3	3	4	2	3
MMN	9	7	11	10	11	7
PRB			13	14		16
MLM	**1**	2	5	5	3	2
AML	4	5	6	6	6	5
NEM	16	16	16	15	16	17
NOI	14	15	15	16	14	15
WMN	6	8	12	11	8	8
WMNN	8	6	9	9	7	6
LRC	10	9	8	7	10	9
LRD	2	4	4	3	**1**	4
DTS	13	13	17	17	15	13
DSM	7	**1**	**1**	**1**	4	**1**
SAMM	5	12	2	2	5	12

Only the four original stereo pairs (Tsukuba, Venus, Cones, and Teddy) have ground truth discontinuity maps.

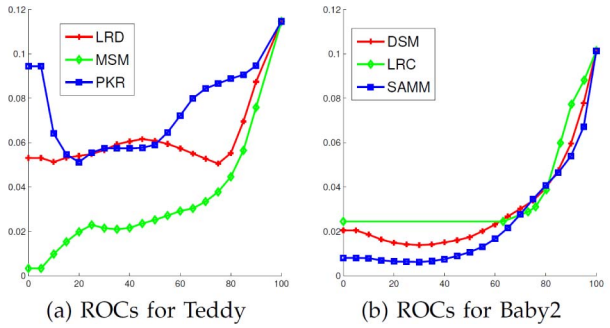


Fig. 6. Evaluation of **occlusion detection**. (a) ROCs for LRD, MSM, and PKR using SAD in 15×15 windows on Teddy. (b) ROCs for DSM, LRC and SAMM using NCC in 5×5 windows on Baby2.

algorithms were able to select the right window size for each pixel, they would perform significantly better than they currently do. To test whether the confidence measures are useful in the selection process, we compute disparity maps using window sizes ranging from 1×1 to 15×15 using SAD and 3×3 to 15×15 using NCC and also compute confidence maps for each disparity map using all methods. LRC has been excluded from this experiment since it results in ties that cannot be broken using this simple selection mechanism.

These computations provide 62 datasets: one each for SAD and NCC, for each stereo pair. Each dataset comprises eight (seven for NCC) disparity maps. A confidence method is applied on a dataset, e.g., Cones-SAD, to estimate the confidence of all eight 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.

There are two benchmarks with which we compare the error rate of the obtained disparity maps: the error rate of the optimal selection and the minimum error rate among 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 Lampshade1, Lampshade2, Midd1, Midd2, Monopoly, Plastic, Tsukuba, and Venus, while many of the methods fail for all datasets. In Table 7, we report results for DSM, LRD, MMN, MSM, PKRN, and SAMM, 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 three quarters. See Fig. 7 for examples of disparity selection. Table 8 reports the success rate of all methods, where success is defined as surpassing the quality of the best input disparity map. Note that the success rate is much higher for NCC than SAD and it often exceeds 60 percent for the former. We attribute this to the fact that NCC values can be transferred more easily across window sizes.

TABLE 6
Rank and Performance for Each Confidence Measure for **Occlusion Detection** on All 31 Stereo Pairs

Method	Rank	SAD Overall Performance	Rank	NCC Overall Performance	All Rank
MSM	**1**	0.264	4	0.225	**1**
CUR	13	0.053	13	0.069	13
PKR	11	0.067	9	0.107	11
PKRN	5	0.159	6	0.207	6
MMN	10	0.087	8	0.117	9
PRB			17	0.039	
MLM	9	0.110	12	0.082	10
AML	8	0.127	11	0.090	8
NEM	16	0.047	15	0.044	16
NOI	15	0.050	16	0.041	15
WMN	12	0.060	10	0.094	12
WMNN	7	0.134	7	0.148	7
LRC	2	0.188	5	0.217	5
LRD	4	0.176	2	0.257	3
DTS	14	0.052	14	0.067	14
DSM	6	0.137	**1**	0.279	4
SAMM	3	0.180	3	0.254	2

Performance is defined in Table 3.

TABLE 7
Some Results on Disparity Selection Using Confidence

Cost	Image	Confidence	Optimal	Input	Output
SAD	Baby3	DSM	0.059	0.203	0.206
		LRD	0.059	0.203	0.109
		MMN	0.059	0.203	0.199
		MSM	0.059	0.203	0.503
		PKRN	0.059	0.203	0.356
		SAMM	0.059	0.203	0.135
SAD	Dolls	DSM	0.023	0.113	0.110
		LRD	0.023	0.113	0.063
		MMN	0.023	0.113	0.100
		MSM	0.023	0.113	0.305
		PKRN	0.023	0.113	0.172
		SAMM	0.023	0.113	0.072
SAD	Rocks2	DSM	0.031	0.111	0.128
		LRD	0.031	0.111	0.057
		MMN	0.031	0.111	0.089
		MSM	0.031	0.111	0.524
		PKRN	0.031	0.111	0.244
		SAMM	0.031	0.111	0.063
NCC	Baby3	DSM	0.020	0.137	0.047
		LRD	0.020	0.137	0.048
		MMN	0.020	0.137	0.078
		MSM	0.020	0.137	0.069
		PKRN	0.020	0.137	0.057
		SAMM	0.020	0.137	0.056
NCC	Dolls	DSM	0.025	0.122	0.056
		LRD	0.025	0.122	0.055
		MMN	0.025	0.122	0.077
		MSM	0.025	0.122	0.063
		PKRN	0.025	0.122	0.061
		SAMM	0.025	0.122	0.075
NCC	Rocks2	DSM	0.014	0.107	0.024
		LRD	0.014	0.107	0.025
		MMN	0.014	0.107	0.034
		MSM	0.014	0.107	0.036
		PKRN	0.014	0.107	0.029
		SAMM	0.014	0.107	0.039

Only results for DSM, LRD, MMN, MSM, PKRN, and SAMM are listed. The last three columns report the error rates: after optimal selection, of the best input disparity map and the one obtained by selection. See text for details.

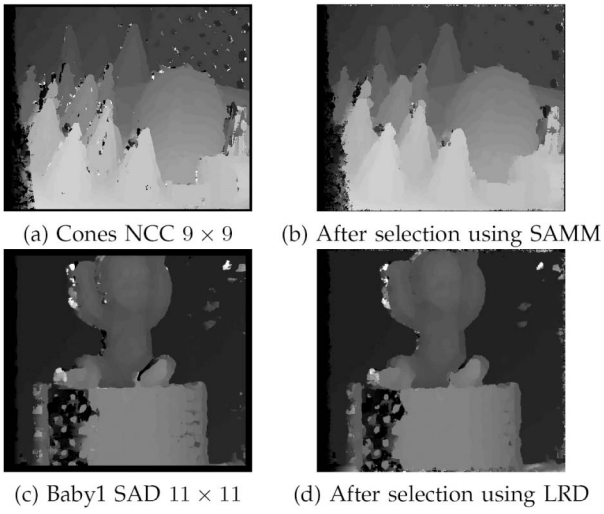


Fig. 7. Disparity selection. Left column: The input disparity map with minimum error rate. Right column: Results of selection according to confidence.

5 Experiments on Multiview Data

Here, we present results on the multibaseline version of the binocular Middlebury benchmark [2] as well as the *fountain-P11* dataset, courtesy of Strecha et al. [3].

It should be noted here that not all confidence methods can be computed in a straightforward manner in a multiview setting. LRC and LRD require the computation of at least one more depth map and rendering of depths estimated from this depth map to the reference view. We decided to exclude them from the comparison due to the unfair advantage they derive from multiple depth maps. (The left and right cost volumes in the canonical binocular configuration contain identical sets of values, providing no additional information.) Distinctiveness-based methods (DTM, DSM, and SAMM) compute self-matching costs along the epipolar lines, requiring the selection of a target view to define an epipolar geometry and were also excluded.

We implemented the plane sweeping algorithm according to [45] and performed multibaseline matching using one sweeping direction (fronto-parallel). SAD and NCC are computed using the same window sizes as in Section 4, while the parameter settings for all confidence measures are also the same. Plane sweeping generates a cost value for each depth candidate and the depth value associated with lowest cost is selected as the final depth. In other words, the cost volume and WTA selection are the same as in the binocular case.

TABLE 8
Disparity Selection Success Rate

Method	Rank	SAD Success Rate	Rank	NCC Success Rate
MSM	12	0.000	9	0.484
CUR	12	0.000	14	0.226
PKR	10	0.065	2	0.677
PKRN	11	0.032	6	0.581
MMN	5	0.290	11	0.419
PRB			12	0.355
MLM	12	0.000	16	0.129
AML	4	0.355	7	0.548
NEM	12	0.000	15	0.194
NOI	3	0.387	8	0.516
WMN	6	0.258	3	0.645
WMNN	7	0.226	9	0.484
LRD	**1**	0.581	**1**	0.710
DTS	8	0.194	13	0.323
DSM	9	0.161	5	0.613
SAMM	2	0.452	3	0.645

TABLE 9
Performance on the Multibaseline Middlebury Data Set

Method	Rank	SAD Performance	Rank	NCC Performance	Ave Rank
MSM	9	0.208	9	0.127	9
CUR	5	0.265	5	0.139	4
PKR	7	0.240	4	0.146	7
PKRN	3	0.306	**1**	0.179	3
MMN	4	0.268	7	0.133	5
PRB			6	0.135	
MLM	2	0.368	3	0.177	2
AML	**1**	0.373	2	0.178	**1**
NEM	10	0.196	11	0.103	10
NOI	11	0.173	12	0.088	11
WMN	8	0.218	10	0.120	8
WMNN	6	0.263	8	0.131	6

5.1 Detection of Correct Matches on Controlled Data

The multibaseline Middlebury dataset comprises seven images for each of the 27 scenes released in 2005 and 2006. Since ground truth disparity maps are only provided for views 1 and 5, the ground truth for the central view (view 3) is calculated by computing the 3D coordinates of every pixel on both views using their disparity maps and then projecting those points onto the central view. These projections are possible using the information provided by the authors of the data: The views are equally spaced, the baseline is 160 mm, the focal length of the camera is 3,740 pixels, all images planes are coplanar, and their axes are aligned. A per pixel depth test is performed if more than one 3D point is projected to the same pixel of the central view and the smaller depth value is kept. If a pixel on the central view is not covered by either views 1 or 5, its depth is considered missing and it is excluded from the evaluation. All experiments here were performed using images at one third of the full resolution. To calculate the cost volume for the central view, we place the near and far plane at depth 2.3467 and 14.96, respectively, and generate 1,000 depth candidates evenly spaced between the near and far plane.

We evaluate all methods on the detection of correct matches, but neither on performance near discontinuities, due to the unavailability of ground truth, nor on occlusion detection, since all pixels of the central view are visible in at least one other view. The results are summarized in Table 9 using performance, defined as in the binocular case, as the criterion. See also Fig. 8 for ROCs and confidence map examples. AML, MLM, and PKRN are the best performing methods in this experiment, while most methods again surpass MSM.

5.2 Detection of Correct Matches on Outdoor Data

The *fountain-P11* dataset consists of 11 images and is one of the few publicly available outdoor datasets with ground truth. Strecha et al. [3] provide an online evaluation tool

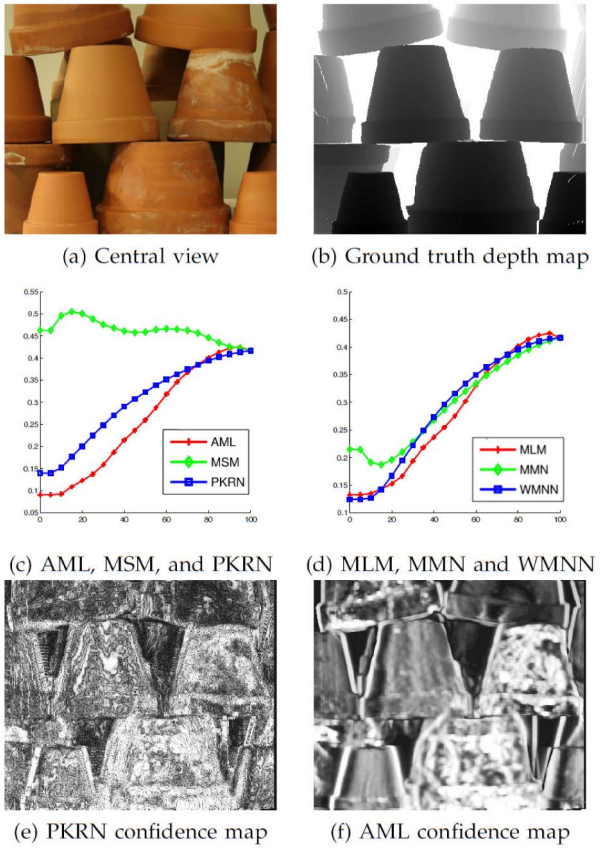


Fig. 8. The *Middlebury* dataset. (a) Central view of the Flowerpots image set. (b) Rendered depth map using ground truth disparity maps of views 1 and 5. (c) and (d) ROCs for Flowerpots for SAD in 7×7 windows. (e)-(f) Confidence maps using PKRN and AML. Bright pixels correspond to higher confidence.

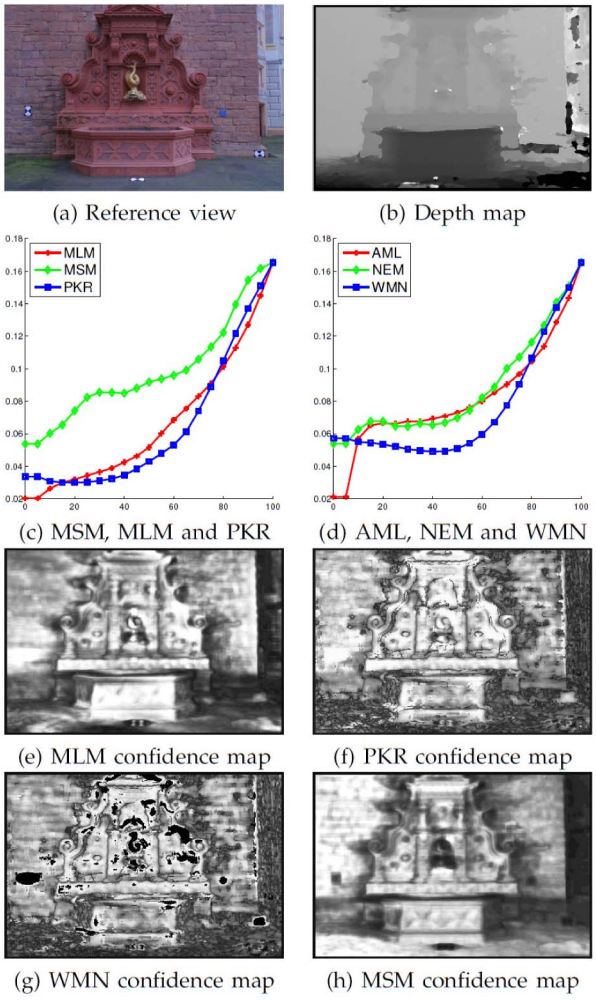


Fig. 9. The *fountain-P11* dataset. (a) One of the input images. (b) The depth map using SAD in 11×11 windows. (c) and (d) ROCs for *fountain-P11* for SAD in 11×11 windows. (e)-(h) Confidence maps.

which does not serve our need to evaluate depth maps at less than full density. Therefore, we generated ground truth depth maps by rendering the provided 3D model of *fountain-P11*. In the following experiments, we estimate depth maps for the central image using all 10 other images as matching targets. All images were downsampled to 615×410 and a plane was swept in 1,000 steps along the optical axis of the reference camera.

We used two error metrics in this experiment: the average distance from the ground truth, as well as the percentage of bad pixels, defined as those with error above a certain threshold, set here to 1 percent of the depth range of the scene. Occlusion and discontinuity maps are not available for this dataset, making evaluations similar to Sections 4.2 and 4.3 impossible. Fig. 9 shows the reference view, a depth map, confidence maps, and ROCs for *fountain-P11*. Tables 10 and 11 show the performance of each method under SAD and NCC according to the two error metrics. AML, MLM, PKR, and WMN are the best performing methods.

6 Conclusions

The most significant conclusions from our experiments are the following:

- Most confidence measures meet the requirements of Section 1 and they typically outperform the baseline method (MSM), except on occlusion detection. According to our results, LRD, DSM, PKRN, SAMM, AML, MLM, WMNN, and MMN perform better than MSM on binocular data, while MLM, AML, PKR, PKRN, and WMN work well for multibaseline inputs.

TABLE 10
Overall Performance of SAD and NCC Cost Volumes for the *fountain-P11* Data Set Using **Average Distance from Ground Truth** as the Error Metric

Method	Rank	SAD Performance	Rank	NCC Performance
MSM	7	0.191	5	0.232
CUR	10	0.151	10	0.113
PKR	**1**	0.285	4	0.288
PKRN	6	0.204	6	0.198
MMN	8	0.166	9	0.123
PRB			7	0.167
MLM	2	0.251	**1**	0.294
AML	4	0.232	3	0.293
NEM	5	0.219	11	0.109
NOI	11	0.099	12	0.074
WMN	3	0.246	2	0.293
WMNN	9	0.154	8	0.125

TABLE 11
Overall Performance of SAD and NCC Cost Volumes for the *fountain-P11* Data Set Using **Fraction of Bad Pixels** as the Error Metric

Method	Rank	SAD Performance	Rank	NCC Performance
MSM	6	0.079	5	0.064
CUR	10	0.051	10	0.024
PKR	**1**	0.129	4	0.087
PKRN	7	0.077	6	0.050
MMN	8	0.061	9	0.027
PRB			7	0.042
MLM	3	0.100	**1**	0.097
AML	5	0.088	2	0.095
NEM	4	0.096	11	0.024
NOI	11	0.031	12	0.015
WMN	2	0.106	3	0.088
WMNN	9	0.054	8	0.028
Random		0.043		0.022

The expected performance by random choice is also shown.

- Methods that consider the entire cost curve (PRB, MLM, AML, NEM, NOI) assign abnormally high confidence to pixels with very small numbers of valid disparity choices. This is usually not an issue for multibaseline data, but affects binocular results.
- PKRN and WMNN, which do no require the second cost in the ratio to be a local minimum, outperform PRK and WMN on the binocular data because, in some sense, they combine the criteria on the flat local neighborhood around the minimum cost and on the presence of competing hypotheses. This relationship is reversed in the multibaseline 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 infeasible in general.
- Eight of the methods (LRD, PKR, WMN, SAMM, DSM, PKRN, AML, and NOI) are successful in more than 50 percent of the disparity selection experiments using NCC as the cost, while two more (MSM and WMNN) succeed more than 48 percent of the time. This does not hold when SAD is used and only LRD succeeds at a 58 percent rate, while SAMM is second at 45 percent. We believe that this is due to NCC being a normalized metric that can be transferred among cost curves computed using different window sizes. On the other hand, SAD, even if it is normalized by the total number of pixels in the window, is bound to produce smaller costs for small window sizes. A common failure mode, regardless of underlying cost, is bias for small or large window sizes. The former results in salt and pepper noise and the latter in "foreground fattening." The nature of each stereo image also affects the performance of disparity selection, as no confidence measure works for Lampshade1, Lampshade2, Midd1, Midd2, Monopoly, Plastic, Tsukuba, and Venus. On the other hand, all measures produce good disparity maps for Cloth1, Cloth2, Cloth3, and Cloth4 using NCC.
- Our results confirm that MSM and LRC/LRD are well suited for occlusion detection. DSM and SAMM also perform well.
- In all experiments, NCC shows better performance than SAD as a matching function. Of course, this comes at increased computational cost.
- 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 multiview, stereo predictability may be preferable to lower error at full density.

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

The **Matching Score Measure** 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 tends to rank some errors very highly because it assigns high confidence to pixels near discontinuities due to the accompanying large discontinuity in the cost curve. As a result, it performs especially poorly near discontinuities. CUR performs worse than expected given its popularity and it is a poor choice for the multibaseline data due to the uneven spacing of the depth candidates on the target images.

The **Peak Ratio** is *one of the top methods on the multibaseline data*, but performs poorly on the binocular experiments, in which it is much worse than PKRN, especially using SAD as the cost function.

The **Naive Peak Ratio (PKRN)** is one of the top methods on the binocular data, especially near discontinuities. It is not very effective in disparity selection, however, due to bias for small windows that leads to salt and pepper noise. On multibaseline inputs, it suffers from an inherent weakness similar to CUR.

The **Maximum Margin** is reliable, but not outstanding in any particular task.

The **Probabilistic Measure** 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 Measure** is the second best method near discontinuities and arguably *the best method on multiview data.* It generates confidence maps with the sharpest boundaries, but it performs surprisingly poorly in disparity selection.

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

The **Negative Entropy Measure** does not perform well on binocular data, as noted also in [15]. It is significantly better in multiview experiments, particularly on the $fountain-P11$ dataset using SAD. We have not been able to explain this inconsistency.

The **Number of Inflection Points** does not work well because it merely considers the number of local minima, not all of which are viable disparity candidates.

The **Winner Margin** is usually worse than PKR, but still among the top methods in the multibaseline setting. It is worse than WMNN on binocular data, but better on multiview (see PKR and PKRN). It is effective for disparity selection.

The **Naive Winner Margin (WMNN)** is worse than PKRN. It is better on binocular data and worse on multiview data than WMN. It is worse than WMN at disparity selection due to bias for small windows.

The following methods were only evaluated on binocular data, as explained in Section 5.

The **Left Right Consistency** achieves average performance due to quantization. More than 50 percent of the matches in almost all experiments are left-right consistent, resulting in a very large set of matches that appear to have equal confidence. LRC cannot discriminate further to select a more reliable subset. LRC is effective in occlusion detection.

The **Left Right Difference** is one of the best overall method for binocular inputs. It also performs very well near discontinuities and in occlusion detection and is the *best method in disparity selection.*

The **Distinctiveness Maps** method does not perform well because it utilizes information from only one input image. Pixels with high confidence may not even be visible in the other image. It is average in disparity selection.

The performance of **Distinctive Similarity Measure** is much better than DTS since it makes use of both input images and considers the similarity between the corresponding pixels, but it is not particularly successful at disparity selection. In almost every experiment, the results on NCC cost volumes are much better than SAD.

The **Self-Aware Matching Measure** is the fourth best method for binocular inputs on average. It typically trails DSM, with which it is relatively similar theoretically, except for disparity selection, where it is one of the top methods.

More effective 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 measures within a learning approach should lead to significant progress, but the design of appropriate training and testing conditions that will allow the algorithm to generalize to different types of scenes is far from trivial. Training a classifier for selecting the most appropriate confidence measure for a particular stereo pair or multiple-view set, as in the work of Aodha et al. [46] and Reynolds et al. [47] for optical flow and time-of-flight sensors, respectively, may be a less ambitious but more promising path for future research.

ACKNOWLEDGMENTS

This work has been supported in part by the US National Science Foundation (NSF) Computing Research Infrastructure grant (CNS-0855218) and Google, Inc., via a Google Research Award.

REFERENCES

[1] D. Scharstein and R. Szeliski, "A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms," *Int'l J. Computer Vision,* vol. 47, nos. 1-3, pp. 7-42, 2002.

[2] H. Hirschmüller and D. Scharstein, "Evaluation of Cost Functions for Stereo Matching," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2007.

[3] 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," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2008.

[4] P. Kohli and P.H.S. Torr, "Measuring Uncertainty in Graph Cut Solutions," *Computer Vision and Image Understanding,* vol. 112, no. 1, pp. 30-38, 2008.

[5] M. Gong and Y. Yang, "Fast Unambiguous Stereo Matching Using Reliability-Based Dynamic Programming," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 6, pp. 998-1003, June 2005.

[6] X. Hu and P. Mordohai, "Evaluation of Stereo Confidence Indoors and Outdoors," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2010.

[7] D. Scharstein and R. Szeliski, "High-Accuracy Stereo Depth Maps Using Structured Light," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. I-195-I-202, 2003.

[8] P. Mordohai, "The Self-Aware Matching Measure for Stereo," *Proc. IEEE Int'l Conf. Computer Vision,* 2009.

[9] A. Bobick and S. Intille, "Large Occlusion Stereo," *Int'l J. Computer Vision,* vol. 33, no. 3, pp. 1-20, 1999.

[10] Q. Chen and G. Medioni, "A Volumetric Stereo Matching Method: Application to Image-Based Modeling," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. I-29-I-34, 1999.

[11] Y. Wei and L. Quan, "Region-Based Progressive Stereo Matching," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. I-106-I-113, 2004.

[12] M. Lhuillier and L. Quan, "A Quasi-Dense Approach to Surface Reconstruction from Uncalibrated Images," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 3, pp. 418-433, Mar. 2005.

[13] J. Cech and R. Sara, "Efficient Sampling of Disparity Space for Fast and Accurate Matching," *Proc. CVPR Workshop Towards Benchmarking Automated Calibration,* 2007.

[14] L. Matthies, "Stereo Vision for Planetary Rovers: Stochastic Modeling to Near Real-Time Implementation," *Proc. SPIE,* vol. 1570, pp. 187-200, 1991.

[15] D. Scharstein and R. Szeliski, "Stereo Matching with Nonlinear Diffusion," *Int'l J. Computer Vision,* vol. 28, no. 2, pp. 155-174, 1998.

[16] Z. Zhang and Y. Shan, "A Progressive Scheme for Stereo Matching," *Proc. Second European Workshop 3D Structure from Multiple Images of Large-Scale Environments,* pp. 68-85, 2001.

[17] H. Hirschmüller, P. Innocent, and J. Garibaldi, "Real-Time Correlation-Based Stereo Vision with Reduced Border Errors," *Int'l J. Computer Vision,* vol. 47, nos. 1-3, pp. 229-246, 2002.

[18] 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," *Proc. IEEE Int'l Conf. Computer Vision,* 2007.

[19] K. Yoon and I. Kweon, "Distinctive Similarity Measure for Stereo Matching under Point Ambiguity," *Computer Vision and Image Understanding,* vol. 112, no. 2, pp. 173-183, 2008.

[20] G. Egnal, M. Mintz, and R. Wildes, "A Stereo Confidence Metric Using Single View Imagery with Comparison to Five Alternative Approaches," *Image and Vision Computing,* vol. 22, no. 12, pp. 943-957, 2004.

[21] M. Goesele, N. Snavely, B. Curless, H. Hoppe, and S. Seitz, "Multi-View Stereo for Community Photo Collections," *Proc. IEEE Int'l Conf. Computer Vision,* 2007.

[22] M. Pollefeys, D. Nistér, J.M. Frahm, A. Akbarzadeh, P. Mordohai, B. Clipp, C. Engels, D. Gallup, S.J. Kim, P. Merrell, C. Salmi, S. Sinha, B. Talton, L. Wang, Q. Yang, H. Stewénius, R. Yang, G. Welch, and H. Towles, "Detailed Real-Time Urban 3D Reconstruction from Video," *Int'l J. Computer Vision,* vol. 78, nos. 2/3, pp. 143-167, 2008.

[23] M. Jancosek, A. Shekhovtsov, and T. Pajdla, "Scalable Multi-View Stereo," *Proc. IEEE Int'l Workshop 3-D Digital Imaging and Modeling,* pp. 1526-1533, 2009.

[24] Y. Furukawa and J. Ponce, "Accurate, Dense, and Robust Multiview Stereopsis," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 32, no. 8, pp. 1362-1376, Aug. 2010.

[25] M. Brown, D. Burschka, and G. Hager, "Advances in Computational Stereo," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 25, no. 8, pp. 993-1008, Aug. 2003.

[26] S.M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 519-528, 2006.
[27] D. Neilson and Y. Yang, "Evaluation of Constructable Match Cost Measures for Stereo Correspondence Using Cluster Ranking," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2008.
[28] H. Hirschmüller and D. Scharstein, "Evaluation of Stereo Matching Costs on Images with Radiometric Differences," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 31, no. 9, pp. 1582-1599, Sept. 2009.
[29] M. Gong, R. Yang, L. Wang, and M. Gong, "A Performance Study on Different Cost Aggregation Approaches Used in Real-Time Stereo Matching," *Int'l J. Computer Vision,* vol. 75, no. 2, pp. 283-296, 2007.
[30] F. Tombari, S. Mattoccia, L. di Stefano, and E. Addimanda, "Classification and Evaluation of Cost Aggregation Methods for Stereo Correspondence," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2008.
[31] M. Bleyer, S. Chambon, U. Poppe, and M. Gelautz, "Evaluation of Different Methods for Using Colour Information in Global Stereo Matching Approaches," *Proc. Int'l Soc. for Photogrammetry and Remote Sensing,* pp. 63-68, 2008.
[32] M. Bleyer and S. Chambon, "Does Color Really Help in Dense Stereo Matching?" *Proc. Int'l Symp. 3D Data Processing, Visualization, and Transmission,* 2010.
[33] C. Dima and S. Lacroix, "Using Multiple Disparity Hypotheses for Improved Indoor Stereo," *Proc. IEEE Int'l Conf. Robotics and Automation,* vol. 4, pp. 3347-3353, 2002.
[34] S. Lefebvre, S. Ambellouis, and F. Cabestaing, "A Colour Correlation-Based Stereo Matching Using 1D Windows," *Proc. IEEE Conf. Signal-Image Technologies and Internet-Based System,* pp. 702-710, 2007.
[35] G. Egnal and R. Wildes, "Detecting Binocular Half-Occlusions: Empirical Comparisons of Five Approaches," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 24, no. 8, pp. 1127-1133, Aug. 2002.
[36] P. Jodoin, C. Rosenberger, and M. Mignotte, "Detecting Half-Occlusion with a Fast Region-Based Fusion Procedure," *Proc. British Machine Vision Conf.,* pp. I-417-I-426, 2006.
[37] D. Marr and T. Poggio, "Cooperative Computation of Stereo Disparity," *Science,* vol. 194, no. 4262, pp. 283-287, 1976.
[38] J. Sun, Y. Li, S. Kang, and H. Shum, "Symmetric Stereo Matching for Occlusion Handling," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 399-406, 2005.
[39] Y. Deng, Q. Yang, X. Lin, and X. Tang, "Stereo Correspondence with Occlusion Handling in a Symmetric Patch-Based Graph-Cuts Model," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 29, no. 6, pp. 1068-1079, June 2007.
[40] L. Xu and J. Jia, "Stereo Matching: An Outlier Confidence Approach," *Proc. European Conf. Computer Vision,* pp. 775-787, 2008.
[41] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int'l J. Computer Vision,* vol. 60, no. 2, pp. 91-110, 2004.
[42] R. Manduchi and C. Tomasi, "Distinctiveness Maps for Image Matching," *Proc. Int'l Conf. Image Analysis and Processing,* pp. 26-31, 1999.
[43] A. Bruhn and J. Weickert, "A Confidence Measure for Variational Optic Flow Methods," *Geometric Properties from Incomplete Data,* R. Klette, R. Kozera, L. Noakes, and J. Weickert, eds., pp. 283-297, Springer, 2006.
[44] J. Kostliva, J. Cech, and R. Sara, "Feasibility Boundary in Dense and Semi-Dense Stereo Matching," *Proc. CVPR Workshop Towards Benchmarking Automated Calibration,* 2007.
[45] D. Gallup, J.-M. Frahm, P. Mordohai, Q. Yang, and M. Pollefeys, "Real-Time Plane-Sweeping Stereo with Multiple Sweeping Directions," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2007.
[46] O. Aodha, G. Brostow, and M. Pollefeys, "Segmenting Video into Classes of Algorithm-Suitability," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 1054-1061, 2010.
[47] M. Reynolds, J. Doboš, L. Peel, T. Weyrich, and G.J. Brostow, "Capturing Time-of-Flight Data with Confidence," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2011.

Xiaoyan Hu received the BS degree in electrical engineering from Hunan University, China, in 2008. He is currently working toward the PhD degree in the Computer Science Department, Stevens Institute of Technology, where he conducts research in stereo vision-related topics. He is a student member of the IEEE.

Philippos Mordohai received the diploma in electrical and computer engineering from Aristotle University of Thessaloniki, Greece, in 1998 and the MS and PhD degrees, both in electrical engineering, from the University of Southern California in 2000 and 2005, respectively. He is an assistant professor of computer science at Stevens Institute of Technology. Prior to joining Stevens, he held postdoctoral researcher positions at the University of North Carolina and the University of Pennsylvania. His research interests include 3D reconstruction from images and video, range data analysis, perceptual organization, and manifold learning. He serves as an associate editor for the *Journal of Image and Vision Computing* and as a reviewer for numerous international journals and conferences. He has also organized several workshops and symposia. He received best reviewer awards from the Asian Conference on Computer Vision in 2010, the IEEE Conference on Computer Vision and Pattern Recognition in 2011, and the IEEE International Conference on Computer Vision in 2011. He is a member of the IEEE.