Evaluation of Stereo Confidence Measures on Synthetic and Recorded Image Data

Ralf Haeusler and Reinhard Klette
Computer Science Department, The University of Auckland
Tamaki Innovation Campus, Auckland
rhae001@aucklanduni.ac.nz

***Abstract*—We comparatively discuss a set of confidence measures for stereo analysis by testing them on semi-global matching (SGM) cost functions. The aim is a prediction of (potentially) erroneous areas in calculated disparity maps. The evaluation is done by using the sparsification technique which provides more information than commonly used RMS or NCC measures. We also present an approach for combining different confidence measures. This allows us to perform a quantisation of confidence estimates in terms of disparity errors.**

I. Introduction

Stereo matching is the computation of matching image locations across images with a defined parallax. This is a very active research area having newly designed algorithms published frequently due to still unsatisfying results for particular *situations* [10] when recording stereo images, for example over hours, days or months in the context of vision-based driver assistance. Performance evaluation on such *long* stereo image sequences is still a challenging subject. The use of a "third-eye strategy" provides one option, and it already helped to identify challenging *events* in recorded stereo image sequences [10].

Available disparity ground truth (for a few stereo image frames) supports performance rankings of stereo algorithms by a single number, such as the root-mean squared error (RMS) where *error* is defined by the absolute distance between calculated stereo result and measured or calculated ground truth [16]. Ground truth is either a rendered depth map of potentially very high accuracy for synthetic scenes, or a measured disparity map obtained by using alternative methods, for example a structured lighting approach, or a laser range finder. Those two methods have the potential to provide reasonably good depth estimates also on untextured areas, though not, for example, on highly specular surfaces.

Extensive stereo analysis studies in outdoor scenes cannot be supported by structured light, and the use of laser range-finders appears to be more suitable [12], [14]. High-accuracy laser range-finders allow us to reconstruct static 3D scenes at very high accuracy [9]. However, laser range-finder devices as applied for dynamic scenes do have insufficient spatial resolution for high-accuracy depth recoveries, especially at depth discontinuities [12]. Thus, calculating depth ground truth in dynamic outdoor scenes is still a challenging subject.

In summary, available stereo images with accurate depth ground truth are insufficient for challenging current stereo matching algorithms due to their simplicity, singularity, or missing diversity of situations or events, as they occur, for example, when running a stereo imaging system for days or months in a vehicle. As a result of currently used test environments, stereo matching algorithms are tuned to perform well on singular test images or very short sequences. The test environment [2] provides video sequences of up to a few hundreds of stereo frames.

Interestingly, current stereo matching of aerial images still applies often a simple correlation-based scheme, just because the correlation coefficient in block matching appears to be a quality measure for a match: A coefficient close to one indicates a "good match", and disparity values at pixels with small coefficients are discarded. This leads to sparse depth maps.

Common confidence measures on dense depth maps use, for example, the opening of local parabola fits on the matching cost function minimum, or the slope of an Okutomi fit [17], [5]. We show that these local measures carry less information about stereo confidence than other simple features.

A "more informative" confidence measure has been proposed for 3D reconstruction in [11]. For more proposals of confidence measures, see [8]. It appears that there is still a need for a comprehensive evaluation of these measures. We approach this by using the sparsification strategy, as used in [1] for discussing optical flow techniques.

Due to space limits, we can only discuss one stereo matcher in this article. The evaluation of depth confidence will be by using results of a "very well" performing stereo matcher. Our aim is to identify image areas posing substantial problems to stereo matchers in general, not just to poorly performing ones. We decided for semi-global matching (SGM) stereo analysis [6] using a census data term as a cost function. To our experience [10], this configuration performs "pretty well" both on synthetic as well as on recorded images. Best overall performance of this scheme was also reported in [7]. Moreover, computational costs of this stereo matcher configuration are sufficiently low to make realtime processing possible [3]. Figure 1 illustrates used test data.

Section II reflects some issues when comparing stereo results using a laser range-finder, verifying our statements above (see also [12]), thus highlighting the need of alternative evaluations. Sections III and IV present confidence measures used for evaluation. Many of those are based on the accumulated SGM cost cube. Section V briefly explains the

Fig. 1: A synthetic (*left*) and a recorded scene. Disparity ground truth is only available for the synthetic scene.

sparsification strategy used for our evaluation. Sections VI and VII contain results and a discussion. Section VIII concludes.

II. Evaluation using LRF data

One option to generate ground truth for recorded sequences is using depth measurements with a statistically independent sensor such as a laser range-finder (LRF). Figure 3 illustrates the approach of comparing stereo results with laser measurements: For each pixel where a laser measurement is available, the error to stereo estimates is computed.

We used a Velodyne laser scanner with 64 scanning lines. The spatial density of measurements is not as high as in the calculated stereo analysis results. Hence, only points with LRF measurement can be used for comparison. Major drawbacks of laser measurements are the "rolling shutter" problem due to a low scanning frequency of 2 Hz, ambiguous measurements on translucent surfaces such as glass, and poor accuracy at depth discontinuities due to viewing occlusions and difficulties in precise calibration of the extrinsics.

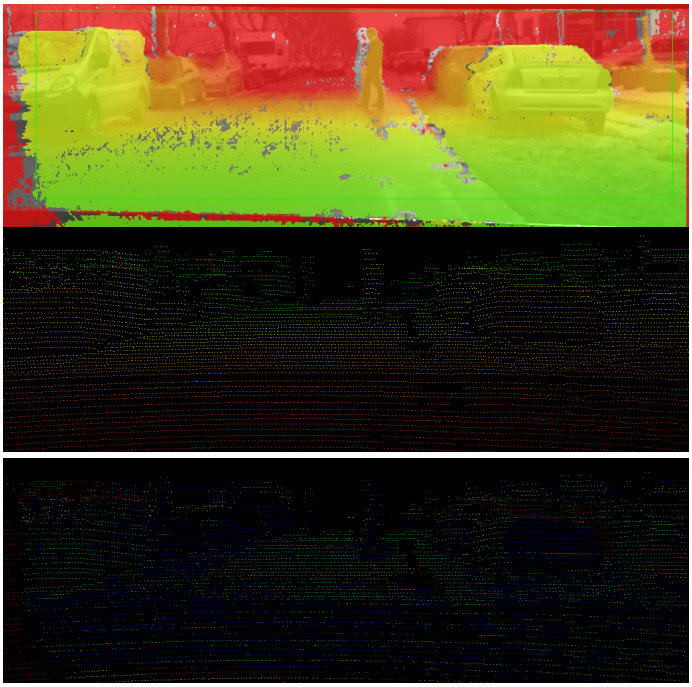

Fig. 3: Stereo result (*top*), LRF measurement (*middle*), and difference between both.

Class	Error range
1	0 – 0.5
2	0.5 – 1
3	1.0 – 2.0
4	2.0 – 5.0
5	5.0 – ∞

TABLE I: Definition of error classes.

The stereo result on the top of Figure 3 is colour coded, green for close objects to red for distant objects. The green box defines the region of interest for comparison. We avoid border areas, especially those with objects that are known to be too close for matching such as the car bonnet. In the middle of the figure, laser points projected into the image space (same image as in the top figure) indicate far (green) and near (red) objects. Note the erroneous measurements in the car windows. Finally, the bottom part of the figure shows differences between stereo and laser measurements. They are colour coded as follows: blue and cyan - disparity difference less than 1, green - disparity difference between 1 and 2, yellow - disparity error between 2 and 5, red - disparity error more than 5 px. This corresponds to the five classes outlined in Table I.

III. Measures based on (accumulated) cost

For each calculated depth value, we want to assign a label for the likelihood of being defined by a correct match. In some applications, such as 3D reconstruction, subsequent processing steps may use this information to assign a lower weight to calculated depth values having a low confidence weight. For an example of such a strategy, see [11].

Some of the measures used here are described more in detail in [8], but this paper lacks a comprehensive evaluation. No quantitative comparison of measure's performance is supported. We will provide such a technique based on sparsification plots (to be detailed in the next section). Figure 2 illustrates the used measures.

Let d be a disparity. We denote by $c(d)$ the accumulated cost associated with disparity d. We also use the notation that d_0 is the disparity with the minimum cost (i.e. defining not only

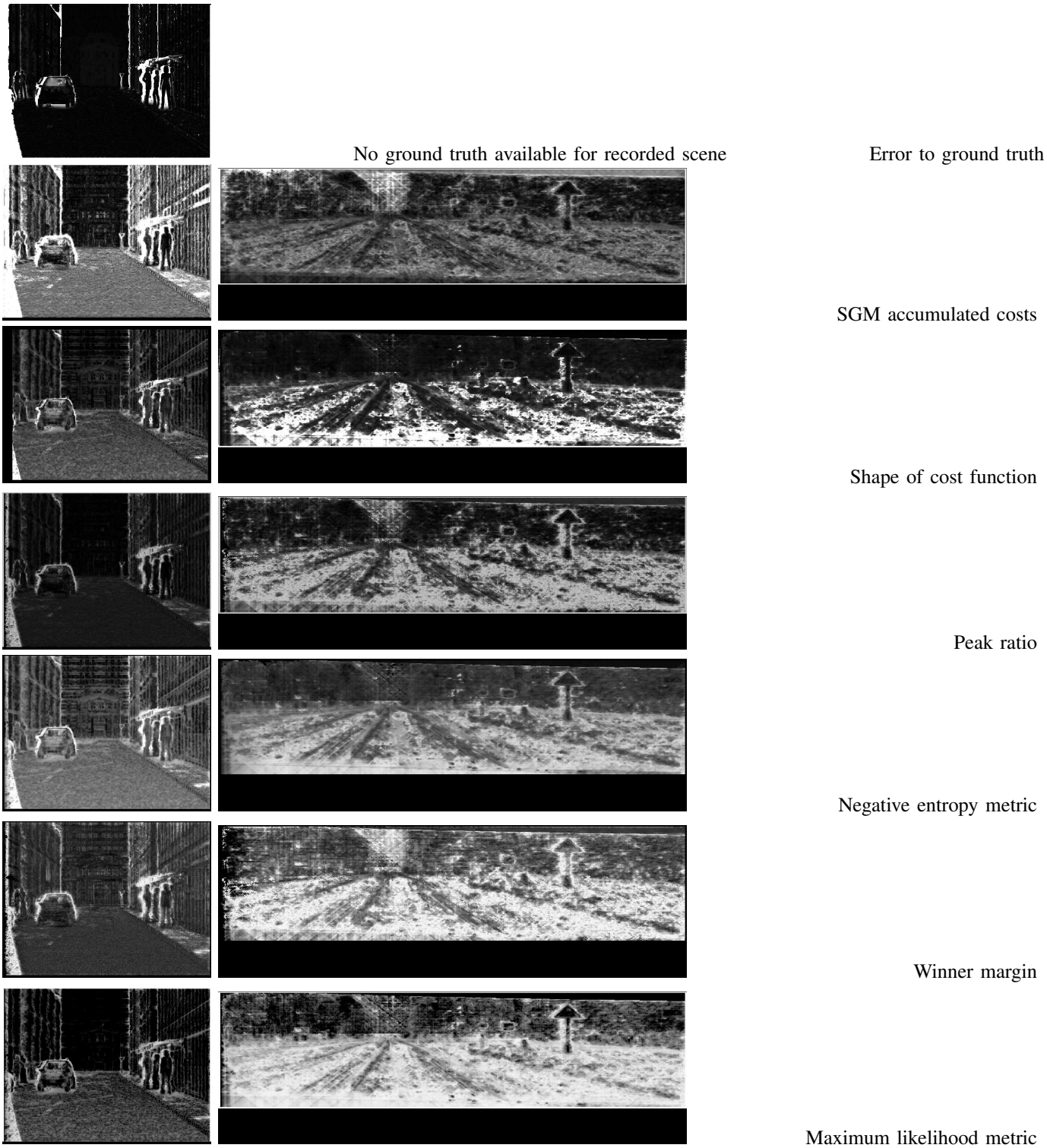

Fig. 2: Features (named on the right) for the synthetic and recorded scene as shown in Fig. 1. Darker grey values are associated with higher confidence (i.e. with a smaller expected error).

a local minimum but even a global minimum), followed by d_1 which has the second lowest cost at another local minimum, i.e. $c(d_1 - 1) > c(d_1)$ and $c(d_1) < c(d_1 + 1)$. We use a parameter σ for scaling.

The following measures are defined for the used cost function (to be computed for each pixel independently), see [8] for measures (1) to (7), and [11] for (8):

1) Minimum of accumulated costs: $c(d_0)$

2) Local curvature: $-2c(d_0) + c(d_0 - 1) + c(d_0 + 1)$
3) Peak ratio: $\frac{c(d_0)}{c(d_1)}$
4) Negative entropy metric: $-\sum_d p(d) \log p(d)$ with $p(d) = \frac{\exp(-c(d_0)/\sigma)}{\sum_d \exp(-c(d)/\sigma)}$
5) Winner margin: $(\frac{c(d_1)-c(d_0)}{\sum_d c(d)})$
6) Nonlinear margin: $\exp \frac{c(d_1)-c(d_0)}{2\sigma^2}$
7) Maximum-likelihood metric: $\frac{\exp(-\frac{(c(d_0)-c(d_1))^2}{2\sigma^2})}{\sum_d \exp(-\frac{(c(d)-c(d_1))^2}{2\sigma^2})}$
8) Shape of cost function: $\sum_{d \neq d_0} \exp - \frac{(c(d)-c(d_0))^2}{\sigma^2}$

Rationales for defining these measures are as follows:

1) Minimum of accumulated costs: High costs indicate reduced similarity between patches in the search range. This can be associated with a difficult matching situation. However, in SGM the accumulated cost values are not independent from cost values at neighbouring pixels.

2) Local curvature: This is one of the most commonly used features in literature. The definition may vary slightly. It can be the opening value of a parabola fit, or the slope of an Okutomi fit for subpixel estimation.

3) Peak ratio: The peak ratio measures whether there is another close candidate for matching. If the costs of both candidates (which are considered to define local minima) are very similar (i.e. their ratio is close to one), then matching is ambiguous, and thus unreliable.

4) Negative entropy metric: This measure computes the entropy of a cost measure that is transformed into a probability distribution function. Cost functions producing only noise carry little information (low entropy).

5) Winner margin: Here, not the cost ratio between the first and the second minimum is used (as it was done in the peak ratio feature), but the difference of these costs. Furthermore, this number is put into relation with the sum of all costs. This prefers cost functions with high overall costs, indicating that there are few disparities with low cost (i.e. good matches).

6) Nonlinear margin: The difference to the winner margin is that there is no weighting with the sum of all costs, making it computationally much cheaper. A nonlinear transformation is applied for obtaining a suitable range of values.

7) Maximum likelihood metric: This metric is part of the negative entropy metric (see above). Instead of computing the entropy, this metric is parametrized with a value for assumed noise. The underlying idea is again to convert the costs into a probability density function.

8) Shape of cost function: Here, the aim is to implement a measure being low if the cost function has a singular, "sharply defined" minimum and high if the cost function is flat or has more than one pronounced minimum.

IV. Measures not based on cost

The following measures can be defined without knowledge of the cost function values. They are based on image intensity, depth map, or both.

9) Disparity variance: Variance image of the disparity map with a specified patch window size. The very simple idea is to assume that errors in stereo mostly occur at depth discontinuities or noisy patches. The problem with this measure is that actually correctly estimated depth edges also have a high variance.

10) Foreground fattening (also referred to as surface overextension): This occurs at depth discontinuities with a contrast change in the image. It is rather difficult to model in general. An analysis for the sum of squared difference costs can be found in [13]. We propose the following in this paper: For a pixel at location (x, y) with associated disparity d_r, consider a patch around this location of the same size as the one used for cost calculation, and another patch of the same size at location $(x - i, y)$. Here i is half of the horizontal extent of the cost calculation block size (i.e. $i = 3$). The associated disparity at location $(x - i, y)$ be d_l. Our confidence measure is now $\left|(d_l - d_r) \times (\varsigma(I_{(x-i,y)}) - \varsigma(I_{(x,y)}))\right|$. Here $\varsigma(I_{(x,y)})$ is the variance of the image patch with the same size and location as the cost calculation window at location (x, y). – The idea behind this definition is that foreground fattening occurs at disparity discontinuities with contrast change at the same location of the image.

11) Normalized cross correlation: We also include the (usual) matching measure used in local stereo matching for confidence estimates. Intuitively, if the correlation between patches is not close to one, the match might be "bad".

12) Error compared to ground truth. This is actually an accuracy measure, but we include it here for comparison.

V. Evaluation of confidence measures

These 12 measures have advantages and disadvantages, depending on the actual cost function for a certain pixel. It appears to be desirable to find a way of combining selected

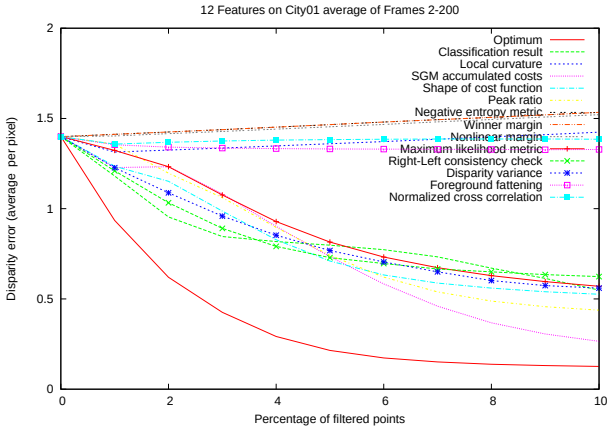

Fig. 4: Sparsification plot for 12 measures on the used synthetic sequence; see Fig. 1, left. Plot values are averaged over 200 frames of the sequence.

measures for obtaining "stronger" results. We tried to do so by supervised learning using a Gaussian mixture model for these measures. Classes of confidence are defined by the deviation of stereo results from ground truth (if there is ground truth available). We define five classes for magnitudes of stereo errors as shown in Table I.

For a more detailed description of the learning procedure, using an expectation maximization scheme, we refer to [4], where the idea was applied to the classification of optical flow errors.

The evaluation scheme of [15] is ranking algorithms by accumulating an error metric over three different fixed subsets of pixels. These subsets are (1) all pixels except border areas, (2) areas near occlusions and depth discontinuities, and (3) areas excluding occlusions and depth discontinuities.

A confidence measure for recorded images should indicate itself which image areas are "problematic", as ground truth for depth discontinuities or semi occlusions is not available in general. Thus, we use subsets of pixels which are included in the error measures as follows:

Initially, an error measure and a confidence measure are computed for all pixels. Next, a certain fraction of all pixels (e.g. 1%) is filtered out, and the error metric is computed on the remaining percentile. The fraction to be filtered out in each iteration, is determined by the confidence measure (i.e. the 1% of pixels with lowest confidence score are excluded).

The process terminates when all pixels are excluded. The values of the error measure (the average absolute error per pixel in our case) for each filtering stage constitute a *sparsification plot*. If ground truth is available, it is possible to plot an optimal line by using the error to ground truth instead of the confidence measure for filtering. Then, a "more convincing" quantitative comparison of performance for different confidence measures can be produced: The closer the plot of a measure to the optimal line, the better the measure performs in detecting errors.

VI. Results

The sparsification plot in Fig. 4 indicates the performance in detecting stereo errors on a synthetic sequence. Although not reaching the optimal curve, the following features show good performance: Minimum of accumulated costs, disparity variance, maximum likelihood metric, shape of cost function, and left-right consistency check. The following features do not perform better than excluding pixels randomly: Nonlinear margin, winner margin, negative entropy metric, foreground fattening, and normalized cross correlation.

By combining measures it is possible to slightly outperform any single measure in the top 3% of erroneous pixels.

The confusion matrix in Table II suggests that classification based on Gaussian mixture models trained on "favourable" measures seems to work well for small errors (below 0.5 px) and large errors (above 5 px). However it does not outperform all the individual measures.

The LRF data consists of 65 frames that were manually selected such that the laser measurement was visually in good accordance with the corresponding image, i.e. laser measurements of "obviously poor quality" is not part of the evaluation.

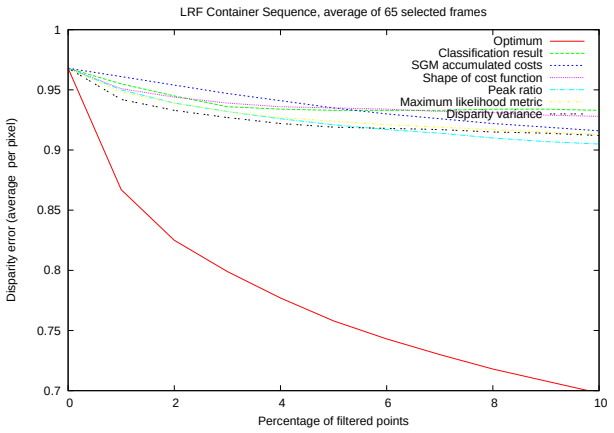

Fig. 5: Sparsification plot of five measures on recorded sequence and sparsification results of classification derived from these measures. Comparison for stereo results is done with measurements from a laser range-finder as illustrated in Fig. 3.

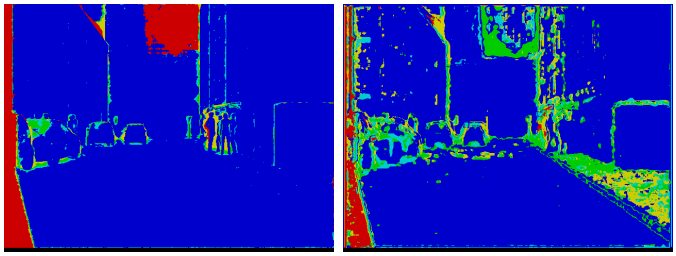

Fig. 6: Colour coded error image and classification results using six features and a Gaussian mixture model with two splits, trained in three iterations. Colour coding for error classes is as follows: blue - less than 0.5 px error, cyan - 0.5 px to 1 px error, yellow - 1 px to 2 px error, green - 2 px to 5 px error, red - more than 5 px error.

The comparison of measure's performance, shown in Fig. 5, suggests that none of the measures that work well on the synthetic scene have any contribution in the recorded scene when compared to the LRF data.

	c=0	c=1	c=2	c=3	c=4
k=0	19560298	64397	18663	16497	87124
k=1	843612	128005	36957	31428	106371
k=2	923537	85864	76631	61198	139531
k=3	863632	34761	28635	112344	222802
k=4	75439	16135	13245	30326	944184
c=k	87.8%	38.9%	44.0%	44.6%	62.9 %

TABLE II: Confusion matrix of classification results for all pixels of one image from the City01 sequence. k is the classification result, and c the correct class.

VII. Discussion

The measures listed in Sections III an IV have a performance comparable to a right-left consistency check. However, some of these are computationally much cheaper. The peak ratio, for example, only needs to find the second smallest local minimum of the accumulated costs.

The bad performance of the parabola fit measure can be explained as follows: A cluttered cost function has a sharp local peak (high curvature) at its minimum location, but it is very likely that there not just a single one. In contrary, an unambiguous global minimum does not necessarily have high curvature. The normalized cross correlation between image patches as a confidence measure is problematic for a number of reasons: Patches of homogeneous image regions may have high correlation in synthetic images, but low correlation in recorded images due to sensor noise. Results are therefore arbitrary. The foreground fattening measure may not be successful as it can fail to properly locate the erroneous pixel.

There is a general weakness of sparsification plots that should be mentioned: If a measure tends to overestimate an error, it can still indicate good performance. For example, the disparity variance is high even on correctly computed disparity steps. Then, a wrong removal is not obvious in the sparsification plot.

The LRF comparison suggests that the chosen measures may not work at all on recorded sequences. However, a visual comparison of stereo errors with the measure images cannot confirm this assumption. In consequence, sparsification plots are not suitable for comparison of LRF measurement with stereo results. This is due to the poor quality of laser data as explained in Section II. They are "most dramatic" on disparity discontinuities. As most stereo errors occur in such areas, the LRF measurements would have to be excluded in those, leaving no relevant data for comparison.

On recorded images we observed (visually) good results also for the winner margin and negative entropy measures (see Fig. 2). This, however, cannot be shown accurately due to missing ground truth. Yet, it indicates that the characteristics of the presented measures on synthetic data do not allow conclusions about their performance on real world data; a statement already made similarly for other evaluations in [10].

The unsatisfactory result of classification is likely due to poor fitness of data to the Gaussian mixture model. In particular, we observed large intra-class variances that result in more or less arbitrary decisions for a specific class. It may be worthwhile to examine the results of different classification methods such as support vector machines, or to attempt the definition of more discriminative measures.

The advantage of classification is given by a quantification of stereo errors. These cannot be deduced from the measures themselves due to their continuous nature.

VIII. Conclusion

We used a method that allows to quantitatively compare confidence measures that were derived from SGM accumulated costs. However, such evaluation can be done only on synthetic data or data with very accurate ground truth. The characteristics of errors are substantially different from those ones on recorded data, especially when recorded outdoors. In particular, errors occurring in our synthetic data can be all attributed to the well known class of foreground fattening [13]. However, foreground fattening is not introduced by the SGM accumulation step, hence cannot be detected from cost function values. Another approach for detecting this problem has to be incorporated.

Acknowledgment

The authors thank Uwe Franke and his research group at Daimler A.G., Germany, for providing support for performing the discussed experiments involving a laser range-finder and for providing image data.

References

[1] Bruhn, A., Weickert, J.: A confidence measure for variational optic flow methods. In *Geometric Properties for Incomplete Data* (Klette, R., Kozera, R., Noakes, L., Weickert, J., editors), pages 283–298, Springer, 2006.

[2] EISATS (*.enpeda..* image sequence analysis test site): The University of Auckland, www.mi.auckland.ac.nz/EISATS, last visit: 13 Feb 2012.

[3] Ernst, I., Hirschmüller, H.: Mutual information based semi-global stereo matching on the GPU. In Proc. *ISVC*, LNCS 5358, pages 228–239, Springer, 2008.

[4] Gehrig, S., Scharwächter, T.: A real-time multi-cue framework for determining optical flow confidence. In Proc. *CVVT:E2M*, ICCV Workshops, pages 1978–1985, 2011.

[5] Hermann, S., Vaudrey, T.: The gradient - a powerful and robust cost function for stereo matching. In Proc. *IVCNZ*, IEEE 978-1-4244-9631-0, 2010.

[6] Hirschmüller, H.: Accurate and efficient stereo processing by semi-global matching and mutual information. In Proc. *CVPR*, volume 2, pages 807–814, 2005.

[7] Hirschmüller, H., Scharstein, D.: Evaluation of cost functions for stereo matching. In Proc. *CVPR*, pages 1–8, 2007.

[8] Hu, X., Mordohai, P.: Evaluation of stereo confidence indoors and outdoors. In Proc. *CVPR*, pages 1466–1473 , 2010.

[9] Huang, F., Klette, R., Scheibe, K.: *Panoramic Imaging - Sensor-Line Cameras and Laser Range-Finders*. John Wiley & Sons Ltd, Chichester, 2008.

[10] Klette, R., Krüger, N., Vaudrey, T., Pauwels, K., van Hulle, M., Morales, S., Kandil, F., Haeusler, R., Pugeault, N., Rabe, C., Lappe, M.: Performance of correspondence algorithms in vision-based driver assistance using an online image sequence database. *IEEE Trans. Vehicular Technology*, 60:2012–2026, 2011.

[11] Merrell, P., Akbarzadeh, A., Wang, L., Mordohai, P., Frahm, J.-M., Yang, R., Nister, D., Pollefeys, M.: Real-time visibility-based fusion of depth maps. In Proc. *ICCV*, pages 1–8, 2007.

[12] Morales, S., Klette, R.: Ground truth evaluation of stereo algorithms for real world applications. In Proc. *CVVT:E2M*, ACCV Workshops, LNCS 6469, pages 152–162, 2011.

[13] Okutomi, M., Katayama, Y., Oka, S.: A simple stereo algorithm to recover precise object boundaries and smooth surfaces. *Int. J. Computer Vision*, 47: 261–273, 2002.

[14] Reulke, R., Luber, A., Haberjahn, M., Piltz, B.: Validierung von mobilen Stereokamerasystemen in einem 3D-Testfeld. In Proc. 3D-NordOst, DLR, Berlin, 2009.

[15] Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Computer Vision*, 47: 7–42, 2002.

[16] Scharstein, D., Szeliski, R.: High-accuracy stereo depth maps using structured light. In Proc. *CVPR*, pages 195–202, 2003.

[17] Wedel, A., Brox, T., Vaudrey, T., Rabe, C., Franke, U., Cremers, D.: Stereoscopic scene flow computation for 3D motion understanding. *Int. J. Computer Vision*, 95:29–51, 2011.