Disparity Confidence Measures on Engineered and Outdoor Data

Ralf Haeusler and Reinhard Klette

Department of Computer Science, The University of Auckland
Auckland, New Zealand
rhae001@aucklanduni.ac.nz

Abstract. Confidence measures for stereo analysis are not yet a subject of detailed comparative evaluations. There have been some studies, but still insufficient for estimating the performance of these measures. We comparatively discuss confidence measures whose performance appeared to be 'promising' to us, by evaluating their performance on commonly used stereo test data. Those data are either engineered and come with accurate ground truth (for disparities), or they are recorded outdoors and come with approximate ground truth. The performance of confidence measures varies widely between these two types of data. We propose modifications of confidence measures which can improve their performance on outdoor data.

1 Errors or Confidence Values in Stereo Analysis

With the current application of stereo vision to a variety of imaging tasks, the *reliability* of stereo vision became also a research topic in itself. On rendered or engineered data (e.g., data discussed in [14], or Sets 2 and 7 on [2]), state-of-the-art stereo analysis algorithms are capable of computing depth maps of satisfying quality. However, this differs on image data taken under adverse lighting conditions as they are common for outdoor scenes (e.g., real-world stereo video data on [2]). Outdoor scenes are classified in *situations* in [8], defined by *events* such as lighting artefacts, traffic scenes in the night and sun strikes. Stereo reconstruction appears to be impossible with current methods for such situations. In an abstract sense, critical situations are, for example, if both camera recordings do not satisfy the brightness constancy assumption, or if (e.g. around a recorded light such as in the "Night" sequence in Set 5 on [2]) intensities are nearly constant in some image regions.

The quality of disparity maps is usually rated globally (i.e. summarizing for one disparity map of a given stereo frame). If disparity ground truth is available, common *error measures* are the *root-mean squared error* (RMS) or the *normalized cross-correlation* (NCC) between given and calculated disparities. [14] initiated a ranking of a large number of stereo matchers but only on a small number of stereo frames. Current results on those stereo frames show that they do not represent a true challenge anymore for state-of-the-art stereo algorithms. Prediction error analysis in [16] for the case of optical flow analysis has been adapted

L. Alvarez et al. (Eds.): CIARP 2012, LNCS 7441, pp. 624–631, 2012.

to disparity error analysis in [10] for stereo frames where disparity ground truth is not available. The used error measures provide again only one summarizing value for each stereo frame of a stereo video sequence. Such values are likely to be meaningless in critical situations as specified above.

Confidence measures are designed to provide local (i.e. *pixelwise*) evaluations for identifying regions of failure. Such pixelwise measures are defined based on values of the data cost function (used in the stereo matcher) at a pixel; the global minimum of those values defines usually the selected disparity (ignoring for simplicity the influence of a smoothness term in the stereo matcher). See Fig. 1 for an example; in our experiments we use semi-global matching [5] with the census cost function and 8-path optimization (SGM). An example is shown in Fig. 2.

Confidence measures may be defined on data derived from the cost function (e.g. "around" the global minimum) of the used stereo matcher. The *left-right consistency check* is a common way in stereo matching to accept only results where left-to-right and right-to-left matching defines (about) the same disparity. For a compilation of a number of confidence measures, see [1,7]. Both papers do not provide a comprehensive evaluation of the performance of confidence measures, in particular for stereo data recorded outdoors.

Sophisticated, but computationally more expensive disparity confidence measures have been proposed by [13] (used here directly for the stereo matching process), or in [9] (a perturbation measure) when aiming at improved 3D reconstructions of outdoor scenes. For these two measures, every single cost value contributes, and this makes these measures computationally expensive.

A popular confidence measure is the *opening of the parabola* fitted to the global minimum (and its immediate neighbours) of the cost function. This opening is equivalent to the curvature at the vertex of the parabola. It is assumed that a "wide valley" around the minimum indicates a mismatch, whereas a "narrow valley" is likely to indicate a correct match. [17] used this confidence measure for improving scene flow by enhancing the used stereo-analysis module.

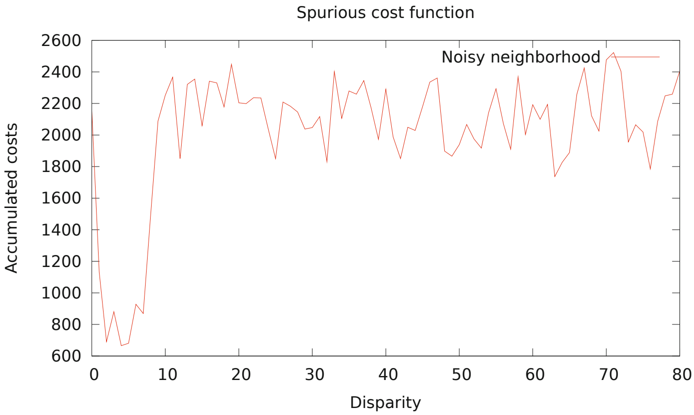

Fig. 1. Values of the census cost function at one pixel (of a recorded stereo video)

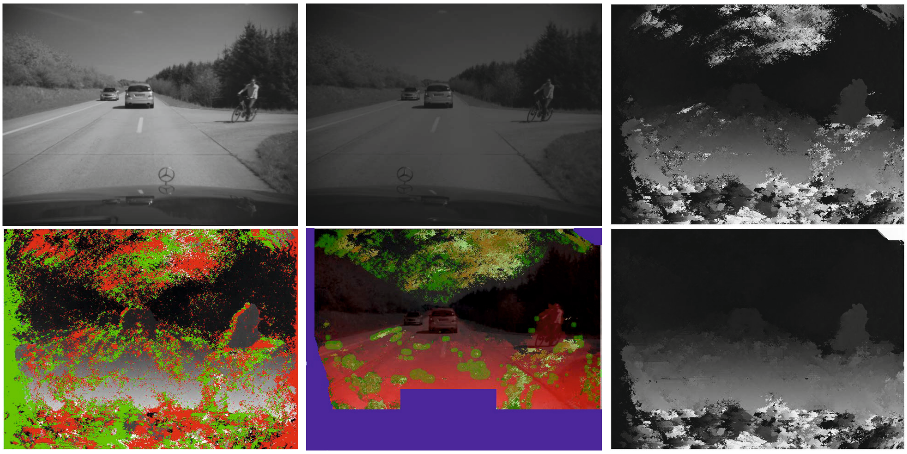

Fig. 2. Top row: Outdoor stereo pair (from Set 1 on [2]) and depth map (SGM as specified in the text without any post-processing). Bottom row: Labels from the left-right consistency check (green for occluded, red for other mismatches), manually assigned labels of bad matches (green), and of areas excluded from evaluation (blue); depth map after interpolating manually labelled areas.

The *peak ratio* (basic idea as known from feature matching) is a confidence measure based on comparing two values of a cost function. The second-smallest cost value is usually a neighbour of the global minimum; the peak ratio uses that local minimum of the cost function having the second smallest cost value (considered to be a competing matching candidate). Figure 1 illustrates a pixel where the peak ratio would not be a 'reasonable' confidence value.

There is still a lack of comparative evaluations of such *disparity confidence measures*, and also of a more systematic approach towards the possible design of new measures. Due to the general lack of disparity ground truth for outdoor scenes, evaluations are typically restricted to rendered or engineered indoor stereo data, and only in a few rare cases to outdoor stereo data with approximated disparity ground truth. This paper contributes to such comparative evaluations, suggests new ways for designing confidence measures, and also shows ways how to use recorded outdoor stereo data more widely in these studies.

The paper is structured as follows: Section 2 provides formal definitions of confidence measures with our proposed modifications. Section 3 explains the evaluation method for confidence measures. Section 4 contains results and discussion. Section 5 concludes.

2 Disparity Confidence Measures

A confidence measure C is defined pixelwise for the selected disparity values. These disparities are usually defined by the global minimum of a cost function c. In our case, c is resulting from enforcing smoothness constraints to the disparity values by aggregating according to the semi global matching heuristic. For a pixel in the left image, c is defined for disparities in an interval $[d_{min}, d_{max}]$; $c(d)$ is the cost for disparity d.

We identify two special disparities: d_0, where $c(d_0)$ is the global minimum, and d_1, where $c(d_1)$ also defines a local minimum but which is only the second smallest globally.

Curvature. Local curvature of c at the cost minimum d_0 is a widely used confidence measure. We use the inverse of the opening of a fitted parabola:

$$C_0 = \frac{1}{-2c(d_0) + c(d_0 - 1) + c(d_0 + 1)} \tag{1}$$

Perturbation. The perturbation measure, proposed in [9], computes the deviation from an ideal cost function that has a single minimum at location d_0 and is 'very large' everywhere else. Nonlinear scaling is applied:

$$C_1 = \sum_{d \neq d_0} e^{-\frac{(c(d_0) - c(d))^2}{\sigma^2}} \tag{2}$$

Parameter σ is chosen to obtain a valid range of confidence values regarding numerical precision limits.

Peak ratio. The peak ratio indicates low confidence if there are two candidates with similar matching costs. It is defined as

$$C_2 = \frac{c(d_0)}{c(d_1)} \tag{3}$$

Right-left consistency check. Right-left consistency compares the selected disparities of left-to-right and right-to-left matching. Let d_0^R be the global minimum of the cost function for right-to-left. Then,

$$C_4 = |d_0 - d_0^R| \tag{4}$$

Large disparity differences between both views show an incorrect match, at least for one of both. In practice, a difference of more than one pixel is considered to be an indication of a mismatch. Defining values smaller than 1 is of questionable value for a confidence measure.

Proposed Modifications. Inspired by the observation that the neighbourhood of the global cost function minimum on recorded stereo frames contains little information about the correctness of the match (see the plot of the curvature measure in Fig. 4), we exclude a neighbourhood of size n from the confidence computation, where n equals (about) half of the matching window size. The perturbation measure is then defined as follows:

$$C_1(x, y) = \sum_{d \in [d_{min}, d_0 - n] \cup [d_0 + n, d_{max}]} e^{-\frac{(c(d_0) - c(d))^2}{\sigma^2}} \tag{5}$$

For the peak ratio, in addition to being a local minimum of c, the following constraint is applied for the selection of d_1: $d_1 < d_0 - n$ and $d_1 > d_0 + n$.

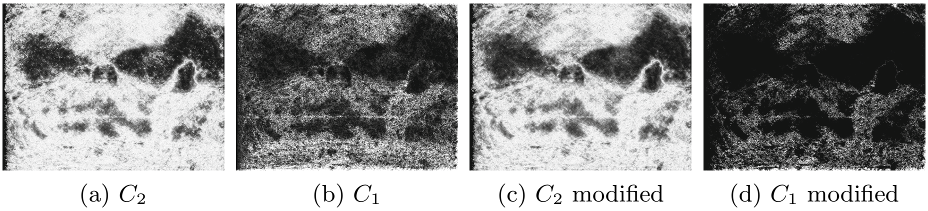
(a) C_2 (b) C_1 (c) C_2 modified (d) C_1 modified

Fig. 3. Visualisation of peak ratio (a) and perturbation measure (b) and their modified counterparts (c),(d) for the stereo frame displayed in Fig. 2. Lighter grey values indicate locations with lower confidence into the calculated disparity.

3 Synthetic, Engineered and Recorded Data

We tested confidence measures defined on accumulated cost of SGM (defined above) using a census [3] cost function instead of the originally used mutual information [5]. This choice is justified by good overall performance of census costs, as, for example, reported in [6].

We use a *sparsification strategy* for comparing the performance of measures: initially, the number of bad pixels is counted on the disparity map with full density; successively, pixels with lowest confidence (or highest score assigned, respectively) are removed, resulting in semi-dense disparity maps; for each disparity density, the number of bad pixels is counted, until the set of points in the disparity map is empty.

This evaluation requires disparity ground truth. For synthetic scenes, ground truth is available with very high accuracy, without any mismatches; see Set 2 on [2]. Accurate ground truth for indoor scenes can be obtained using the structured lighting technique [15]. Subpixel accuracy is available on downsampled images. Several data sets with ground truth, generated using structured lighting, have been published in conjunction with [14,12].

For outdoor scenes, one of the few methods to generate depth measurements is using a laser range-finder [4,11]. Drawbacks of this technique include: misregistration of camera and range-finder sensors, non-overlapping occluded areas, low density of laser measurements, and numerous measurement artefacts on specular surfaces.

For the evaluation of disparity confidence measures we count the number of bad pixels (according to some criterion), not the total deviation from ground truth (e.g., as done with RMS). Therefore it is more important to have accurate maps of gross mismatches than very accurate disparity measurements itself. For the evaluation of recorded data, we choose to manually enhance a disparity map by labelling bad pixels; see Fig. 2. To compensate for unavoidable inaccuracies on real-world data, we identify as bad matches only disparities that differ from the ground truth value by eight units or more. On synthetic and engineered data, a difference of more than 1 defines a bad match.

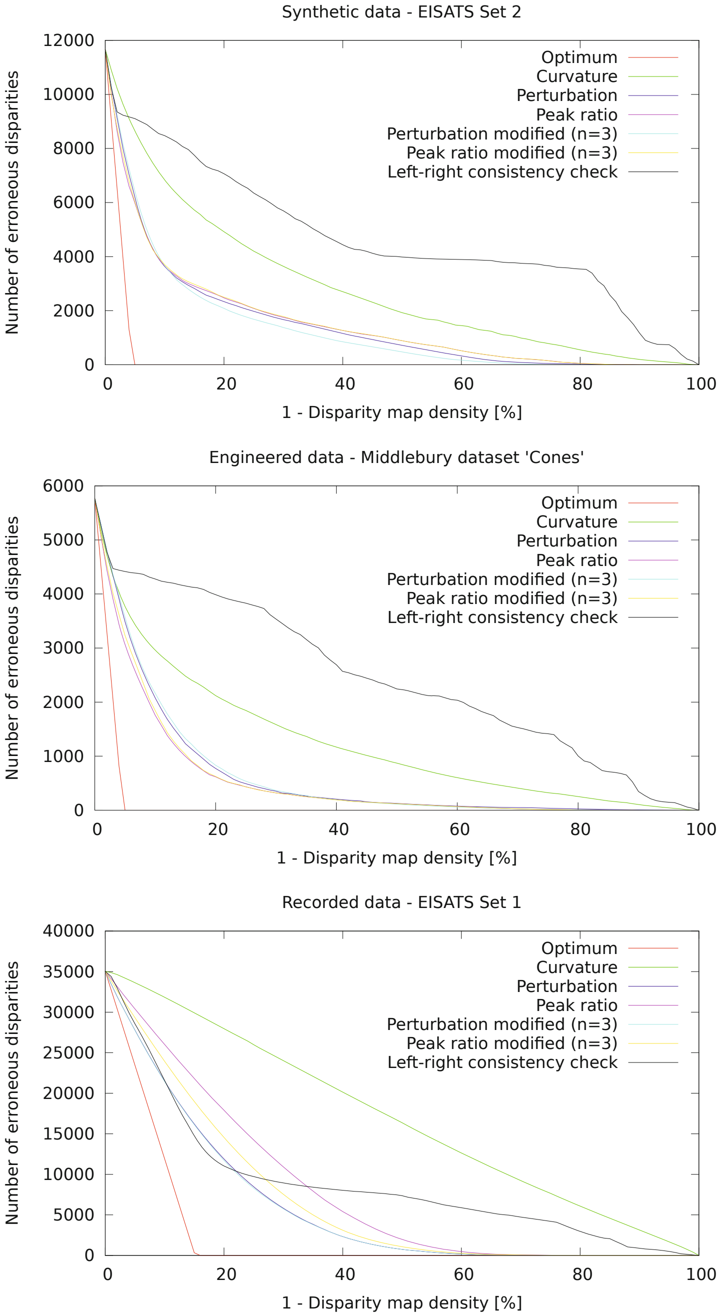

Fig. 4. Sparsification plots of confidence measures. Top to bottom: synthetic, engineered, and outdoor-recorded stereo input data.

4 Results

Results from sparsification are plotted in Fig. 4. The curvature measure is generally the worst performing one, despite its popularity. On the recorded data set it even provides no information about confidence at all. The proposed modifications of perturbation and peak ratio measure are not advantageous on synthetic and engineered data. The modified peak ratio significantly outperforms the original peak ratio measure on the recorded dataset. Improvements in sparsification (see Fig. 4)for the modified perturbation measure are minor for the used dataset, in contrast to what the visualization in Figure 3 suggests. With the modification of the peak ratio feature, an effective noise reduction in feature space can be achieved: See, e.g., in Fig. 3 that areas depicting trees in the recorded images (see Fig. 2) are well estimated by the used stereo matcher, and the modified peak ratio measure better reflects this than the original one. The main advantage of the proposed feature modification is avoiding false positives in detecting stereo errors.

In the following, we discuss reasons for deviations in confidence measure accuracies. For the curvature measure, it is important to note that due to limited sharpness in recorded data, the global minimum of the accumulated cost function is never a sharp peak. Therefore, parabola fits at the cost minimum (and immediate neighbours) never yield large values for curvature, except at noisy patches. However, such patches may not generate a correct match. Extending the parabola fit to, for example, a least-squares fit of a wider neighbourhood might help. In the perturbation measure, for the same reason as explained above, excluding a neighbourhood of the global cost minimum gives more weight to other minima, and can enhance the distinctiveness of this feature. The largest potential for confidence measure accuracy improvements seems to be in peak ratio modifications: In recorded data, due to inherent noise, there is often another local minimum only a few (e.g. two) disparities away from the global minimum (see, e.g., Fig. 1). This local minimum usually has an associated cost very close to that of the global minimum, hence produces a very high peak ratio, or a low confidence, respectively. However, such matches are often correct or have a minor disparity inaccuracy. So, it is not desirable to exclude them from subsequent computations using these disparities. It may be of interest to scale n, the exclusion window size, depending on disparity d, as matching errors in more distant objects produce larger absolute errors in object space. Note, however, that this does not influence evaluations based on metrics using the number of bad pixels.

5 Conclusion

We have shown that popular confidence features behave significantly different on synthetic or engineered data on the one hand, compared to outdoor data on the other hand. We conclude that conclusions from evaluations of such measures on synthetic or engineered data are of no value for outdoor data.

However, characterizing different confidence measures on outdoor data in terms of their properties (e.g. performance in dependence of given situations, as discussed in [8]) requires more extensive experiments than reported in this paper. This will also help to identify particular signal or geometry cases where stereo matchers may fail and need to be improved.

References

1. Banks, J., Corke, P.I.: Quantitative evaluation of matching methods and validity measures for stereo vision. Int. J. Robotic Research 20, 512–532 (2001)
2. EISATS (.enpeda.. image sequence analysis test site): The University of Auckland, `www.mi.auckland.ac.nz/EISATS` (last visit: February 13, 2012)
3. Egnal, G.: Mutual information as a stereo correspondence measure. Computer and Information Science, University of Pennsylvania, Philadelphia, Tech. Rep. MS-CIS-00-20 (2000)
4. Haeusler, R., Klette, R.: Evaluation of Stereo Confidence Measures on Synthetic and Recorded Image Data. In: Proc. IEEE ICIEV (to appear, 2012)
5. Hirschmüller, H.: Stereo processing by semiglobal matching and mutual information. IEEE Trans. Pattern Analysis Machine Intelligence 30, 328–341 (2008)
6. Hirschmüller, H., Scharstein, D.: Evaluation of cost functions for stereo matching. In: Proc. CVPR, pp. 1–8 (2007)
7. Hu, X., Mordohai, P.: A quantitative evaluation of confidence measures for stereo vision. In: IEEE Trans. Pattern Analysis Machine Intelligence (2012), doi:10.1109/TPAMI.2012.46
8. 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)
9. 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, pp. 1–8 (2007)
10. Morales, S., Klette, R.: A Third Eye for Performance Evaluation in Stereo Sequence Analysis. In: Jiang, X., Petkov, N. (eds.) CAIP 2009. LNCS, vol. 5702, pp. 1078–1086. Springer, Heidelberg (2009)
11. Morales, S., Klette, R.: Ground Truth Evaluation of Stereo Algorithms for Real World Applications. In: Koch, R., Huang, F. (eds.) ACCV Workshops 2010, Part II. LNCS, vol. 6469, pp. 152–162. Springer, Heidelberg (2011)
12. Scharstein, D., Pal, C.: Learning conditional random fields for stereo. In: Proc. CVPR, pp. 1–8 (2007)
13. Scharstein, D., Szeliski, R.: Stereo matching with nonlinear diffusion. Int. J. Computer Vision 28, 155–174 (1998)
14. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Int. J. Computer Vision 47, 7–42 (2002)
15. Scharstein, D., Szeliski, R.: High-accuracy stereo depth maps using structured light. In: Proc. CVPR, pp. 195–202 (2003)
16. Szeliski, R.: Prediction error as a quality metric for motion and stereo. In: Proc. ICCV, pp. 781–788 (1999)
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)