Analysis of KITTI Data for Stereo Analysis with Stereo Confidence Measures

Ralf Haeusler and Reinhard Klette

The University of Auckland
Computer Science Department
Tamaki Campus

Abstract. The recently published KITTI stereo dataset provides a new quality of stereo imagery with partial ground truth for benchmarking stereo matchers. Our aim is to test the value of stereo confidence measures (e.g. a left-right consistency check of disparity maps, or an analysis of the slope of a local interpolation of the cost function at the taken minimum) when applied to recorded datasets, such as published with KITTI. We choose popular measures as available in the stereo-analysis literature, and discuss a naive combination of these. Evaluations are carried out using a sparsification strategy. While the best single confidence measure proved to be the right-left consistency check for high disparity map densities, the best overall performance is achieved with the proposed naive measure combination. We argue that there is still demand for more challenging datasets and more comprehensive ground truth.

1 Introduction

Computational stereo vision is in general an ill-posed problem, and solutions are approximate in some sense. Situations with no unique solutions are commonly found in recorded stereo images. Also with strong assumptions imposed to resulting disparity maps, such as piecewise smoothness, no stereo matcher can guarantee correct results.

Widely used test datasets often use input data of limited complexity [1]. On these, recently developed stereo matchers deliver dense results of satisfactory quality. However, in practice, recorded real-world data are considerably more challenging; see, e.g., [2,3]. There may be no satisfactory solution to the stereo problem in many situations displayed in these datasets. For treating unmatchable regions in disparity maps, there are some attempts to identify these using so-called confidence measures.

To evaluate quality of stereo matching results and accuracy of detected mismatches by means of confidence measures, availability of accurate ground truth is necessary. However, real-world stereo data in [2,3] are provided without ground truth. Recently published stereo-vision test data [4], in the paper briefly called *KITTI data*, show real-world scenes and come with ground truth provided by a laser range-scanner. In this paper, we try to rate the challenge given by KITTI data and quantify the value of confidence measures in dealing with such datasets.

A. Fusiello et al. (Eds.): ECCV 2012 Ws/Demos, Part II, LNCS 7584, pp. 158–167, 2012.

Confidence measures can be used in stereo processing to remove spurious disparity matches and interpolate these locations with surrounding disparity values. The left-right consistency check is the most prominent example. Various confidence measures have been proposed for stereo analysis; see [5,6]. Experimental studies of quantitative evaluations of confidence measures have been conducted (on very limited datasets) in [6]. By using the KITTI data, we also discuss confidence measures on a larger set of real-world data.

Many confidence measures are derived from the graph of cost functions. For example, consider a parabola fit of the cost function at the taken minimum and its neighbourhood. The minimum of this parabola can be used to interpolate disparity with subpixel accuracy. Parameter a defines the curvature of such a parabola $ax^2 + bx + c$, and this value is often considered to be a confidence measure. In [7] it is applied with the intention to improve results for scene flow computations. An equivalent confidence measure can be defined using the slope of an Okutomi fit [8], and this may be more appropriate, depending on the stereo matching cost function used.

However, a confidence measure that only operates locally on the cost function cannot assign, for example, low confidence to ambiguous matches within repetitive textures of the source image. Improved results can be expected when the costs of all disparity values are taken into account. This was demonstrated by [9] for the 3D reconstruction of large outdoor scenes where confidence is high if the cost function has a single well-defined minimum, and low if the cost function is flat or has multiple strong local minima. It is an open question which confidence measure has the best performance in a particular matching situation.

Very little has been published on improving the accuracy of confidence measures by combining different measures. For example, see [10] for confidence measures in optical flow computations, using a supervised learning process of Gaussian mixture models with classification criteria being the magnitude of the end point error in optic flow. However, it remained open whether this confidence measure (obtained as a result of classification in feature space) could actually improve accuracy. In [11], flow data is assigned to algorithms found most suitable by classification, based on a combination of flow confidence features such as photo consistency and texture properties. [12] used similar features for detecting difficult to match flow regions in order to test whether these regions are occluded. In both, classification results are obtained using random decision forests.

In this paper we include a naive approach for confidence measure combinations for stereo, which seems to be useful in increasing accuracy, i.e. is reducing numbers of false positives. Also, we define a simple bound for limits of potential accuracy improvements by measure combinations. We apply this measure and a number of popular stereo confidence measures to the recently published KITTI dataset [4], and evaluate them using a similar sparsification strategy as used in [6].

Section 2 contains the definitions of confidence measures used in this paper, including a multiplicative combination and an upper bound for accuracy improvements. Section 3 presents the evaluation method for the performance of

measures and provides details about experiments conducted. Section 4 presents results. Section 5 discusses limitations of confidence measures and of the used KITTI dataset. Section 6 concludes.

2 Confidence Measures for Stereo Analysis

For our experiments, we use *semi-global matching* (SGM) for stereo analysis [13] due to its best overall performance on synthetic and recorded outdoor data in terms of quality and computational costs. The data term in the original version of SGM stereo is derived from a mutual information measure [14]. However, [15] concludes that the census cost function leads to the best overall performance. Confidence measures used in this study are not defined on census costs directly, but on census costs aggregated using SGM with penalty parameters $p_1 = 20$ and $p_2 = 100$. In the following, we denote these aggregated cost A.

A compilation of different confidence measures can be found in [5,6]. We selected measures which appear to be pairwise "fairly different" by their definitions, with an intention to obtain the best possible performance boost by combining "orthogonal measures". This is, of course, possible only to a limited extent. All confidence measures presented here are defined pixelwise. No dependency from neighbouring pixels is introduced, except the dependency defined by cost aggregation. Contrary to the common practice, we scale values such that lower values indicate a smaller likelihood of an erroneous disparity estimate (i.e. a higher confidence).This facilitates the comparison to actual disparity errors. We define and use the following measures, with d_0 representing the estimated disparities $D(x, y)$:

- *Curvature.* Curvature derived from parabola fitting is a very popular stereo confidence measure. It is the only confidence measure in the source code of [1]. We are using the inverse of the opening parameter for scaling:

$$\Gamma_0(x,y) = \frac{1}{-2A_{(x,y)}(d_0) + A_{(x,y)}(d_0-1) + A_{(x,y)}(d_0+1)} \tag{1}$$

- *Perturbation.* The perturbation measure, introduced 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:

$$\Gamma_1(x,y) = \sum_{d \neq d_0} e^{-\frac{\left(A_{(x,y)}(d_0) - A_{(x,y)}(d)\right)^2}{\varsigma^2}} \tag{2}$$

  The parameter ς should be chosen such that no numerical underflows appear. A good choice depends on the range of possible valid cost values.
- *Peak ratio.* The peak ratio indicates low confidence if there are two candidates with similar matching score. It is defined as

$$\Gamma_2(x,y) = A_{(x,y)}(d_0)/A_{(x,y)}(d_1) \tag{3}$$

where d_1 is the second smallest local minimum, i.e. $A_{(x,y)}(d_1-1) > A_{(x,y)}(d_1)$ and $A_{(x,y)}(d_1+1) > A_{(x,y)}(d_1)$. Not requiring a local minimum but only the second smallest cost value yields a measure similar to parabola curvature.

– *Entropy.* Entropy defined on the disparity space was used in steering a diffusion process for cost aggregation in stereo matching [16]. It can be used as a confidence measure itself. Cost values are normalized for computing the entropy. Here, p indicates that cost values are converted into a probability distribution function:

$$\Gamma_3(x,y) = -\sum_d p(d)\log p(d) \quad \text{with} \quad p(d) = \frac{e^{-A_{(x,y)}(d)}}{\sum_{d'} e^{-A_{(x,y)}(d')}} \tag{4}$$

Similar to the perturbation measure Γ_1, cost values for all disparities influence the result. All cost values undergo nonlinear scaling, making these measures computationally rather expensive.

– *Right-left consistency check.* Right-left consistency compares the disparities of the left and right view, denoted by D and $D^{(r)}$:

$$\Gamma_4(x,y) = \left| D_{(x,y)} - D^{(r)}_{(x,y)} \right| \tag{5}$$

Large discrepancies, intuitively, should indicate a higher likelihood of an incorrect match. We are also interested in finding out whether there is confidence information extractable from disparities given with subpixel precision, i.e. for $\left| D_{(x,y)} - D^{(r)}_{(x,y)}) \right| < 1$.

– *Combination of measures.* We define a measure combination Γ by multiplying all five measures defined above:

$$\Gamma(x,y) = \Gamma_0(x,y)\Gamma_1(x,y)\Gamma_2(x,y)\Gamma_3(x,y)\Gamma_4(x,y) \tag{6}$$

Note that this is not justified in the sense of joint probabilities. Measures are not pairwise independent. Correlation is quite strong as their definition is derived from same cost values A.

– *Accuracy.* Ideally confidence measures exactly predict the magnitude of the error in disparity estimation, so as to discard largest errors first. For comparison with $\Gamma_0, \ldots, \Gamma_4$ and Γ we include this ideal measure, given by the absolute difference Δ between disparity estimate D and ground truth G:

$$\Delta(x,y) = \left| D_{(x,y)} - G_{(x,y)} \right| \tag{7}$$

In practise, finding such a confidence measure was equivalent to the stereo problem being solved.

Thus, it is more realistic to compare a combination of confidence measures rather against a measure that identifies an error if at least one of the contributing confidence measures can detect it. The amount of remaining errors s is then the cardinality of the union set of all undetected erroneous pixels for a specific sparsification ϕ of the disparity map D:

$$s(\phi) = \mathrm{card}\left\{ D_{(x,y)} | (x,y) \in \Omega \wedge \Delta(x,y) > 3 \wedge \forall i \left(\Gamma_i(x,y) \leq \gamma_i(\phi)\right) \right\} \tag{8}$$

Here, $\gamma_i(\phi)$ is the value of the confidence measure Γ_i which induces a sparsification ϕ. Note, that values from this definition cannot be derived without ground truth. It merely gives a lower limit of remaining disparity errors. The bad-pixel criterion is deviation of estimated disparities from ground truth being larger than 3. This is in accordance with the default in [4] and due to limited accuracy of laser range finder measurements.

3 Evaluation of Confidence Measures

We test SGM stereo and derived confidence measures on a small subset of the recently introduced KITTI Dataset [4]. We choose stereo frames which expose some interesting behaviour, e.g. the ones challenging for SGM stereo. Challenging are considered the frames which produce most errors in comparison to ground truth. We also include two frames where only few errors between SGM results and ground truth can be observed.

For comparisons, we use raw disparities D, i.e. do not attempt to improve results by consistency checks, interpolation, median filtering, or any other kind of postprocessing. Any such steps may introduce some bias to findings. However, we exclude such disparities from evaluation, which have been found to be occluded according to the method described in [13]. It may be beneficial, to use occlusion ground truth instead, if such was available with [4].

Furthermore, we restrict the evaluation to pixels where:

- Ground truth is available. In KITTI data there is no ground truth for upper parts of images, areas usually depicting sky in road vehicle based recordings.
- The cost function contains valid values for all possible disparities. This excludes the left image border in the left view with its width equalling the number of disparities, the reason being scaling issues in the presense of unknown cost values in the definition of confidence measures based on the entire cost function.
- Disparity estimates are supported by data term values, i.e. we do not rely on extrapolation introduced by SGM. This excludes a narrow band along the image border (e.g. 3 pixels for a 7×7 cost matching window).

A sparsification function is computed by summing the remaining errors of matching results when a certain percentile of the disparity map is filtered according to the locations having least confidence assigned. Locations with lowest confidence are removed first.

Measure comparisons often refer to the area under the sparsification curve. However, errors for very sparse maps are of little interest and should not bias a comparison of confidence measures. Thus, we do not compare the areas under the sparsification curves. We are merely interested in measure performance for disparity map densities in a range corresponding to proportions of good disparity pixels, typically between 70 % and 95 % when using the SGM stereo algorithm without postprocessing disparities on the KITTI dataset (see intersection of ground truth optimum with x-axis in Figures 2 and 3).

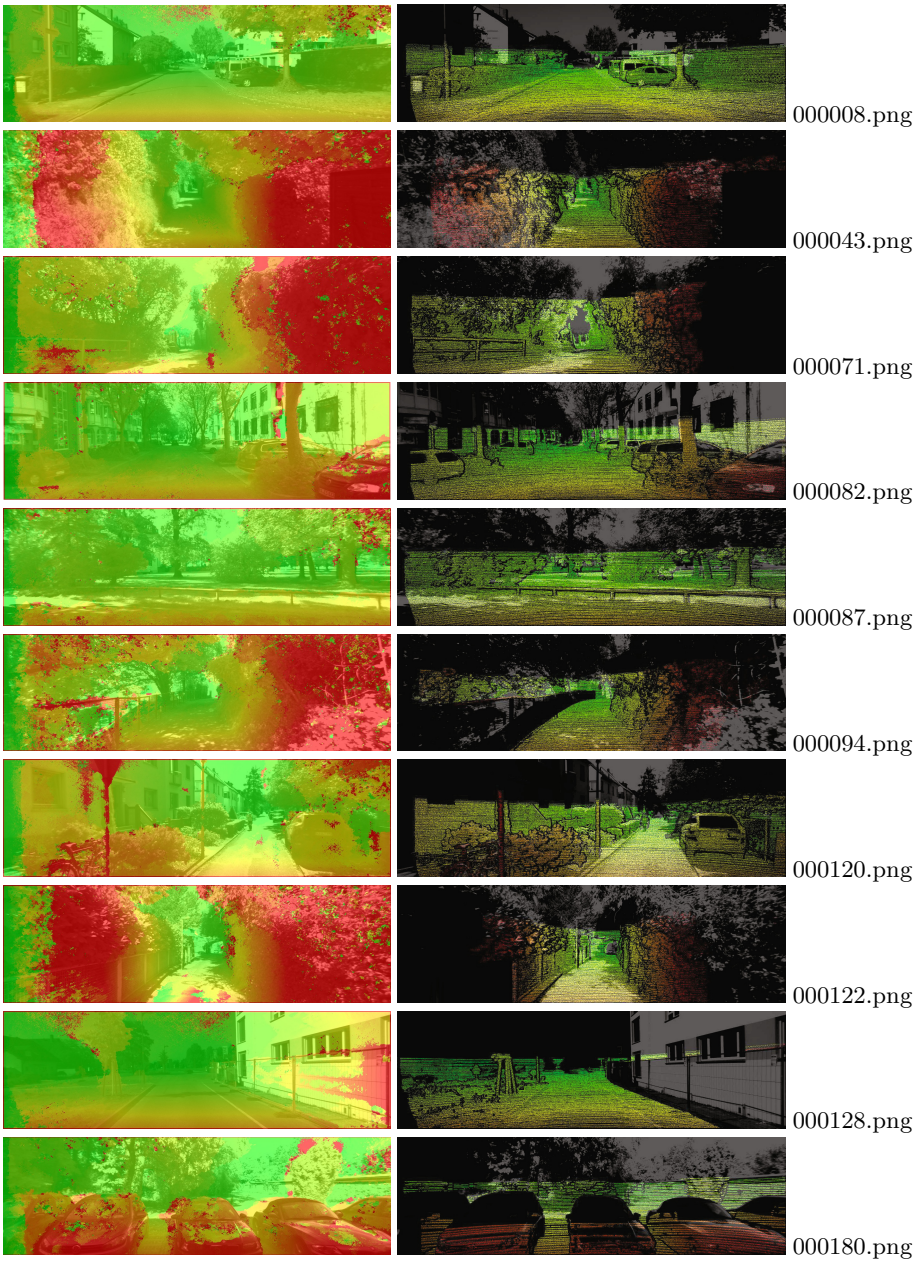

Fig. 1. Visualization of stereo results without postprocessing (left column) and laser range finder ground truth (right column) of frames included in this study. Close objects are red, distant objects green.

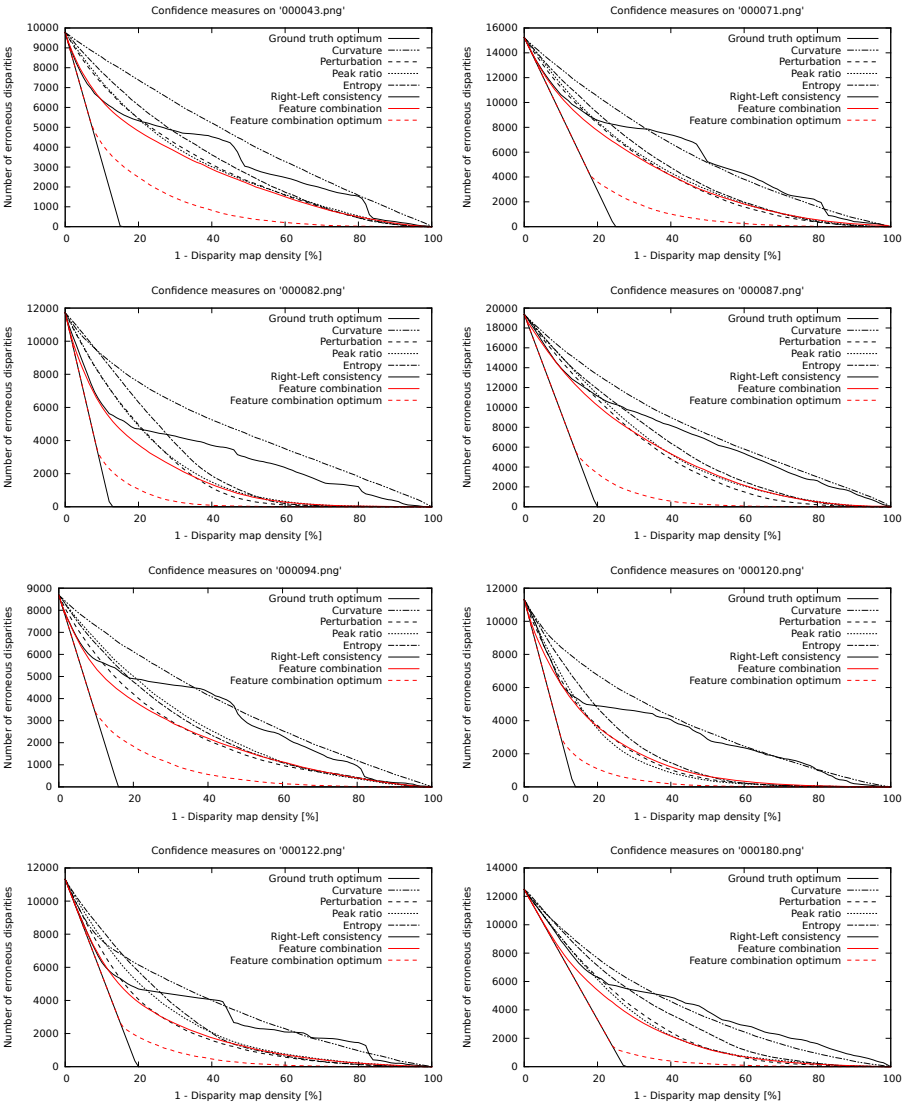

Fig. 2. Sparsification plots for selected frames of KITTI data with high amount of bad pixels in the evaluation domain excluding locations detected as being occluded by the stereo matcher. Numbers of remaining bad pixels are plotted against disparity map density.

4 Results

Recall that Frames 8 and 128 were selected due to their apparent amenability to the stereo problem. The illustration in Fig. 1, however, reveals that low error rates are only a result of missing ground truth in locations where stereo solutions

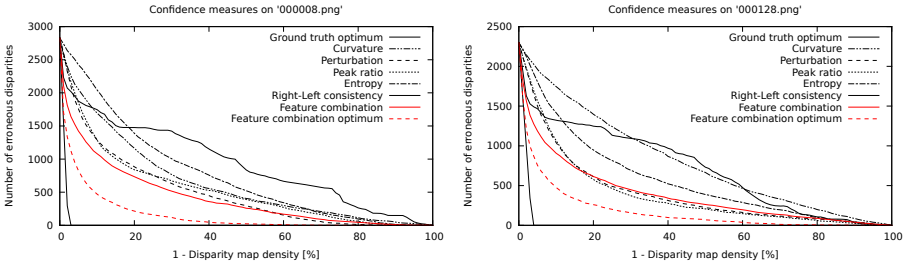

Fig. 3. Sparsification plots for selected frames of KITTI data with low amount of bad pixels in the evaluation domain excluding locations detected as being occluded by the stereo matcher. Numbers of remaining bad pixels are plotted against disparity map density.

are not accurate. Nevertheless, this may not corrupt an evaluation of confidence measures, as these areas are ignored.

Sparsification plots in Figures 2 and 3 show that the curvature confidence measure Γ_0 in general identifies disparity errors much less accurately than other confidence measures included in this study. In Frame 43 the selection of bad pixels appears to be almost random (straight line). However on Frame 43, depicting some vegetation in close proximity, other confidence measures also do not seem to work well. Useful accuracy of Γ_0, according to our experiments, can be observed in Frames 8 and 122.

The remainder of confidence measures, regarding accuracy, is almost on a par at low sparsifications. From the measures Γ_1 to Γ_4, the entropy confidence measure Γ_3 is most often the worst one. The left-right consistency check is not helpful within higher sparsifications, especially not if $\Gamma_4 < 1$.

For most frames, the naive combination Γ of confidence measures outperforms single confidence measures at a sparsification determined by numbers of bad pixels. However, the left-right consistency check Γ_4 can be on a par at this sparsification level.

Regarding the proposed bound for combined confidence measures it can be found that at the level of sparsification necessary to remove all bad pixels, there is usually a remainder of half the amount of bad pixels found by the proposed combination Γ, which cannot be identified by any confidence measure.

5 Discussion

In the following, we discuss accuracy of confidence measures on the frames included in this study and try, where possible, to link their behaviour to scene structure and quality of ground truth and stereo results.

The curvature confidence measure, although widely used, is not much better than randomly removing pixels. Reasons may include limited sharpness of

recorded images. This measure only performs rather well on Frames 8 and 122. This finding may not be significant for Frame 8, due to low numbers of bad pixels involved. The reason for apparently improved accuracy of curvature on Frame 122 may be the detection of homogeneously textured areas on road surfaces, which are wrongly estimated by SGM stereo. For detection of textureless areas, however, there may exist better indicators, such as the structural tensor [17].

That, in all frames, the consistency check Γ_4 does not perform well at higher sparsifications is not surprising, as subpixel estimation only is based on interpolated cost function values and therefore compromised by so-called pixel locking. However, the consistency check is also not successful in detecting errors resulting from surface over-extension or foreground fattening [18], as these occur likewise in left-right and right-left matching.

Frame 87 visually does not seem to contain 20 percent bad pixels showing in Fig. 2. Also, confidence measures do not detect these bad pixels well. Such results pose the question whether a fixed tolerance of 3 pixel for disparity errors is appropriate. In particular, in practise, it may be advantageous to define error tolerance based on physical distance and with a lower limit in disparity space.

Another problem of the KITTI dataset is missing ground truth in large image areas. It is most often absent where image regions are challenging for stereo. This is, e.g., the case on fences (Frames 94 and 128) and car windows (Frames 82, 120 and 180). However, it should be noted that there is no agreement or even no approach regarding evaluation strategies for such image regions exposing stereo matchers to semi-transparencies.

6 Conclusion

Compared to previous benchmark data which was recorded under controlled lighting conditions, some frames of KITTI data are more challenging for computational stereo due to outdoor scenes with suboptimal lighting conditions. However, extremely difficult recordings of road scenarios are absent. To our experience, these include tunnels, back-light, night scenes with street lights and lights of other traffic participants, wet road surfaces and precipitation.

Missing ground truth at locations challenging for stereo vision reduces the significance of our findings and makes it hard to draw comprehensive conclusions.

A comprehensive analysis, however, can be conducted if sufficiently diversified datasets with ground truth are available. Apart from the low coverage with ground truth at interesting image locations, another shortcoming of the KITTI dataset is the low dynamic range and luminance resolution of the imagery. Being eight bit only, preventable stereo problems are introduced, which may disguise more interesting problems. We conclude that a sufficiently challenging and rich dataset for stereo benchmarking with ground truth is not yet available.

References

1. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International Journal of Computer Vision 47, 7–42 (2002)
2. The University of Auckland: .enpeda.. image sequence analysis test site (EISATS), http://www.mi.auckland.ac.nz/EISATS
3. Heidelberg Collaboratory for Image Processing: Robust Vision Challenge, http://hci.iwr.uni-heidelberg.de/Static/challenge2012/
4. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The KITTI vision benchmark suite. In: Computer Vision and Pattern Recognition (CVPR), Providence, USA (2012)
5. Banks, J., Corke, P.I.: Quantitative evaluation of matching methods and validity measures for stereo vision. I. J. Robotic Res. 20, 512–532 (2001)
6. Hu, X., Mordohai, P.: Evaluation of stereo confidence indoors and outdoors. In: [19], pp. 1466–1473
7. Wedel, A., Brox, T., Vaudrey, T., Rabe, C., Franke, U., Cremers, D.: Stereoscopic scene flow computation for 3D motion understanding. International Journal of Computer Vision 95, 29–51 (2011)
8. Shimizu, M., Okutomi, M.: Precise sub-pixel estimation on area-based matching. In: ICCV, pp. 90–97 (2001)
9. Merrell, P., Akbarzadeh, A., Wang, L., Mordohai, P., Frahm, J.-M., Yang, R., Nistér, D., Pollefeys, M.: Real-time visibility-based fusion of depth maps. In: ICCV, pp. 1–8 (2007)
10. Gehrig, S.K., Scharwächter, T.: A real-time multi-cue framework for determining optical flow confidence. In: ICCV Workshops, pp. 1978–1985 (2011)
11. Aodha, O.M., Brostow, G.J., Pollefeys, M.: Segmenting video into classes of algorithm-suitability. In: [19], pp. 1054–1061
12. Humayun, A., Aodha, O.M., Brostow, G.J.: Learning to find occlusion regions. In: CVPR, pp. 2161–2168. IEEE (2011)
13. Hirschmüller, H.: Stereo processing by semiglobal matching and mutual information. IEEE Trans. Pattern Anal. Mach. Intell. 30, 328–341 (2008)
14. Egnal, G.: Mutual information as a stereo correspondence measure. Computer and Information Science, University of Pennsylvania, Philadelphia, USA, Tech. Rep. MS-CIS-00-20 (2000)
15. Hirschmüller, H., Scharstein, D.: Evaluation of cost functions for stereo matching. In: CVPR (2007)
16. Scharstein, D., Szeliski, R.: Stereo matching with nonlinear diffusion. International Journal of Computer Vision 28, 155–174 (1998)
17. Shi, J., Tomasi, C.: Good features to track. In: CVPR, pp. 593–600. IEEE (1994)
18. Okutomi, M., Katayama, Y., Oka, S.: A simple stereo algorithm to recover precise object boundaries and smooth surfaces. International Journal of Computer Vision 47, 261–273 (2002)
19. The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, June 13-18. IEEE (2010)