Ground Truth Evaluation of Stereo Algorithms for Real World Applications

Sandino Morales and Reinhard Klette

.enpeda.. group, Dept. Computer Science, University of Auckland, New Zealand

Abstract. Current stereo algorithms are capable to calculate accurate (as defined, e.g., by needs in vision-based driver assistance) dense disparity maps in real time. They have become the source of three-dimensional data for several indoor and outdoor applications. However, ground truth-based evaluation of such algorithms has been typically limited to data sets generated indoors in laboratories. In this paper we present a new approach to evaluate stereo algorithms using ground-truth over real world data sets. Ground truth is generated using range measurements acquired with a high-end laser range-finder. For evaluating as many points as possible in a given disparity map, we use two evaluation approaches: A direct comparison for those pixels with available range data, and a confidence measure for the remaining pixels.

Keywords: Performance evaluation, stereo algorithms, laser range finder

1 Introduction

Vision-based stereo algorithms are designed to generate three-dimensional (3D) information from two-dimensional (2D) data recorded with two or more video cameras. State-of-the-art stereo algorithms are capable to perform in real-time "accurate" disparities for almost all the points visible in the input images. Current applications for stereo algorithms, among many others, are vehicle navigation (robots [17], forklifts [21], wheelchairs [20], and so forth) or industrial safety equipment.[1]

We are interested in the evaluation of stereo algorithms in the context of vision-based Driver Assistance Systems (DAS) [11] for improving those techniques. DAS requires that the detection of depth is accurate on every road, under all kinds of weather conditions, and in any traffic context. Therefore, stereo algorithms need to be evaluated in the real-world, and not only on data representing a few seconds of recording but hours or days.

The evaluation of stereo algorithms is either based on ground truth data, allowing direct comparisons between true disparity values and those obtained with the algorithms; or it is performed in the absence of ground truth using various ideas for still ensuring some kind of objective testing. For real-world

[1] `http://www.pilz.com/products/sensors/camera/f/safetyeye/`

video data it is the ultimate goal to provide ground truth as well. Synthetic (i.e., computer generated stereo pairs) or engineered (i.e., images captured under highly controlled conditions, using structured light for generating ground truth) data do have their own characteristics [8], and do not cover the "challenges" as occurring in real-world data.

Real-world data do not come (typically) with ground truth. Therefore, diverse methods have been proposed to evaluate the algorithms even in absence of ground truth. In [1], the evaluation was done by measuring the number of successfully matched pixels using a left-right consistency check [9]. Some authors used an extra image (e.g., prediction error in [23]) or a third video sequence (see the third view in [15]) as ground truth. Confidence measures are another example of evaluation in the absence of ground truth [6, 16]. The idea is to measure the reliability of the calculated disparity value for each pixel. Techniques, specifically designed for DAS, were proposed in [14, 22]; these evaluation schemes can only be applied if some conditions are satisfied in the recorded scenes.

We generate ground truth using precise depth measurements acquired with a laser range-finder (LRF). The generation of ground truth (or of accurate 3D models) using LRF's has been investigated before [2, 10, 17]. However, those publications do not report about the evaluation of stereo algorithms using laser range data. Stereo algorithms are discussed together with laser range data in [19] at selected feature areas.

The evaluation scheme in this paper analyzes stereo algorithms on recorded video sequences based on available 'sparse' (but uniformly distributed) ground truth and also applying a confidence measure for dealing with the 'gaps'. We use Velodyne's HDL-64E S2 range-finder [24]. For the distance interval of interest (about 5 to 120 m), the available accuracy is defined by possible errors of less than 10 cm (the producer even sees the error at 1.5 cm at most in 5 to 120 m).

The obtained range data are insufficient for evaluating an entire dense disparity map, e.g. a VGA image has $640 \times 480 = 307,200$ pixels, and the used LRF generates up to 24,000 points in the field of view of the reference camera in our stereo set up; see Fig. 1. Thus, we combine two approaches for the evaluation.

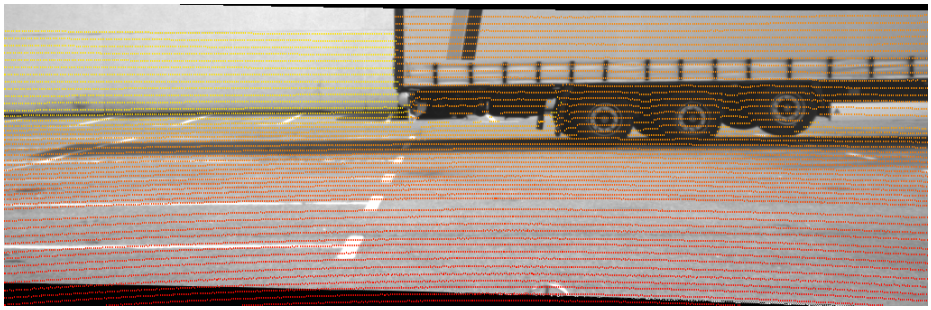

Fig. 1. Sample image showing combined laser range-finder and image data. Ground truth points (i.e., points acquired with the laser range-finder) are color encoded from red (for close) to green (for further away).

If ground truth data are available at a specific pixel, we perform a direct comparison between the calculated disparity value and the ground truth. For the remaining points, we use a geometrical approach using "close" range readings to generate a confidence measure. This approach allows us to evaluate stereo algorithms for outdoor real-world data based on true measurements. Data sets can be recorded in all kinds of weather where the LRF will work in, or road conditions.

The main contributions of this paper are the measures proposed to evaluate dense algorithms against sparse (less than 10%) ground truth. The data provided contain sub-pixel accurate ground truth for real-world scenes, and this was not available prior to the use of a laser range-finder. This data set has been made publicly available for future research considerations, see [4].

The structure of this paper is as follows. In Section 2 we present the proposed approach. We continue with experiments in Section 3, and finalize with conclusions in Section 4.

2 Approach

We generate sparse ground truth disparity maps with the LRF, and perform the evaluation by fusing a direct comparison approach (where true values from the LRF were available) and a confidence measure (for the remaining points). See Fig. 2 for a flow chart of the proposed approach.

Ground Truth Disparity Map Generation. We record range data of the surrounding environment of the *ego-vehicle* (i.e. the vehicle carrying the stereo camera and the LRF) using a high-end LRF [24]. The provided accuracy data (precision of 1·5 centimeters within a range from one to 120 metres) needs to be slightly corrected, and 10 cm can be used as an upper bound in our experiments.

The rotational architecture of the LRF allows us to obtain readings from 64 lasers in a full 360° rotation. Its optimum resolution (depending on the rotational speed) is of 0·09° (horizontal) times 0·4° (vertical). The vertical field of view of 26·8° provides sufficient information for modeling the road and the objects that would be of interest in a driving scene.

Assume for now that the coordinate systems of the LRF and the stereo camera have been calibrated and aligned. Then, we are able to project the output of the LRF (a set of 3D points) onto a 2D image G using the (internal) parameters of the stereo camera. The ground truth disparity value $G(\mathbf{x})$ of a pixel $\mathbf{x} \in G$ is

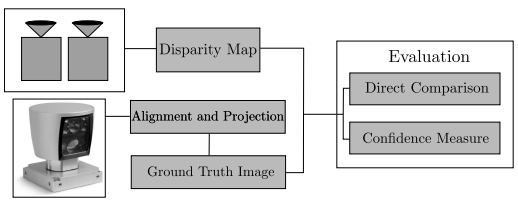

Fig. 2. Flow chart of the used approach.

defined by

$$G(\mathbf{x}) = \frac{f \cdot b}{Z(\mathbf{x})} \tag{1}$$

where f denotes the focal length of the stereo camera, Z the distance from the camera (at pixel $\mathbf{x}$) in the depth direction, and b is the distance between the optical centers of the cameras (the length of the baseline). For pixels where there is no distance measure available, a distinctive negative value is assigned (as disparity values are strictly positive). For the images that we use for our experiments (i.e., $1024 \times 334 \approx 342,000$ pixels), we are able to obtain ground truth values for almost 7% (about 24,000) of the pixels. These are the only points we are able to perform a direct comparison.

In the context of DAS, the final goal is to analyze the performance of stereo (or any) algorithms in outdoor dynamic environments. Thus, it does not make sense to scan the same scene multiple times to get more range readings. Instead, we use the available measurements to generate a confidence measure to evaluate the remaining points.

Direct Comparison. Where range data is available we use the percentage of badly calculated pixels (BCP) as quality metric. Let D be a disparity map obtained with a given stereo algorithm, and G the generated ground truth image. Let Ω denote the set of pixels in G and D such that $G(\mathbf{x}) > 0$ (i.e., pixels with a valid measurement from the LRF) and $D(\mathbf{x}) > 0$ (i.e. pixels with invalid disparities were also identified with a negative value). Let T be a predefined tolerance threshold, and

$$\delta(\mathbf{x}) = \begin{cases} 1, & \text{if } |G(\mathbf{x}) - D(\mathbf{x})| \geq T \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

Then, the BCP of D is as follows:

$$B = \frac{100\%}{|\Omega|} \sum_{\mathbf{x} \in \mathbf{\Omega}} \delta(\mathbf{x}) \tag{3}$$

where $|\cdot|$ denotes the cardinality of a set.

Confidence Measure. To complement the direct comparison (i.e., to evaluate also points where no range data are available), we use a simplified version of the approach presented in [3]. In that paper, the authors used a probabilistic scheme to deal with non organized point clouds generated by a LRF of small objects under controlled conditions (i.e., indoor scenes).

Given three "close" pixels in the ground truth image G, we define a patch $P_G \subset G$ and its 3D version $\overline{P_G}$ by back projecting the three pixels into the 3D space. Using the corresponding pixels in the disparity map D, we generate the respective patches P_D and $\overline{P_D}$. The evaluation is then made by comparing the geometric properties of the 3D patches.

The selection of the three "close" pixels is as follows. Given a pixel $\mathbf{x} \in G \cap \Omega$, its closest neighbors are the points generated by the same laser beam $L_{\mathbf{x}}$ (recall that the horizontal resolution of the LRF is $0{\cdot}09°$) in the previous or in the

next shot, followed by the points generated by one laser beam below or above $L_{\mathbf{x}}$ (there are 64 lasers in the LRF). Thus, we choose to generate the patches P_G using two pixels from the same laser and one either from the laser above or below (creating a triangle). A patch is only be defined if the disparity value of all the selected pixels is within a predefined range. If the selected pixels are also elements of $\Omega \cap D$, we generate the corresponding patch P_D. This patch also contains the pixels in D within the triangle defined by the three selected pixels. Once both patches have been defined, we analyze the geometric properties of their respective back projections (i.e 3D sets), $\overline{P_G}$ and $\overline{P_D}$.

Let $\overline{P} \subset \mathbb{R}^3$ be one of this patches, the *centroid*

$$c\left(\overline{P}\right) = \frac{1}{\left|\overline{P}\right|} \sum_{\overline{\mathbf{x}} \in \overline{P}} \overline{\mathbf{x}} \tag{4}$$

is calculated, as well as the *deviation* of the points in $\overline{P}$ with respect to $c(\overline{P})$:

$$\mathrm{Dev}\left(\overline{P}\right) = \sqrt{\frac{1}{\left|\overline{P}\right| - 1} \sum_{\overline{\mathbf{x}} \in \overline{P}} \left(\overline{\mathbf{x}} - c\left(\overline{P}\right)\right)^2} \tag{5}$$

Note that $\overline{\mathbf{x}} \in \mathbb{R}^3$. Now, let P_G and P_D be corresponding patches in G and D, respectively. The confidence measure is calculated based in the distance between the centroid of the back projected patches, $\overline{P_G}$ and $\overline{P_D}$, and the ratio of their respective deviations. Let Δ_P be the Euclidean distance between $c\left(\overline{P_G}\right)$ and $c\left(\overline{P_D}\right)$, and

$$\rho = \frac{\mathrm{Dev}\left(\overline{P_G}\right)}{\mathrm{Dev}\left(\overline{P_D}\right)} \tag{6}$$

Then, the confidence measure index for P_D is calculated as

$$CM(P_D) = \frac{2\rho}{\rho^2 + 1}\left(1 - \frac{\Delta_P}{\Delta_{\max}}\right) \tag{7}$$

where $\Delta_{\max}$ is the maximum possible Euclidean distance between the centroids.

The range of CM is $[0, 1] \subset \mathbb{R}$; where a value close to one indicates that both patches are geometrically alike, and thus that the disparity results are reliable. Low values imply a low confidence in the calculated disparity values. To obtain a high confidence value (i.e., a value close to one), it is necessary that the centroids of both patches are close to each other and the ratio ρ of the variances is close to one.

The first factor in Eq. (7) penalizes the index more if $\rho < 1$, as it is expected that $\overline{P_G}$ would be a more homogeneous set than $\overline{P_D}$. See Fig. 3 for an example of two pairs of analyzed patches in a sample 3D scene as viewed from above.

3 Experiments

Our experimental data set was captured using the LRF and two grey-scale (12 bits per pixel) cameras, all mounted in the same ego-vehicle. The cameras were

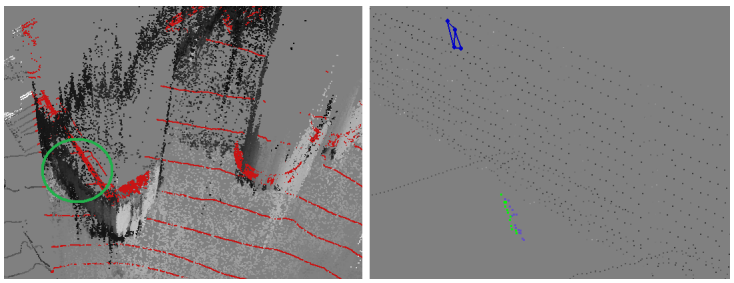

Fig. 3. A 3D test scene from the containers sequence (see Section 3) as viewed from above. *Left:* The red dots are the points returned with the LRF within the field of view of the reference camera. The grey points are the back projected points from a sample disparity map. *Right:* The gray points in here represent the LRF points. The projection of the highlighted blue points define two patches in the ground truth image; while the green and purple dots are the back projection of the two corresponding patches in the sample disparity map.

placed behind the windshield, while the LRF was attached to a rack on the roof. The coordinate system from the LRF was calibrated according to the external parameters of the reference camera (the left camera) of the stereo set up using the method proposed in [13], where a closed-form solution of the Perspective-n-Point problem was presented. We use the internal parameters of the stereo camera to project the 3D points from the LRF to generate the ground truth image.

For defining the patches, we use a disparity threshold of one, so that they were generated with points that are really close to each other. The threshold for the BCP quality metric was set to one.

Data Set. We illustrate the presented approach by using three sequences recorded in "simple" environments. The objective of using these sequences is to "grow" a first experience using this approach and to validate if there is a good correlation between the direct comparison and the confidence measure's indexes.

The size of the images is of 1024×334 pixels, reduced to 930×289 due to the rectification procedure for stereo analysis. Range data were recorded using the five revolutions per second configuration of the LRF, in order to obtain the maximum number of measurements (around 24,000 pixels with positive value in the ground truth image). All the sequences are stop-and-go ones, in order to minimize synchronization issues between the camera (set to 20 frames per second) and the LRF. Developing and approach to generate ground truth in dynamic scenes is out of the scope of this paper. Sample frames of each sequence are shown in Fig. 1 and 4.

Wall sequence. Recorded while driving towards a wall that covers the entire field of view of the cameras. In the lower right corner of the images there is a small car and a trailer. Both objects are only present in the first part of the sequence.

Wall-trailer sequence. Recorded while driving towards the same wall as in the wall sequence. In this case there is a trailer that covers almost half of the

reference and match image. This sequence turned out to be a good example for miscalculated disparities, as the trailer's cover has areas with no texture at all. There are also two areas below the trailer where it is possible to see the road behind the trailer.

Container sequence. In this scene two different kinds of containers are present, one with a square base and two with a circular one. There is also a small part of a building with an intensity that it is very similar to the intensity of the curved containers so we are expecting "not so good" results from the stereo algorithms for this sequence. There are also two staircases with thin handrails that even the LRF had problems to detect.

Results. The stereo algorithms used in this work are briefly identified below. We use a local standard *dynamic programming* (DP) stereo algorithm [18]. Two global algorithms: *belief propagation stereo* (BP), with a coarse-to-fine approach [5] and a quadratic cost function [7], and a graph cut (GC) [12] algorithm. Finally, a *semi-global matching* (SGM) approach with mutual information as the cost function [9] was also used.

The algorithms are tested with respect to pixel accuracy. But, the approach presented here, as well as the data set, are well suited to test sub-pixel accuracy disparities. We are not aware of an existing real-world data set that can evaluate the performance of sub pixel accurate algorithms. See Table 1 for a summary of the results for all algorithms and both sequences.

Wall sequence. For this sequence, we expect the disparity values to get better as the ego-vehicle approaches the wall. This is due to the inverse proportionality of distance to disparity, thus small errors in disparity have a large effect at large distances. The algorithms behave as expected with respect to the BCP index; the percentage of badly calculated pixels decreases as the ego-vehicle gets closer to the wall. SGM had a high peak among the last five frames, where it has a poor performance on the road area. For CM, an average of 14,500 patches were analyzed (so above 50% of of the points in the disparity maps were considered for the evaluation). For GC and SGM, the CM index showed a consistent behavior with BCP, even the same peak for SGM in the last five frames can be identified here. The DP algorithm showed a relatively constant CM index. But, there is a low peak in the last five frames (the same set of frames that made SGM have a high BCP peak). In these frames the disparities obtained for the wall are not as homogeneous as it is expected, this can be barley detected with the BCP index. For BP, the CM index decreases over the sequence, in contrast with the behavior of the BCP index, an this indicates that there are less miscalculated points (according to the low BCP score) but the miscalculations are larger.

Fig. 4. Sample images of the *wall* (left) and *container* (right) sequences. For a sample image of the *wall-trailer* sequence see Fig. 1.

	Wall				Wall-Trailer				Containers			
	CM			BCP	CM			BCP	CM			BCP
Alg.	Avg.	"> 0·9"	"< 0·5"		Avg.	"> 0·9"	"< 0·5"		Avg.	"> 0·9"	"< 0·5"	
BP	0·43	4·2	64·6	28·5	0·43	6·0	63·2	33·5	0·44	4·8	64·1	29·5
DP	0·49	7·1	56·8	14·8	0·47	10·1	56·5	22·1	0·51	13·9	54·1	16·3
GC	0·29	0·5	85·8	35·1	0·34	0·6	82·2	38·7	0·28	0·7	88·2	42·6
SGM	0·37	3·2	67·9	35·3	0·38	2·5	75·9	50·2	0·36	2·4	76·6	59·5

Table 1. Summarized results for the three sequences with both quality metrics. The results for the confidence measure (CM) are presented as the average over the entire sequence (first column), and the percentage of the number of patches with an index below 0·5 (second respective column), again over the entire sequence. For BCP is only shown the average percentage over each one of the sequences.

Wall-trailer sequence. As expected, most of the algorithms have problems with the trailer's cover, as it is almost textureless. With respect to BCP, the algorithms had a similar performance, showing the worst results at the end of the sequence, when the trailer occupied almost the half of the stereo images. The exception was SGM. SGM handles this area better than the other algorithms. Its BCP index showed an improvement on its performance in the last part of the sequence. However, this algorithm had a poor performance in the road area making it the worst performing algorithm.

The results for CM show a good correlation with BCP. The confidence index decreases for DP and BP as the trailer is getting closer to the cameras, but increases for SGM. The GC algorithms did not follow the same pattern as with BCP; the last frames are the ones with highest CM value (but still very low). This can be explained as in the first half of the sequence, where the two areas below the trailer are visible; as this sequence goes forward, one of these zones goes out of the field of view of the cameras. The GC algorithms had more trouble detecting those background zones than the other algorithms. This can be detected with the CM index. However, the BCP index keeps going higher indicating that the disparity maps are still affected by the trailer's cover, but that the accuracy of the disparity values are better. The average number of patches calculated for this sequence was 14,300 (almost 50% of the points).

Container sequence: While both staircases and the building on the right side are present in the stereo images, the results for all the metrics for DP, GC and SGM show a failure. They all have problems detecting the thin structures from the staircases and the almost equal intensities of the circular containers on the right and the building next to it.

The BP algorithm behaved differently, but consistently for the two metrics. Its best performance is on the first part of the sequence, and starts decreasing from frame five. It looks like it had more trouble than the others with an almost saturated background area that grows as the sequence goes forward.

The GC and SGM algorithms swap their ranking under different metrics, see Table 1. For BCP and SGM, the GC algorithm had a better performance than SGM. This does not represent a drawback for our approach as one metric counts the number of pixels that were miscalculated (BCP) while the other one focuses

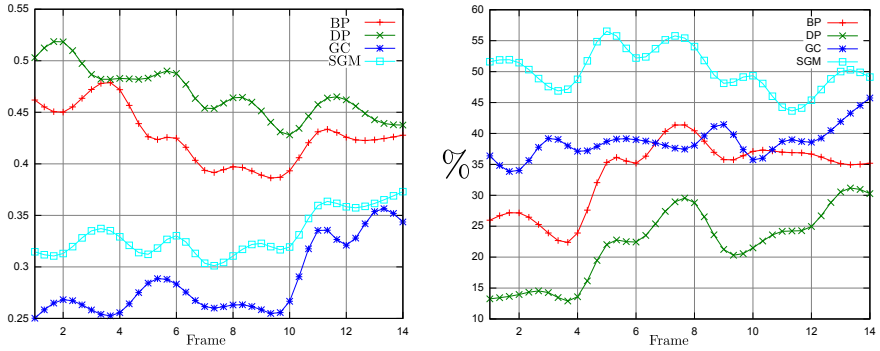

Fig. 5. Results for the wall-trailer sequence. *Left:* Results for CM, a value close to one indicates a high confidence in the disparity map. *Right:* Plot for the BCP, larger values implies a larger number in the miscalculated points.

on how accurate the disparity values are (CM). The average number of analyzed patches for this sequence was 14,600 implying that there were evaluated more than 50% of the pixels in the disparity map.

4 Conclusions and Future Work

In this work we present a ground truth-based approach to evaluate stereo algorithms over real-world sequences. We evaluate the algorithms by comparing the calculated disparity maps against ground truth images generated using a high-end LRF. As the ground truth images are not dense enough to evaluate all the pixels in the disparity maps, we follow two evaluation criteria: Where ground-truth data are available, we use a well-known quality metric to evaluate the corresponding disparity values. For the remaining points, we use a confidence measure that compares the geometric properties of corresponding point sets in the ground truth images and in the disparity maps. We also include a few experiments to show the effectiveness of the presented approach. In the experiments we noticed a good correlation between the measures used.

Using the direct comparison approach, we were capable to evaluate around 7% of the pixels in a disparity image. However, when we also use the confidence measure, we could evaluate the majority of the points. The exact number depends on the scene.

The obtained evaluation results need to be addressed in work aiming at improvements of stereo matching algorithms. We have a lot more experimental data, and those accumulated data will help further to identify weakness and strength of particular matching strategies, cost functions, or further algorithmic "ingredients" of stereo matching.

Acknowledgements: The first author thanks Dr. Uwe Franke for the opportunity to be a part of his research team at Daimler A.G. for six months, and Dr. Stefan Gehrig for valuable guidance.

References

1. Banks, J., Corke, P.: Quantitative evaluation of matching methods and validity measures for stereo vision. Int. J. Robotics Research **20** (2001) 512–532
2. Eid, A., Farag, A.: A unified framework for performance evaluation of 3-d reconstruction techniques. Proc. Comp. Vision Pattern Recognition Workshop **3** (2004) 33–41
3. Egnal, G., Mintz, M., Wildes, R.: A stereo confidence metric using single view imagery with comparison to five alternative approaches. Image Vision Computing **22** (2004) 943–957
4. *.enpeda..* Group, University of Auckland: EISATS (*.enpeda..* sequence analysis test site), `http://www.mi.auckland.ac.nz/EISATS` (2010)
5. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient belief propagation for early vision. Int. J. Comp. Vision **70** (2006) 41–54
6. Gherardi, R.: Confidence-based cost modulation for stereo matching. Proc. ICPR (2008) 1–4
7. Guan, S., Klette, R., Woo, Y.W.: Belief propagation for stereo analysis of night-vision sequences. Proc. PSIVT, LNCS 5414 (2009) 932–943
8. Haeusler, R., Klette, R.: Benchmarking stereo data (not the matching algorithms). Proc. DAGM, LNCS 6376 (2010) 383–392
9. Hirschmüller, H.: Accurate and efficient stereo processing by semi-global matching and mutual information. Proc. CVPR (2005) 807–814
10. Huang, F., Klette, R., Scheibe, K.: Panoramic Imaging: Sensor-Line Cameras and Laser Range-Finders. Wiley, Chichester (2008)
11. Klette, R., Vaudrey, T., Wiest, J., Haeusler, R., Jiang, R., Morales, S.: Current challenges in vision-based driver assistance. In: Progress in Combinat. Image Analysis, Research Publ. Services, Singapore (2010) 3–32
12. Kolmogorov, V., Zabih, R.: Multi-camera scene reconstruction via graph cuts. Proc. ECCV, LNCS 2352 (2002) 82–96
13. Lepetit, V., Moreno-Noguer, F., Fua, P.: Accurate non-iterative o(n)solution to the PNP problem. Proc. ICCV (2007) 874–885
14. Liu, Z., Klette, R.: Approximate ground truth for stereo and motion analysis on real-world sequences. Proc. PSIVT, LNCS 5414 (2009) 874–885
15. Morales, S., Klette, R.: A third eye for performance evaluation in stereo sequence analysis. Proc. CAIP, LNCS 5702 (2009) 1078–1086
16. Mordohai, P.: The self-aware matching measure for stereo. Proc. ICCV (2009) 1841 –1848
17. Murray, D., Little, J.J.: Using real-time stereo vision for mobile robot navigation. Aut. Robots **8** (2000) 161–171
18. Ohta, Y., Kanade, T.: Stereo by two-level dynamic programming. Proc. Int. Joint Conf. Artificial Int. (1985) 1120–1126
19. Reulke, R., Luber, A., Haberjahn, M., Piltz, B.: Validierung von mobilen Stereokamerasystemen in einem 3D-Testfeld. Proc. 3D-NordOst (2009)
20. Satoh, Y., Sakaue, K.: An omnidirectional stereo vision-based smart wheelchair. J. Image Video Processing **2007** (2007) 1–11
21. Seelinger, M., Yoder, J.D.: Automatic pallet engagement by a vision guided forklift. Proc. IEEE Int. Conf. Robotics Automation (2005) 4068–4073
22. Steingrube, P., Gehrig, S., Franke, U.: Performance evaluation of stereo algorithms for automotive applications. Proc. Int. Conf. Comp. Vision Systems (2009) 285–394
23. Szeliski, R.: Prediction error as a quality metric for motion and stereo. Proc. ICCV (1999) 781–788
24. Velodyne Lidar Inc.: Velodyne's HDL-64E S2 user manual - `http://www.velodyne.com/laserrangefinder/ManualList.aspx`