Differences Between Stereo and Motion Behaviour on Synthetic and Real-World Stereo Sequences

Tobi Vaudrey[1], Clemens Rabe[2], Reinhard Klette[1] and James Milburn[1]

[1] The *.enpeda..* Project, The University of Auckland, New Zealand
[2] Environment Perception Group, Daimler Research, Daimler AG, Germany

Abstract

Performance evaluation of stereo or motion analysis techniques is commonly done either on synthetic data where the ground truth can be calculated from ray-tracing principals, or on engineered data where ground truth is easy to estimate. Furthermore, these scenes are usually only shown in a very short sequence of images. This paper shows why synthetic scenes may not be the only testing criteria by giving evidence of conflicting results of disparity and optical flow estimation for real-world and synthetic testing. The data dealt with in this paper are images taken from a moving vehicle. Each real-world sequence contains 250 image pairs or more. Synthetic driver assistance scenes (with ground truth) are 100 or more image pairs. Particular emphasis is paid to the estimation and evaluation of scene flow on the synthetic stereo sequences. All image data used in this paper is made publicly available at `http://www.mi.auckland.ac.nz/EISATS`.

Keywords: performance evaluation, stereo analysis, optical flow analysis, synthetic vs. real-world image sequences, scene flow

1 Introduction

Many algorithms have been proposed and carefully studied for stereo and motion analysis; see, for example, [6, 8, 22] and the Middlebury website [2]. Performance evaluation is an important subject in computer vision [18]. Evaluation of stereo and optical flow algorithms are usually performed on computer rendered image pairs where ground truth is easily obtained using ray-tracing principles (e.g., [9, 12]). Evaluation can also be done on short *engineered* real-world scenes with labouriously obtained ground truth (e.g., [7]), and those real-world scenes are not relevant to industrial applications such as driver assistance. For the purposes of this paper, we group both engineered and rendered scenes under the label *synthetic scenes*. Image data in these evaluations only have 8-bit grey or 3×8-bit colour images. However, latest industrial cameras can obtain 10-bit (e.g., [4]) or 12-bit (e.g., [3]) grey-scale accuracy.

Driver assistance image sequences pose some of the most difficult challenges in current computer vision. The cameras are mounted on a moving platform, and the environment is not static; cars and pedestrians move independently of the static scene. This makes motion analysis and temporal stereo improvements very difficult. Stereo analysis has problems, due to low contrast, and optical flow algorithms have problems, due to large motion vectors on low-textured regions. These scenes are what we consider the most interesting for evaluating both stereo and optical flow algorithms.

After years of research on camera calibration and stereo rectification, the automotive industry has all the tools for producing rectified stereo image sequences. In Section 2, we will introduce seven 12-bit grey scale rectified image sequences, provided by Daimler AG Germany (Set 1 on [1]; see also [19]). We briefly highlight main features for each of those sequences, which define goals when analyzing those sequences. For downloads, see our website [1].

We also introduce a new synthetic driver assistance scene in Section 2. This fills one of the gaps in publicly available scenes, of a long stereo sequence with ground truth available. This scene is also made publicly available; see Set 2 on [1]. Currently, another set of three long sequences of rectified stereo colour images is made publicly available; see Set 3 on [1]. However, explanations for those data are not yet provided in this paper.

In Section 3 we identify why evaluations on synthetic data may not be the best option for stereo and optical flow estimations. Improvements on synthetic data may, in fact, not improve results

978-1-4244-2582-2/08/$25.00

on real world data. Section 4 describes *scene flow*, and this identifies another important class of algorithms that can be evaluated using a set of stereo sequences. We propose an evaluation approach and provide some results. The final section provides conclusions and future work.

2 Driver Assistance Scenes

In 2007, Daimler AG Germany provided the authors with seven stereo sequences for research purposes. They have been captured with a calibrated stereo pair of 12-bit grey-scale cameras near Stuttgart. These cameras have been optimised for infrared detection at night, but still obtain good quality images during daylight. Each sequence contains 250 or 300 pairs of frames, and features different driving environments; including highway, urban road and rural areas. Camera calibration is used for geometric rectification, such that image pairs are characterized by "standard epipolar geometry" [15]. Furthermore, the ego-motion of the stereo platform can be correctly estimated and compensated by using [5]. Figure 1 shows an example of one stereo pair in such a sequence.

The resolution of images is 640×481. They are saved in PGM (Portable Grey Map) format. Inertial sensor information is available in the image headers. – Here is a brief introduction for each sequence, including driving environment, main objects, or special features. More information can be found at the website [1] and in [19].

Construction-Site Sequence (Figure 1) features normal traffic density on an Autobahn, with a standard safety fence between opposite traffic directions. However, normal lanes have been cut in width and shifted to the right, with several slow large trucks in the right lane.

Safe-Turn Sequence (Figure 2(a)) features medium traffic in an urban area; the ego-vehicle goes straight and then turns to the left. There are a few pedestrians walking on the footpath, some cars in parking lots, or waiting to join the traffic.

Squirrel Sequence (Figure 2(b)) features a rural road with only oncoming traffic. Starting from the 156th frame, a squirrel appears in the scene and runs across the road, in front of the ego-vehicle.

Figure 1: *Construction-Site Sequence.*

(a) *Safe-Turn*

(b) *Squirrel*

(c) *Dancing-Light*

(d) *Intern-on-Bike*

(e) *Traffic-Light*

(f) *Crazy-Turn*

Figure 2: Selected frames from the image sequences.

The scene appears dark and wet. Vehicles in a distance can only be distinguished by head lights, with reflection on vehicles and road surface.

Dancing-Light Sequence (Figure 2(c)) shows a one-way road (two-lane highway) in a mountainous area. The ego-vehicle follows a small car and passes a few larger trucks. There are many shadows from the trees on the road and vehicles. Illumination changes significantly several times. The lighting between the stereo left and right images also varies.

Intern-on-Bike Sequence (Figure 2(d)) features a straight country road with both incoming and outgoing vehicle traffic. A cyclist drives toward the main road from a perpendicular direction and turns left at the intersection. This simulates a dangerous situation between the cyclist and the vehicle.

Traffic-Light Sequence (Figure 2(e)) has the ego-vehicle stopped in front of a traffic light, for the first half of the scene. It then drives into the opposite lane because of a construction site which blocks the other lane. At the end of the road construction area, another vehicle and a cyclist appear in the scene. The sequence features a road in a forest, with fine texture caused by trees.

Crazy-Turn Sequence (Figure 2(f)) has the ego-vehicle waiting on the roadside for the first 30 frames. It then goes into the central lane and turns left at the intersection while an oncoming vehicle is driving toward it. The sequence features a very dangerous situation where both vehicles almost collide.

These sequences, selected by Daimler AG, Ger-

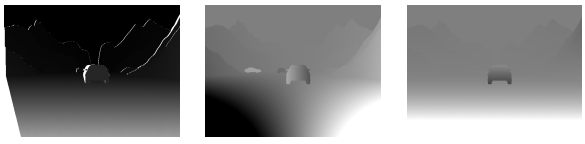
many, provide a multi-faceted challenge for stereo and motion analysis algorithms. These sequences will eventually have available some approximate ground truth. This is outside of the scope of this paper, but more information can be found in [20].

Synthetic sequences have been used for evaluation over the recent history. This is because the ground truth is easily obtained, and comparisons can be done using metrics such as Root Mean Square (RMS) for stereo, and Angular Error or End Point Error for optical flow. The Middlebury evaluation has done a good job at providing these type of sequences and evaluating the metrics to compare accuracy of different algorithms, for both stereo and optical flow.

For stereo and motion, these scenes have provided good bench-marking, allowing comparisons of algorithms for accuracy. These scenes however, are relatively easy for modern algorithms, and the latest algorithms are squeezing tiny improvements in sub-pixel accuracy from the images. The motion sequences are also short (8 frames), so filtering and integration techniques are seldom used. Furthermore, there is a very limited amount of stereo sequences available, so again very little work has been done with temporal integration for stereo, such as [23].

We have therefore made publicly available (Set 2 on [1]) a long synthetic stereo sequence. Our sequence is 100 image pairs with perfectly calibrated stereo cameras. The sequence is available in grey-scale or colour. Ground truth available is: optical flow, disparity (with occlusion map), and *scene flow* (see Section 4). Examples from the sequence can be seen in Figure 3.

(a) Sample Images



(b) Disparity (c) Flow X (d) Flow Y

Figure 3: New driver assistance synthetic stereo sequence, available publicly online [1]. Disparity colour encoding; light = close, dark = far, white = occlusion. Flow colour encoding; dark to light = negative to positive value.

3 Synthetic vs. Real-Scenes

Most comparisons that have been performed recently in computer vision compare their stereo and motion analysis algorithms against the Middlebury evaluation test data [2, 7, 22]. This has been a major help to advance the community for stereo and motion algorithms. Examples of accurate methods are *Semi-Global Matching* (SGM) [13] for stereo, and the duality based optical flow technique from [26].

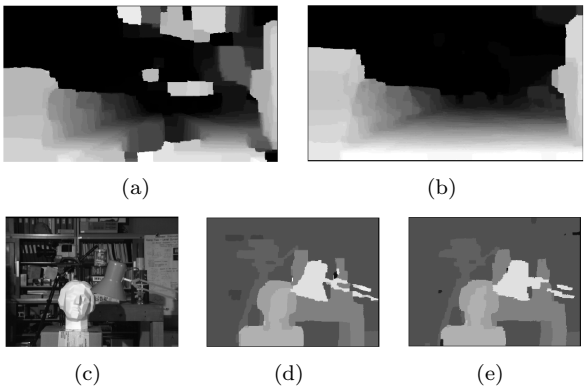
However, the focus in these evaluations is on accuracy for good quality synthetic or engineered real-world scenes. This may not be the best focus for some applications. In the case of driver assistance, robustness has a higher priority than absolute precision. The Middlebury computer vision website has now become the focus of the vision and optical flow community, and appears to be driving the research, rather than being used as a tool to assist it. There is belief in the community that the differences in accuracy for the high ranking algorithms, is within the magnitude of the errors of the ground truth data itself [21].

This section highlights some papers that show that optimising for synthetic data, makes the results on real-world scenes worse, or algorithms that perform well on synthetic data, do not work well on real images and vice-versa.

3.1 Stereo Performance

An evaluation was done to compare a stereo algorithm, top-ranking on [2], against our real-world scenes. The chosen method was Belief Propagation (BP) [25]. When this was tested directly against the real-world scenes the results were not very good, even though it performed well on Middlebury stereo data. An example is shown in Figure 4, where 4(a) and 4(d) show clearly the differences between real-world and synthetic results. The real-world results have a lot of artifacts and clear errors in disparity estimation. However, if you perform BP on the edge images [11], then the results are improved dramatically for real-world results (Figure 4(b)). Testing this edge image BP stereo approach on *Tsukuba* images shows that the results are made (slightly) worse on the synthetic image (Figure 4(d) vs. 4(e), respectively). This is one example showing how focusing on synthetic data as an analysis medium may not produce the desired outcome. A more comparative study is performed in [11].

Another comparison using SGM has been performed in [14]. This study involved comparing differing cost functions, smoothing functions and introducing a second order prior. These results were compared and optimised for the Tsukuba image set (Figure 4(c)). The results were improved for



(a) (b) (c) (d) (e)

Figure 4: Results of Belief Propagation Stereo, comparing results from synthetic images vs real-world images. Light areas are high disparity, dark areas are low disparity. (a) and (b) are performed on the *Construction-Site Sequence*. (d) and (e) are performed on the *Tsukuba* test image, i.e. (c). (a) and (d) show normal BP results. (b) and (e) show BP on the edge images. The differences are explained in the text.

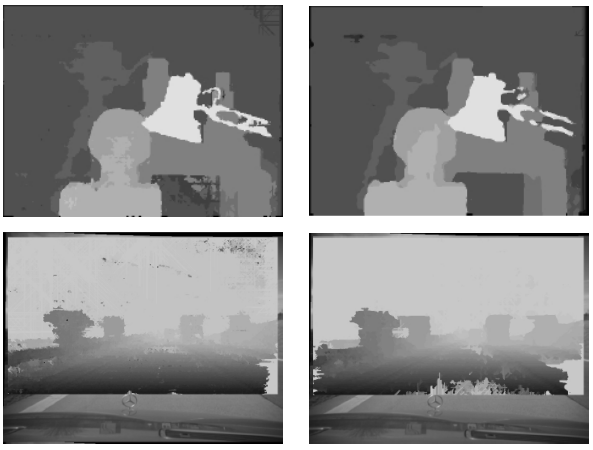


Figure 5: This figure shows the results of SGM Stereo, comparing results from synthetic images vs real-world images. Top: Performance of original (left) vs. optimised (right) SGM on Tsukuba (dark = far, light = close). Bottom: Performance of original (left) vs. optimised (right) SGM on the *Construction-Site* Sequence (light = far, dark = close).

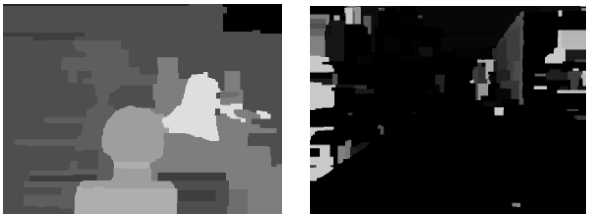


Figure 6: This figure shows the results of BT Stereo on the Tsukuba (left) and *Construction-Site* Sequence (right). Colour encoding: light = close, dark = far.

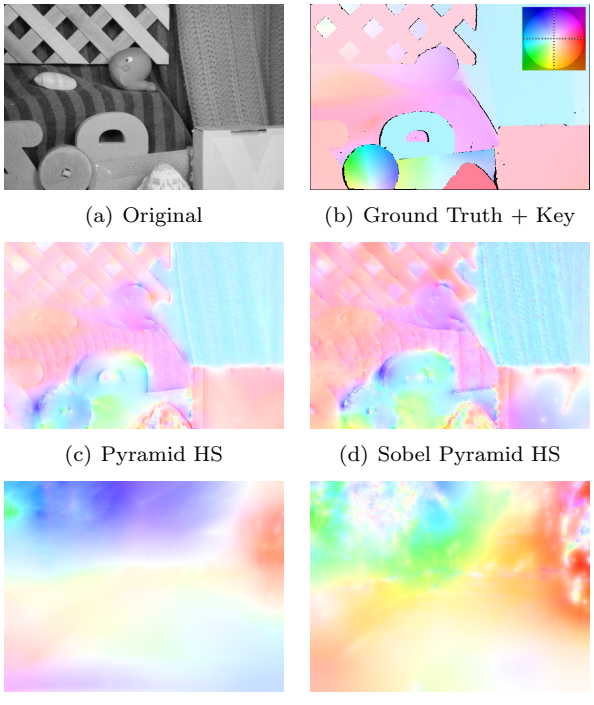
the synthetic data, but they were also obviously made worse against the real-world sequences. Figure 5 identifies this showing that the results have more artifacts on the real-world sequence when SGM is optimised for synthetic data.

Further studies show [20] that algorithms that ranked high on the Middlebury stereo evaluation, such as the Birchfield-Tomasi (BT) algorithm, provide extremely poor results when performed on real-world driving scenes (see Figure 6). This highlights that these algorithms are following a heuristics (e.g., use of discontinuities, vertical smoothing) which does not prove to be suitable for the real-world sequences.

3.2 Optical Flow Performance

Through the literature, there have been few comparisons of real-world results, compared to synthetic results. [10] discusses work in pre-processing showing that their algorithm produces results on both real-world and synthetic data. This is also proven using the duality based approach in [26].

To demonstrate the differences of synthetic and real-world results on optical flow, we used the Horn-Schunck [16] algorithm with a multi-scale pyramid implementation. The pyramid approach helps compensate for large motion vectors and increases the rate of convergence for the algorithm. We tested the effects of using original grey-scale images against Sobel pre-processed images. The reason that the Sobel operator is chosen is that a gra-



(a) Original (b) Ground Truth + Key (c) Pyramid HS (d) Sobel Pyramid HS (e) Pyramid HS (f) Sobel Pyramid HS

Figure 7: This figure shows the results of Pyramid Horn-Schunck on synthetic and real-world scenes. All flow vectors are colour-coded by the key in the top-right corner of (b); direction = colour, intensity = vector length. (a) - (d) show results on *Rubber Whale* from Middlebury. (e) and (f) show the results on the *Dancing-Light* sequence (Figure 2(c)).

dient based edge operator should be more robust to changing illumination, which is a common problem in vision based driver assistance systems. The effects were tested against real-world scenes and synthetic scenes from the Middlebury flow data.

Sample results for this evaluation are seen in Figure 7. The results on the *Rubber Whale* images show that by applying the Sobel operator pre-process step, the quality of the flow gets worse. By comparing Figure 7(c) to 7(d), it can be seen there are more artifacts in the background, image boundaries, and the most obvious error being on the box in the bottom right corner.

However, if the Sobel pre-processing is performed on the *Dancing-Light* scene, the results are improved in on most areas. Figure 7(f) shows the road with vectors mainly going down, which is correct for a forward moving platform, where as Figure 7(e) shows most of the vectors going upward. The truck that is being over-taken on the right is correct in both images as the motion vectors should be pointing to the right. Other areas, such as distant vehicles and points above the horizon seem to have a lot of noise in both images.

3.3 Reasons for Differences

For all the "standard" approaches presented the this section, we have shown that there are major differences between results on real-world data and synthetic data. This subsection presents reasons for this difference.

The approaches rely on good image boundaries. Synthetic data has very obvious image boundaries between objects. This helps both stereo and optical flow algorithms, as it drives the energy minimisation is usually based on a gradient decent of intensity gradients. Real world scenes may not have good image boundaries, especially for optical flow when there is motion blur.

The approaches rely on consistent intensity between images. For synthetic data, the intensity between frames (stereo or optical flow) is usually perfect. However, in real-world scenes, the exposure over time (sequential frames), or between the left and right stereo camera can vary dramatically.

4 Scene Flow Evaluations

In the literature, there have been very little comparisons of scene flow (e.g. [27]), the estimated 3D velocity at every scene point. This is an interesting subject as it combines modern stereo and optical flow algorithms into one framework.

Some evaluation can be found in [24]. This paper uses our synthetic scene introduced above. Available ground truth allows a comparison of scene flow algorithms.

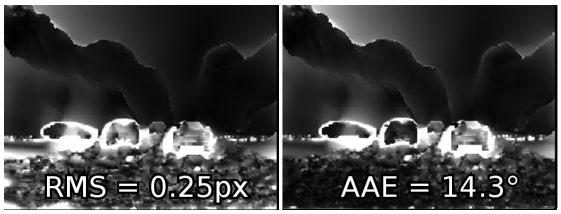


Figure 8: This figure shows results from [24]: RMS (left) and $3AA$ (right) errors on the synthetic sequence. Colour encoding: light = high, dark = low.

The suggested testing criteria are Root Mean Squared Error (RMS) and 3D Absolute Angular Error ($3AA$). These measures are selected to be an extension to the third dimension (disparity change), but remain similar to the evaluation criteria suggested from the research at Middlebury. The defining equations for frames at time t are given below. Let

$$E_{RMS}(t) = \sqrt{\frac{1}{n}\sum_{\Omega}\left\|\begin{pmatrix} u \\ v \\ d \\ d' \end{pmatrix} - \begin{pmatrix} u^* \\ v^* \\ d^* \\ d'^* \end{pmatrix}\right\|^2}$$

where u and v are the x and y optical flow respectively, d is disparity, d' is change in disparity between frames; superscript * denotes the ground truth solution, n is the number of pixels in the image domain Ω, and $\|\cdot\|$ is the L_2 norm. Let

$$E_{3AA}(t) = \frac{1}{n}\sum_{\Omega}\arccos\left(\frac{uu^* + vv^* + d'd'^* + 1}{\sqrt{s(u,v,d')\ s(u^*,v^*,d'^*)}}\right)$$

where $s(a,b,c) = a^2 + b^2 + c^2 + 1$.

Furthermore, errors $E_{RMS}(t)$ and $E_{3AA}(t)$ can be analysed over the entire image sequence, by statistical inference, e.g., standard deviations, medians, etc. This will give a more robust measure to accuracy by taking multiple frames into account. This will place an emphasis on computationally fast algorithms and less parameter tuning. These metrics may not be the best criteria for evaluating scene flow, but further work in evaluating objective criteria to subjective analysis needs to be done. This metric evaluation and statistical analysis of entire sequences is outside the scope of this brief report, but is in the scope for future work.

Examples of RMS and $3AA$ can be seen in Figure 8 using the synthetic sequence in Set 2 on [1].

In [24], their scene flow technique is compared to another good scene flow algorithm [17]. This work can be extended to a larger taxonomy, similar to the work done by [7, 22]. However, this is outside the scope of this brief report, but it is in the scope for future work.

5 Conclusions and Outlook

This paper describes the importance of long stereo sequences by using two publicly available sets of sequences, either real-world sequences without (so far) ground truth, or a long synthetic stereo sequence with ground truth made publicly available. With these two datasets, we show that optimising algorithms for synthetic data may not be the best way to evaluate results. There are multiple examples showing that algorithms that perform well on synthetic data, do not work well in the real world, and vice-versa. Scene flow is another item that can be evaluated using such long sequences, either on long synthetic stereo sequences (with some limited evaluation opportunities), or on long real-world sequences; see also Set 3 on [1]. Future work for our performance evaluation project are: a taxonomy of scene flow algorithms, statistical analysis of errors on long stereo sequences, an evaluation of performance metrics for long scenes, ground truth approximation for real-world driver assistance scenes, and evaluation of robustness using synthetic data.

Acknowledgments:
The sequences of Set 1 were provided by Uwe Franke et al., Daimler AG, Germany. The sequences of Set 3 were provided by Norbert Krüger et al., the European Drivsco project. These sequences can be freely used in academic research.

References

[1] *.enpeda..* Image Sequence Analysis Test Site. `http://www.mi.auckland.ac.nz/EISATS/`

[2] *Middlebury* Computer Vision Website. `http://vision.middlebury.edu`

[3] The Cooke Corporation. Pixelfly VGA Technical Specifications. `www.cookecorp.com`

[4] Point Grey Research. Firefly MV Technical Specifications. `www.ptgrey.com`

[5] H. Badino. A robust approach for ego-motion estimation using a mobile stereo platform. In Proc. *Int. Workshop Complex Motion*, 198–208, 2006.

[6] S. Baker and I. Matthews. Lucas-Kanade 20 years on: a unifying framework. *Int. J. Computer Vision*, **56**:221–255, 2004.

[7] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski. A database and evaluation methodology for optical flow. In Proc. *Int. Conf. Computer Vision*, 1–8, 2007.

[8] J. L. Barron, D.J. Fleet, and S.S. Beauchemin. Performance of optical flow techniques. *Int. Jnl. of Computer Vision (IJCV)*, vol. 12, num. 1, pages 43–77, 1994.

[9] S. S. Beauchemin and J. L. Barron. The computation of optical flow. *ACM Computing Surveys (CSUR)*, **27**:433–466, 1995.

[10] T. Brox, A. Bruhn, N. Papenberg and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In Proc. *ECCV*, 25–36, 2004.

[11] S. Guan, R. Klette, and Y.W. Woo. Belief propagation for stereo analysis of night-vision sequences. In Proc. *PSIVT*, 2009.

[12] P. Handschack and R. Klette. Quantitative comparisons of differential methods for measuring of image velocity. In Proc. *Workshop Aspects Visual Form Proc.*, 241–250, 1994.

[13] H. Hirschmüller. Stereo processing by semiglobal matching and mutual information. *IEEE Trans. Pattern Anal. Mach. Intell.*, **30**:328–341, 2008.

[14] S. Hermann, R. Klette, and E. Destefanis. Inclusion of a second-order prior into semi-global matching. In Proc. *PSIVT*, 2009.

[15] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision,* 2000.

[16] B. K. P. Horn and B. G. Schunck. Determining optical flow. *Artificial Intelligence*, **17**:185–203, 1981.

[17] F. Huguet and F. Devernay. A variational method for scene flow estimation from stereo sequences. Research Report 6267, INRIA, 2007.

[18] R. Klette, S. Stiehl, M. Viergever, and V. Vincken (editors). *Performance Evaluation of Computer Vision Algorithms.* Kluwer, Amsterdam, 2000.

[19] Z. Liu and R. Klette, Performance evaluation of stereo and motion analysis on rectified image sequences. MI-tech-TR-2, University of Auckland, 2007.

[20] Z. Liu and R. Klette. Approximated ground truth for stereo and motion analysis on real-world sequences. In Proc. *PSIVT*, 2009.

[21] R. Mester. Motion beyond Yosemite. Talk at *14th Workshop "Theoretical Foundations Computer Vision"*, Dagstuhl, July 2008.

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

[23] T. Vaudrey, H. Badino and S. Gehrig. Integrating disparity images by incorporating disparity rate. In Proc. *Robot Vision*, 29–42, 2008.

[24] A. Wedel, C. Rabe, T. Vaudrey, T. Brox, U. Franke and D. Cremers. Efficient dense scene flow from sparse or dense stereo data. In Proc. *ECCV*, 2008.

[25] J. Yedidia, W. T. Freeman and Y. Weiss. Understanding belief propagation and its generalizations. In *Exploring Artificial Intelligence in the New Millennium*, 236–239, 2003.

[26] C. Zach, T. Pock and H. Bischof. A duality based approach for realtime TV-L1 optical flow. In Proc. *DAGM*, 214–223, 2007.

[27] Y. Zhang and C. Kambhamettu. On 3D scene flow and structure estimation. In Proc. *CVPR*, Vol. 2, 778–785, 2001.