Performance of Correspondence Algorithms in Vision-Based Driver Assistance Using an Online Image Sequence Database

Reinhard Klette, Norbert Krüger, Tobi Vaudrey, Karl Pauwels, Marc van Hulle, Sandino Morales, Farid I. Kandil, Ralf Haeusler, Nicolas Pugeault, Clemens Rabe, and Markus Lappe

***Abstract*—This paper discusses options for testing correspondence algorithms in stereo or motion analysis that are designed or considered for vision-based driver assistance. It introduces a globally available database, with a main focus on testing on video sequences of real-world data. We suggest the classification of recorded video data into situations defined by a cooccurrence of some events in recorded traffic scenes. About 100–400 stereo frames (or 4–16 s of recording) are considered a basic sequence, which will be identified with one particular situation. Future testing is expected to be on data that report on hours of driving, and multiple hours of long video data may be segmented into basic sequences and classified into situations. This paper prepares for this expected development. This paper uses three different evaluation approaches (prediction error, synthesized sequences, and labeled sequences) for demonstrating ideas, difficulties, and possible ways in this future field of extensive performance tests in vision-based driver assistance, particularly for cases where the ground truth is not available. This paper shows that the complexity of real-world data does not support the identification of general rankings of correspondence techniques on sets of basic sequences that show different situations. It is suggested that correspondence techniques should adaptively be chosen in real time using some type of statistical situation classifiers.**

***Index Terms*—Basic sequences, ground truth, motion analysis, optical flow, performance evaluation, situations, stereo analysis, video data, vision-based driver assistance.**

Manuscript received July 27, 2010; revised December 15, 2010 and February 13, 2011; accepted April 13, 2011. Date of publication May 5, 2011; date of current version June 20, 2011. This work was supported in part by the European Union through the project Learning to Emulate Perception Action Cycles in a Driving School Scenario under Contract 016276-2. The review of this paper was coordinated by Dr. K. Deng.

R. Klette, T. Vaudrey, S. Morales, and R. Haeusler are with the University of Auckland, Auckland 1020, New Zealand.

N. Krüger is with the Mærsk McKinney Møller Institute, University of Southern Denmark, 5230 Odense, Denmark.

K. Pauwels and M. van Hulle are with the Laboratorium voor Neuro- en Psychofysiologie, Faculty of Medicine, Katholieke Universiteit Leuven, 3000 Leuven, Belgium.

F. I. Kandil is with the University of Muenster, 48149 Münster, Germany.

N. Pugeault is with the Centre for Vision, Speech and Signal Processing, University of Surrey, GU2 7XH Surrey, U.K.

C. Rabe is with the Daimler Research, 71059 Sindelfingen, Germany.

M. Lappe is with the Otto Creutzfeldt Center for Cognitive and Behavioral Neuroscience, University of Muenster, 48149 Münster, Germany.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVT.2011.2148134

I. Introduction

DRIVER assistance systems are active safety measures in modern cars, which are also developed for comfort and fuel economy. Vision-based driver assistance systems have been implemented in commercial vehicles since the 1990s, for the first time in the form of a lane-departure warning system by Mitsubishi in 1995 [34]. There is an increasing demand in evaluating such sensor-based components of modern cars, similar to crash tests being performance measures for, e.g., mechanical components of cars.

A. Objectives and Motivation

The main objective of this paper is to report about current work in testing correspondence algorithms (i.e., motion or stereo analysis techniques) on a large variety of test data under realistic recording conditions, as typically occurring in the driver assistance domain.[1] Such "large-scale" testing is not yet supported by popular benchmarks, as used in the computer vision community; these benchmarks are characterized by well-controlled data sets, and as a consequence, their value is only of limited relevance for application domains that deal with outdoor data such as that occurring in driver assistance.

Evaluations of correspondence methods, in general, have already some history in the computer vision literature (for example, see [3], [17], [35], [44], and [47]) and have contributed to the current progress in these algorithms, which is designed for the spatial (S) or temporal matching of image data. For evaluations, particularly in the context of vision-based driver assistance systems, see, for example, earlier work by the authors in [27], [36]–[38], [42], [48], and [49]. This paper summarizes but also extends work reported in these references. In particular, we would like to present the various opportunities in using publicly available test data on the environment perception and driver assistance (*.enpeda..*) image sequence analysis test site (EISATS); see [12].

B. Basic Terms

In this paper, we introduce data sets that are selected or designed to test vision-based driver assistance systems, together

[1]This case certainly also generalizes to other domains such as outdoor robotics, surveillance, human pose recognition, or for mobile platforms in outdoor applications in general. However, this paper only discusses evaluation in the context of driver assistance systems.

Fig. 1. Frames from sequences that show the following three different situations. (Left) Inner city at night. (Middle) Brightness differences (a changing angle toward the sun occurs here). (Right) Close objects.

with a discussion of evaluation methods (when using these data) for stereo or motion analysis. Motion is typically described in computer vision by optic flow (i.e., the visual change of image intensities from one frame to another).

The *ego-vehicle* is the car in which the driver assistance system operates, and *ego-motion* is defined by changes of the ego-vehicle in position, tilt, roll, or yaw angles. Stereo or motion analysis is a low-level task for understanding image data, and results of this analysis are used in subsequent higher order process for understanding traffic scenes. As an example of such a higher order process, we consider in this paper the segmentation of image data into regions where the optic flow is caused either by ego-motion (of the ego-vehicle) or by *independently moving objects*, e.g., other vehicles, pedestrians, or bicyclists.

The EISATS database, in its current form, already supports the evaluation of stereo, motion, or segmentation algorithms. Provided stereo image video data are sequences of some length, typically of about 100–400 frames (or 4–16 s of recording, assuming 25 Hz as the current standard). These sequences are considered to represent particular *situations*. One way of defining a situation is by identifying a particular cooccurrence of *events* in recorded traffic scenes. Examples of events are activities of adjacent traffic (e.g., overtaking, oncoming traffic, and crossing pedestrians), weather and lighting conditions (e.g., rain, sun strike, and patterns of shadow while driving below trees), road geometries (e.g., flat or curved narrow lane, entering a tunnel, driving on a bridge, and a speed bump), or particular events (e.g., traffic signs, a wet road surface, or strong light reflections at night). For example, "driving in daylight on a planar road while overtaking a truck" is an example of a situation that is defined by a concurrent appearance of such events. Another way of defining a situation may be by statistical properties of image data; for example, see the *visual textures* in [1]. However, in this paper, we stay with the first option (i.e., situations defined by events).

Situations typically change every few seconds in normal traffic, and we consider 4–16 s as a reasonable length of a recorded video sequence (which is also called a *basic sequence*) to be identified with one particular situation.

Examples of situations are default driving conditions, inner city traffic at night, brightness differences between both images in the stereo pair, illumination artifacts (e.g., sun through trees along the road), or close objects in front of the ego-vehicle. Fig. 1 shows images that visualize three examples of situations, and these situations are discussed at some length in Section V.

Future testing is expected to be on data that report on hours of driving, and multiple hours of long video data may be segmented into basic sequences and classified into situations. This paper aims at preparing for this expected development.

We are interested in identifying the impact of particular events (i.e., of real-world outdoor issues) on the performance of correspondence algorithms in the context of a given situation. This case allows us to recognize a particular challenge, i.e., often a particular task for research in correspondence algorithms, and, possibly, to also propose ways of overcoming the underlying problem, which is defined by this particular event.

We name a few examples of such challenges as follows: 1) large disparities or motions across frames for image regions that show objects that are close to the ego-vehicle (e.g., caused by passing-by cars) but also due to large rotational motions of the ego-vehicle (i.e., also of the recording cameras); 2) variations in illumination across frames[2] (e.g., slowly due to movements of clouds or a change in the viewing angle relative to the sun or rapidly due to driving through a forested area, having the shade of leaves on cameras for fractions of 1 s); or 3) motion blur in recorded frames. These challenges are only three examples of real-world outdoor challenges. Such challenges often severely affect the complexity of the correspondence problem, which will be solved for stereo or motion analysis.

C. Acronyms Used for Correspondence Algorithms

For stereo analysis, we discuss in this paper loopy belief propagation (BP) stereo [14], dynamic programming (DP) stereo [39] with simple spatial (s) or temporal (t) propagation strategies [30], encoded by the acronyms DPs, DPt, or DPst (note that these propagations can also be applied for the other matching algorithms), semiglobal matching (SGM) stereo [21] with cost functions defined either by mutual information (SGM-MI) or the cost function introduced by Birchfield and Tomasi (SGM-BT) in [7], and graph-cut (GC) stereo matching [28].

Other correspondence algorithms can also be considered, but this set provides a good selection of currently favored stereo-matching approaches. Note that these approaches represent the following three dominant design methodologies of stereo

[2]Reference [22] studies cost functions in stereo algorithms under the particular aspect of brightness differences between stereo frames; see also [20] for a general study of cost functions in stereo algorithms.

Fig. 2. Example of stereo input data (left and middle) and a disparity map obtained with SGM-BT (right). This disparity map is fairly useless in this particular case, and this condition illustrates a "bad performance." Note that the cameras also record some reflections on the windscreen in this example of a situation that might be called "brightness differences between left and right images, and an approaching truck." The first event (brightness differences) causes the BT cost function to fail, and the second event (approaching truck) should be solvable for the SGM matching strategy, in principle, when using a "better" cost function for matching than BT.

matching: 1) optimization along the linear paths of pixels (DP, SGM); 2) GC optimization; or 3) optimization by BP.

Each stereo-matching technique is defined by selected parameters such as weights in cost functions or the size of used neighborhoods when defining cost values. This paper does not intend to evaluate one particular method in detail (e.g., by aiming at optimizing such parameters for a particular situation). This paper points out that matching methods behave very differently for different situations that were recorded in a traffic context, and we define and illustrate ways how we can obtain such evaluation.

The calculation of depth or disparity values (i.e., stereo analysis) is a basic step in the understanding of the surrounding environment of an ego-vehicle. A given stereo algorithm may report very confusing depth values if the input data have not been recorded under ideal conditions. For example, Fig. 2, left and middle, presents a stereo pair in which there is a large difference in brightness between both images. The right image in this figure shows the output (i.e., depth map) of SGM-BT. Note that it is difficult to recognize any 3-D structure of the scene. However this case only shows that SGM-BT is not well suited for image data of this particular situation, and some preprocessing of the input data can possibly change the performance to be better, or SGM-BT may perform "much better" if there would be no brightness differences between the left and right images. In fact, our studies have shown that SGM-BT is "not a good choice" in general due to not only the small neighborhood of contributing pixels, even if there are no brightness differences, but the inherent assumption of brightness constancy in this cost function as well.

We also introduce a few acronyms for the discussed motion analysis algorithms. To compute the optical flow, we will report about Pyramid Horn–Schunck (PyrHS) [23] by extending the basic version of the Horn–Schunck algorithm in OpenCV with a pyramidal control structure, the Brox–Bruhn–Papenberg–Weickert (BBPW) algorithm as described by the four authors in [8], and a total-variation technique using an approximation of the L_1 metric (TVL1) [53]. Again, this selection does not obviously cover the whole diversity of currently discussed motion analysis algorithms, but we aimed at having a methodologically reasonable selection of different techniques for this paper.

Where available, stereo analysis or optic flow sources have been either downloaded from the authors' websites and adopted or fully implemented (partially in contact with these authors as well); the source of basic HS was downloaded from OpenCV.

D. Stability and Robustness

Correspondence methods may be ranked for given situations based on their performance. The *stability* of one method is defined with respect to a particular situation, and a constantly good performance of this method for different basic sequences that all show this situation. The *robustness* of one method is defined by good performance on various situations.

The ranking of correspondence algorithms with respect to stability can be rather different for considered situations. A "top performer" for one situation may not be robust within a given set of situations. For example, BP (which assumes intensity constancy in the data term of its cost function) is often a very good choice if there are no brightness differences in the stereo image data, but its results quickly degrade if there are lighting differences.

Fig. 3 shows complete diagrams of performance values of the stereo algorithms considered for some selected situations, for stereo image (sub)sequences that are 80–140 frames long. The NCC measure used will be specified later. However, with reference to these diagrams, we note that rankings may differ within one sequence from one frame to another and the overall mean performance from one sequence (i.e., situation) another.

For example, the results indicate that SGM-MI may be called "stable" on the *Brightness Difference* sequence, but it ranks low, in general, on the other four sequences. DPt ranks twice fairly high, on the *Ordinary Conditions* and *Close Objects* sequences, and by analyzing the corresponding image data, we noticed that this case correlates to scenes where a high percentage of pixels shows the surface of a planar road (and this case exactly corresponds, on a theoretical level, to the underlying model of the temporal propagation used). Sometimes, there is a strong correlation in the ups and downs of all the stereo algorithms (e.g., at frame 55 in the *Close Objects* sequence), and sometimes, only particular algorithms fail (e.g., depending on the intensity constancy assumption about at frame 110 in the *Brightness Difference* sequence).

Similar variations in performance may be noticed for motion analysis algorithms. Here, TVL1 proved to be superior in general, and the other two motion analysis techniques considered were good only for some situations. TVL1 may be called

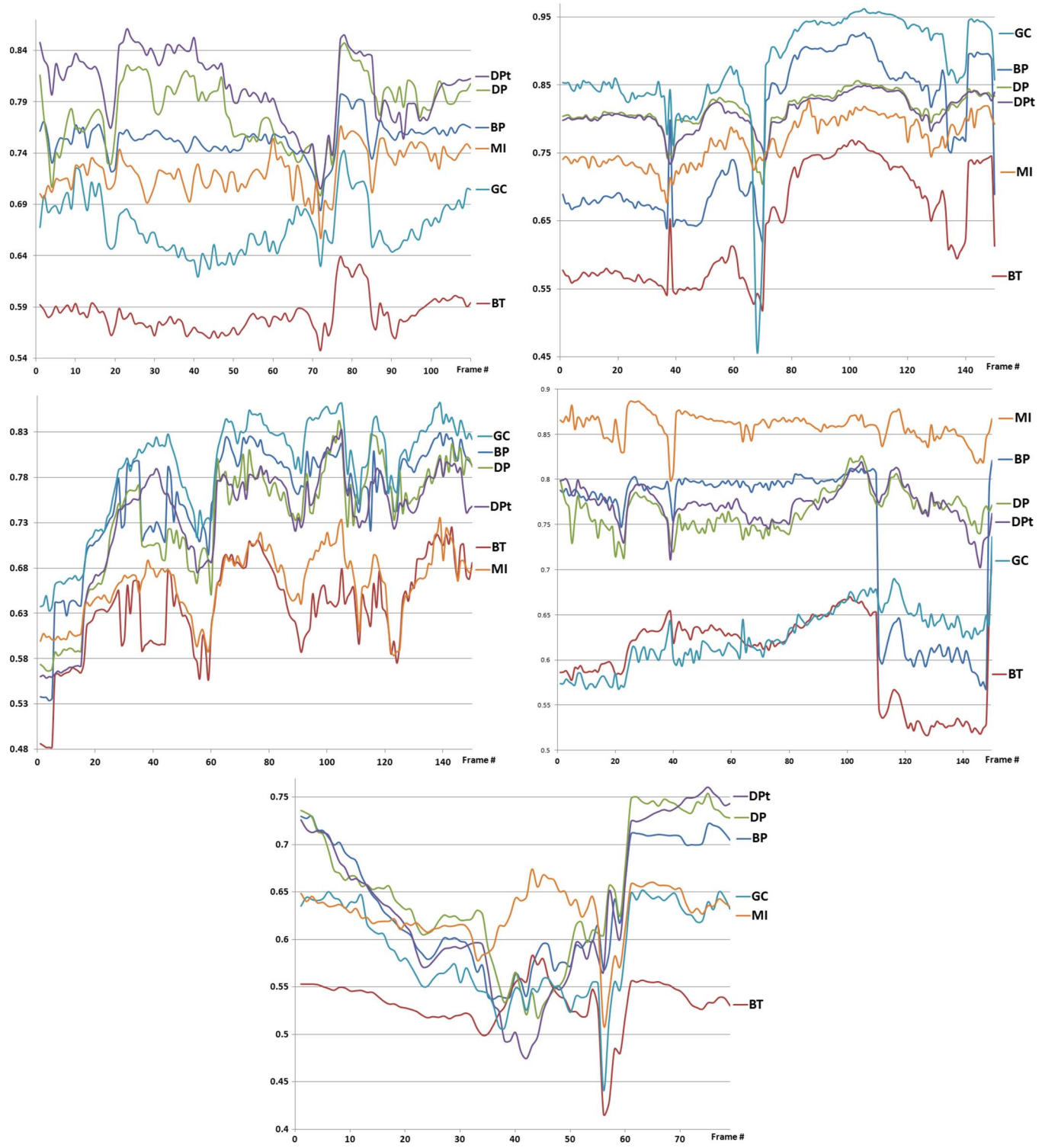


Fig. 3. Performance of stereo algorithms on five different situations using a NCC measure (a larger value is "better") on trinocular sequences. (Top row) *Ordinary Conditions* (left) and *Illumination Artefacts* (right) sequences. (Middle row) *Inner City at Night* (left) and *Brightness Difference* (right) sequences. (Bottom) *Close Objects* sequence.

robust, at least with respect to currently known motion analysis methods.

There was no "clear winner" for stereo analysis algorithms for all the situations considered. Some algorithms that were studied may be called "stable" for some situations, but we do not call any of the stereo-matching algorithms considered "robust" for some kind of extensive class of different situations. However, we have only studied a few different situations so far (which was already quite time consuming), and it might be possible to identify such classes in the future, for example, by sharing more evaluation results for long stereo sequences, as available, on the EISATS database.

E. Contributions of This Paper

As a technical contribution, we introduce the EISATS database [12], which provides appropriate data and evaluation tools for the benchmarking and analysis of correspondence algorithms within a driver assistance context on a much larger scale and variation than currently used benchmarks, which are still defined by small sets of test images. Because the visual information in the driver assistance domain is very different from the commonly used rather-controlled stereo and optic flow benchmarks, their value is limited for practical applications in the driver assistance domain.

We develop tools and perform large-scale benchmarking on long sequences with high variations. This approach allows us to characterize stability or robustness, which was not discussed earlier for small sets of benchmark data. Moreover, there are several situations (e.g., overtaking a truck or driving into a tunnel) in which established correspondence algorithms fail, i.e., their performance drops such that there is no way of ensuring a qualitatively correct 3-D scene analysis. This case points to existing fundamental problems in stereo processing that are not visible in currently used databases.

As a further scientific contribution, our analysis shows that the optimal stereo algorithm, in general, depends on the actual situation, and hence, some categorization of situations is needed to assign whole classes of situations to particular correspondence algorithms. This condition defines a new direction of research.

This paper is structured as follows. First, the next section introduces the EISATS database as available by December 2010, which is structured into sets 1–7. Next, we prepare for stereo and motion performance evaluation; at first, we define error measures, also proposing new summarizing methods for error measures on image sequences and in the following section, we propose prediction error analysis for stereo methods based on calibrated trinocular image sequences.

After all these preparations, we demonstrate comparative stereo analysis on EISATS data, with the aim of highlighting the study of situations. We also use synthesized EISATS data to evaluate motion analysis algorithms. We show that different situations lead to different rankings in stereo or motion algorithms.

Finally, we briefly discuss the use of set 3 for discussing methods for the detection of independently moving objects in traffic scenes. This paper ends with conclusions and comments about future work.

II. EISATS Data

Testing on extensive and varying data sets helps avoid a bias that occurs when using only selective (e.g., "small") sets of data. This paper introduces several sets of "long" test sequences that are made available on the net and used evaluation measures for comparing different algorithms.[3]

[3]In fact, "long" in this paper still translates into durations of sequences in multiple seconds only rather than of minutes, hours, or days; however, sequences of 150 or 400 stereo frames already allow us to illustrate the potentials of such "long" sequences to test and improve correspondence algorithms.

Data that are relevant for driver assistance applications have an unlimited range of variations ("expect the unexpected") due to the potential range of events and, thus, of their combinations into situations. Selective ("small") sets of test data, e.g., with a focus on rendered or engineered (good lighting, indoor) scenes, are insufficient for serious testing. EISATS (see [12]) is not focused on one particular set of data or one particular evaluation strategy but is open to researchers in vision-based driver assistance systems for applying the data in their evaluations, as well as for contributing more (best verified) data. The website has recently contained seven different sets of test sequences, which were provided by different research groups in vision-based driver assistance and of relevance for particular evaluation strategies. We illustrate the use of some of these data sets in this paper. We do not discuss data in sets 6 and 7 of EISATS in this paper; more sequences with accompanying range scans or segmentation ground truth are in preparation, and this case will be a subject of more specialized papers.

For the case of stereo and optic flow, we demonstrate that the performance of established algorithms on small sets of test data, e.g., [35], [44], and [47], does not necessarily describe their performance on data as used in experiments in vision-based driver assistance. This case reflects a common problem that (for example) engineers who work in certain application areas find only little indications on the actual relevance of an algorithm in their specific scenario on a very selective set of test data.

We see the EISATS database as a dynamic forum for relevant data and benchmarks for vision-based driver assistance. See Table I for a brief characterization of available sets of image sequences.

Each of the EISATS image sequences [12], as used in this paper, represents a few seconds of driving (i.e., typically showing one situation). Sets 1–5 provide some diversity of situations, and the goal is to extend these sequences in a more systematic way. These sequences represent only a very small segment of possible situations in vision-based driver assistance (e.g., our discussion does not cover sequences that were recorded in the rain with moving wipers or against the sun).

However, the sets given allow us to go to a new quality of performance evaluation for correspondence algorithms compared to "no-sequence" data sets on sites as shown in [6], [10], [32], and [33], which do not support studies about the influence of illumination artifacts, temporal filtering, or having low-contrast images with rapid changes due to events that happen in real-world driving situations.

III. Error Measures

Because we deal with long sequences, we can analyze results over time, where frames are indexed by t. There are hundreds of error measures available (e.g., see [11]), and we aim to use general error measures that apply to several types of evaluation data.

If the ground truth is available, we may calculate, at a pixel position p, the Euclidean distance

$$E_t(p) = \|A_t(p) - B_t(p)\|_2$$

TABLE I
DATA SETS OFFERED ON THE EISATS WEBSITE IN DECEMBER 2010

Set	Comments
1	Night vision stereo sequences (Daimler AG) These seven stereo night vision sequences (12 bit, between 220 and 300 pairs of frames each) have been provided by Daimler AG, Germany, in June 2007 (group of Dr. Uwe Franke). These sequences come with ego-motion data and time stamps for each frame.
2	Synthesized stereo sequences (*.enpeda..* & Daimler AG) These synthesized stereo sequences (with ground truth) have been provided by Tobi Vaudrey (.enpeda..) and Clemens Rabe (Daimler AG).
3	IMOs in color stereo sequences (Drivsco) These three day-time, color stereo sequences have been provided by the European Drivsco project. Independent moving objects ground truth and gaze data is now available.
4	"Normal camera" binocular stereo sequences (Hella Aglaia Mobile Vision & *.enpeda..*) A few of those day- or night-time, grayscale stereo sequences have been provided by Hella Aglaia Mobile Vision GmbH, Germany; most of them have been recorded by students in the *.enpeda..* project.
5	"Normal camera" trinocular stereo sequences (*.enpeda..*) Three-camera stereo sequences (rectified by pairs) captured with HAKA1.
6	Grayscale stereo sequences with range scans (HU Berlin, *.enpeda..* & Daimler AG) So far three stereo vision sequences where the test vehicle drives through a car park; ground truth from a laser scanner; SGM, block and cross matcher disparity maps are also included.
7	Grayscale stereo sequences for scene labeling (Daimler AG) Stereo sequences with ground truth for scene labeling analysis (segmentation).

of a generated result $A_t(p)$ from the ground truth $B_t(p)$. Doing so for all available (e.g., nonoccluded) pixel positions p, we obtain an error image E_t.

In general, A_t and B_t are both n-valued functions (e.g., for stereo, we have one disparity value, and thus, $n = 1$, the optical flow is a field of 2-D vectors with $n = 2$, or the scene flow combines disparity with the optical flow, and we have $n = 3$ in this case).

From such an error image E_t, we can derive various measures, e.g., the mean μ, standard deviation σ, zero-mean standard deviation σ_0 (which is also referred to as the *root-mean-square error*), or, simply, maxima max.

We can explain this case better by translating to some common error metrics that are already used in the community. In the case of stereo matching, one common metric is the root-mean-square error. For example, in [44], this metric is simplified to

$$\sigma_0(E_t) \text{ with } n = 1.$$

For the optical flow, the common measure is the mean endpoint error [3], i.e.,

$$\mu(E_t) \text{ with } n = 2.$$

To be even more specific, consider the case of stereo algorithms first. Assume that we have to compare two images A_t and B_t (e.g., the calculated depth map with the ground-truth depth map) at time t at all pixel locations p in a set Ω_t (e.g., all nonoccluded pixels). The applied evaluation measures are pointwise root mean square

$$R_p(t) = \sqrt{\frac{1}{|\Omega_t|} \sum_{p \in \Omega_t} \left[A_t(p) - B_t(p)\right]^2} = \sigma_0(E_t)$$

and spatial root mean square, where we compare Gaussian means of local neighborhoods (around the reference pixels) as

$$R_s(t) = \sqrt{\frac{1}{|\Omega_t|} \sum_{p \in \Omega_t} \left[\mu\left(G_\sigma(p) * E_t(p)\right)\right]^2} = \sigma_0(G_\sigma * E_t)$$

where $G_\sigma(p)*$ is a Gaussian convolution. Another measure is normalized cross correlation (NCC), i.e.,

$$N(t) = \frac{1}{|\Omega_t|} \sum_{p \in \Omega_t} \frac{\left[A_t(p) - \mu_t^{(A)}\right]\left[B_t(p) - \mu_t^{(B)}\right]}{\sigma_t^{(A)} \sigma_t^{(B)}}.$$

These errors are calculated along the given sequences, frame by frame, and conclusions are drawn based on mean errors and error variances along sequences or due to particular error patterns at particular subsequences (e.g., for the occurrence of large occlusions or of brightness alterations). In the case of motion vector fields, we calculate either the average angular errors or mean endpoint errors $\mu(E_t)$ between vectors at corresponding pixel positions.

The NCC mean m_N and standard deviation σ_N of stereo-matching techniques

$$m_N = \frac{1}{T} \sum_{t=1}^{T} N(t) \quad \text{and} \quad \sigma_N^2 = \frac{1}{T} \sum_{t=1}^{T} \left[N(t) - m_N\right]^2$$

on individual sequences (e.g., with $T = 100$ or more stereo frames) allows us to identify a *winner* (as always, defined by the order of the means) for the recorded situation and its *steadiness* (standard deviation).

However, the winner algorithm may not be the best in every frame within a given sequence. To measure this approach, we compare each two algorithms using *sums of direct comparison*. Given two algorithms, e.g., C and D, and the corresponding NCC values $C(t)$ and $D(t)$ for every frame t in a given sequence of length T, the sum of direct comparisons between C and D is given by

$$S_{DC}(C, D) = \sum_{t}^{T} \Delta(t)$$

Fig. 4. Application for a sequence in set 5. (Left) Third view. (Middle) Virtual view for the disparity map shown on the right. The matching algorithm applied was BP stereo analysis. The specularity (left), which is apparent both in the recorded left and right images, causes a "defect" in the calculated disparity data.

where

$$\Delta(t) = \begin{cases} 1, & C(t) > D(t) \\ 0, & C(t) = D(t) \\ -1, & C(t) < D(t). \end{cases}$$

Note that the absolute values of $S_{DC}(C, D)$ and $S_{DC}(D, C)$ are equal. Thus, we can define a ranking for each sequence based on sums of direct comparisons. Let

$$s_N(S_j) = \sum_{i=1}^{6} S_{DC}(S_i, S_j) \quad S_j \neq S_i$$

where the S_i's represent the six selected stereo algorithms (SGM-BT, DP, DPt, DPs, BP, and GC), and S_j is the particular stereo algorithm for comparison.

Taking multiple sequences for the same situation or even all sequences for all situations, the NCC mean and standard deviation defines the *robustness* of methods to these data. We illustrate this case for five different situations.

IV. Virtual Views for the Evaluation of Stereo Matching

Set 5 of EISATS offers five trinocular image sequences, where the third camera may be used for prediction error analysis on stereo image sequences [37], similar to the prediction error analysis in [45] for motion analysis.

The prediction error strategy is a valuable tool for objectively evaluating the performance of stereo algorithms when the ground truth is unavailable or impossible to acquire at full range and sufficient accuracy. The prediction error strategy requires only that input data are captured with (at least) three cameras, two of which are used as input of the algorithms, and the remaining camera (third) is used for evaluation. The calculated depth data are used to map one of the stereo images (e.g., of the "left" camera) into the pose of the third camera, thus defining the *virtual* view. The similarity between the virtual and third views characterizes the quality of the stereo algorithm used. Because of possible brightness differences between the left and third views, NCC (rather than the root-mean-square error) is used to quantify this similarity. The set Ω is defined by all pixel positions in the virtual view, which receive a mapped image value of the left image.

See Fig. 4 for an example of a recorded third view, a calculated virtual view, and the disparity map (a result of applying BP stereo matching), which was used for calculating this virtual view. The use of several trinocular sequences for prediction error analysis has been demonstrated in [27], and more such sequences are now available on EISATS. The geometric approach of the prediction error methodology was specified in [36], only using sequence 1 of set 5 in [12] as a long real-world sequence at that time. All contributing cameras are calibrated [18], thus making the mapping of data into defined poses possible.

TABLE II
Overall Results for the Ordinary Conditions Sequence of Set 5 (See Fig. 3 for the Complete Diagram of Values per Frame). Left: Mean and Standard Deviation. Right: Sums of Direct Comparisons for (SGM-BT–BP), (DP–BP), (DP–SGM-BT), and (DPt–BP)

Algorithm	Mean	St. Dv.
DPt	0.81	0.03
DP	0.78	0.03
BP	0.75	0.02
SGM-MI	0.72	0.02
GC	0.67	0.03
SGM-BT	0.58	0.02

	BP	BT	DP	DPt	GC	MI	s_N	Rank
BP	0	-	-	-	-	-	146	3
SGM-BT	-110	0	-	-	-	-	-550	6
DP	76	110	0	-	-	-	334	2
DPt	106	110	70	0	-	-	506	**1**
GC	-110	110	-110	-110	0	-	-328	5
SGM-MI	-108	110	-108	-110	108	0	-108	4

V. Evaluations for Situations

In this section, we illustrate the use of sequences, as provided on EISATS, to evaluate the performance of stereo algorithms for particular situations. The classification of sequences into situations was manually done, only the by subjective evaluation of contributing events.

A. Ordinary Driving Conditions

Ordinary conditions are conditions in which the traffic is relatively light, the brightness differences between the stereo pair are minimum, the sun is still high in the sky, and there are no objects in the borders of the road that may create *illumination artifacts* (see Section V-B). Shadows and specularities are minimum. Note that such conditions can also be present in a cloudy environment. Sequences in sets 1 and 4 of EISATS are mostly in this category.

We show results for the Ordinary Driving Conditions sequence in set 5. The algorithm with better performance (with respect to the NCC mean) was DPt, followed by DP and BP. All the algorithms presented their worst performance in that sequence when the followed and incoming vehicles are closer

TABLE III
RESULTS FOR THE 150 FRAMES OF THE *ILLUMINATION ARTEFACTS* SEQUENCE (SEE FIG. 3 FOR THE COMPLETE DIAGRAM OF VALUES PER FRAME). LEFT: MEAN AND STANDARD DEVIATION. RIGHT: SUMS OF DIRECT COMPARISONS FOR (SGM-BT–BP), (DP–BP), (DP–SGM-BT), AND (DPt–BP)

Algorithm	Mean	St. Dv.
GC	0.88	0.08
DP	0.82	0.03
DPt	0.81	0.02
BP	0.78	0.10
SGM-MI	0.77	0.03
SGM-BT	0.64	0.08

	BP	BT	DP	DPt	GC	MI	s_N	Rank
BP	0	-	-	-	-	-	-6	4
SGM-BT	150	0	-	-	-	-	-446	6
DP	4	150	0	-	-	-	270	2
DPt	4	150	-112	0	-	-	48	3
GC	144	146	138	142	0	-	712	**1**
SGM-MI	4	150	-142	-148	-142	0	-278	5

to the ego-vehicle. DPt, the winning algorithm, did not show the best performance in every single frame, and for around ten frames, it performed worse than DP. See Table II and Fig. 3, top row, left.

B. Illumination Artifacts

Illumination artifacts (e.g., while driving below trees) are present in most of the sequences of sets 1, 4, and 5. We show results for the Illumination Artifacts sequence in set 5. This sequence was recorded over a road that is surrounded by trees. This situation, in general, did not drastically modify the brightness between the stereo pairs but introduced a considerable number of different dark and bright patches (caused by the foliage) in the left and right images. It also introduced a fast change in the lighting conditions between subsequent frames.

For this particular sequence, we noticed that, when the trees are closer to the right side of the road (because this sequence was recorded in the late afternoon, when the sun was in a low position in the left side of the ego-vehicle), the difference in brightness between the stereo pair is reduced, improving the performance for most of the algorithms. The top-performing algorithm was GC, followed by the two DP approaches; see Fig. 3, top row, right. One interesting point to note with this sequence is that BP had a slightly better performance than SGM-MI with respect to the mean; however, the latter algorithm had a better performance in a larger number of frames than the former algorithm, as shown in Table III.

C. Inner City at Night

One sequence of an Inner City at Night situation (set 5) is recorded after sunset with regular to dense traffic on the road; the scene is illuminated by the lights of the other vehicles, and lights and specularities cause large white regions of missing dynamics in intensity values. Using NCC as a quality metric, GC was the algorithm with the best performance on the selected original sequence, whereas SGM-BT showed the worst performance; see Table IV, left. In the first 30 frames, the performance

TABLE IV
RESULTS FOR THE 150 FRAMES OF THE *INNER CITY AT NIGHT* SEQUENCE (SEE FIG. 3 FOR THE COMPLETE DIAGRAM OF VALUES PER FRAME). LEFT: MEAN AND STANDARD DEVIATION. RIGHT: SUMS OF DIRECT COMPARISONS FOR (SGM-BT–BP), (DP–BP), (DP–SGM-BT), AND (DPt–BP)

Algorithm	Mean	St. Dv.
GC	0.79	0.06
BP	0.76	0.06
DP	0.74	0.07
DPt	0.73	0.07
SGM-MI	0.66	0.04
SGM-BT	0.64	0.05

	BP	BT	DP	DPt	GC	MI	s_N	Rank
BP	0	-	-	-	-	-	332	2
SGM-BT	-150	0	-	-	-	-	-696	6
DP	-88	150	0	-	-	-	60	3
DPt	-92	150	-26	0	-	-	0	4
GC	138	150	146	150	0	-	734	**1**
SGM-MI	-140	96	-118	-118	-150	0	-430	5

of all the algorithms is poor, as a consequence of a close object that is present in the scene (see Fig. 3, middle row, left).

For this sequence, the ranking with the NCC mean and the ranking that was obtained with sums of direct comparisons were the same. However, as shown in Table IV, top, DP had a better performance than BP in almost half of the frames.

D. Brightness Differences

Brightness differences between the input stereo pair are a common issue in driver assistance. For example, by changing the viewing angle with respect to the sun, one camera may record a brighter sequence of frames than another camera. Of course, intercamera communication may somehow relax this issue in the future.

The output of correspondence algorithms is severely affected in such a situation of brightness differences. In the Brightness Differences sequence of set 5, there are brightness differences in every frame, and they increase in the last 40 frames. The algorithm with the best performance on this sequence was SGM-MI, followed by DP and DPt (see Table V, top). This ranking is in accordance with the results obtained in [37] in the case of the brightness altered sequence, supporting the idea that the prediction error is a good technique for evaluating stereo algorithms in the absence of the ground truth. BP showed a good performance (second) until the difference in brightness has become extreme, in which its performance is the second worst. This situation was reflected in the ranking defined by the sums of direct comparisons, in which BP was the second best, because its performance was degraded until the last third of the sequence; see Fig. 3, middle row, right.

E. Close Objects

Sequences with close objects (e.g., people, other vehicles, and static structures) are very important to be investigated, because this situation is the main characteristic during a traffic jam or a potential conflict. In this situation, it is likely that the driver assistance systems implemented should contribute to the adaptation and optimization of driving. In the analyzed Close

TABLE V
OVERALL RESULTS FOR THE 150 FRAMES OF THE *BRIGHTNESS DIFFERENCES* SEQUENCE. LEFT: NCC MEAN AND STANDARD DEVIATION. RIGHT: SUMS OF DIRECT COMPARISONS FOR (SGM-BT–BP), (DP–BP), (DP–SGM-BT), AND (DPt–BP)

Algorithm	Mean	St. Dv.
SGM-MI	0.86	0.01
DPt	0.77	0.02
DP	0.77	0.02
BP	0.75	0.08
GC	0.63	0.03
SGM-BT	0.60	0.05

	BP	BT	DP	DPt	GC	MI	s_N	Rank
BP	0	-	-	-	-	-	174	2
SGM-BT	-150	0	-	-	-	-	-584	6
DP	-58	150	0	-	-	-	28	4
DPt	-42	150	64	0	-	-	172	3
GC	-74	-16	-150	-150	0	-	-540	5
SGM-MI	150	150	150	150	150	0	750	**1**

TABLE VI
OVERALL RESULTS FOR 79 FRAMES OF THE *CLOSE OBJECTS* SEQUENCE. LEFT: MEAN AND STANDARD DEVIATION. RIGHT: SUMS OF DIRECT COMPARISONS FOR (SGM-BT–BP), (DP–BP), (DP–SGM-BT), AND (DPt–BP)

Algorithm	Mean	St. Dv.
DP	0.65	0.07
BP	0.64	0.06
DPt	0.63	0.08
SGM-MI	0.63	0.03
GC	0.59	0.05
SGM-BT	0.53	0.03

	BP	BT	DP	DPt	GC	MI	s_N	Rank
BP	0	-	-	-	-	-	123	2
SGM-BT	-71	0	-	-	-	-	-331	6
DP	35	65	0	-	-	-	241	**1**
DPt	-5	61	-43	0	-	-	81	3
GC	-77	55	-67	-57	0	-	-201	5
SGM-MI	-5	79	-31	-11	55	0	87	4

Objects sequence of set 5, two pedestrians appear in front of the vehicle (without any actual danger, because the ego-vehicle stopped earlier).

The ranking of the algorithms differs from the two previous algorithms: DP is the overall best algorithm, as shown in Table VI, top, and Fig. 3, lower row. However, four different algorithms have the best performance throughout the sequence for particular intervals of time. The performance of all algorithms is below the standards around the middle of the sequence (except for SGM-MI) when the two pedestrians are very close to each other. Even the top-performing algorithm, DP, had a worse performance than the worst overall algorithm (SGM-BT) for about 15 frames, as shown in Table VI, bottom.

F. Summary

We briefly summarize the detected robustness of the algorithms across the five situations presented in this paper (see Table VII). Using the mean of NCC over all the situations, DPt outperforms (by a small difference) all the other algorithms, although it only performs the best in two of the situations presented. However, the performance of DPt heavily depends on the percentage of pixels that show a planar road surface.

TABLE VII
OVERALL NCC RESULTS OVER THE FIVE DIFFERENT SITUATIONS CONSIDERED IN SECTION V

Algorithm	Mean	St. Dv.
DP	0.65	0.07
BP	0.64	0.06
DPt	0.63	0.08
SGM-MI	0.63	0.03
GC	0.59	0.05
SGM-BT	0.53	0.03

On the other hand, GC finalized as the second worst algorithm, although it was the best in two of the sequences.

This paper only discusses five sequences (situations). Summarizing our more general experience and taking our experimental results into account, which are not reported in this paper, we may conclude that the following cases hold.

- Cost functions should not depend on brightness constancy (as an alternative, some kind of preprocessing methods may map the given stereo sequences into data where the impact of brightness differences has been reduced, e.g., by using redials with respect to smoothing).
- The well-known streaking-effect of DP also limits the use of this simple matching approach in the given application context.
- SGM can potentially deal with scenes of high-depth complexities.
- BP may be preferred in scenes with larger homogeneous regions.
- GC has a tendency to create convex regions of nearly constant depth.

These findings are also accompanied by progress in real-time implementations of DP, SGM, and BP, but there is the lack of fast implementations of GC variants.

VI. SYNTHESIZED VIDEO SEQUENCES

Synthetic data with the ground truth already have a history in computer vision. Long synthetic sequences, e.g., in set 2 of EISATS, are very useful for studying defined variations in image data and for analyzing their impact on the performance of a selected stereo correspondence algorithm.

In [37], the performance of several stereo algorithms was tested over different adverse conditions (e.g., blurred images, stereo pairs with differences in brightness, and images corrupted with Gaussian noise) by modifying the rendered sequence 1 of set 2 in [12]. An objective evaluation (using the root-mean-square error and the percentage of miscalculated bad pixels), based on the available ground truth for such a rendered sequence, showed that, indeed, the ranking of the studied algorithms varied, depending on the modifications that are applied to the sequence. However, a ranking of methods on such rendered sequences is not well correlated to a ranking on real-world sequences for particular situations. This case is due to, at least, the following two facts: 1) The synthesized sequences have yet to be perfectly photorealistic and physics based and are thus different from the real-world sequences recorded with specific cameras, and 2) these sequences are

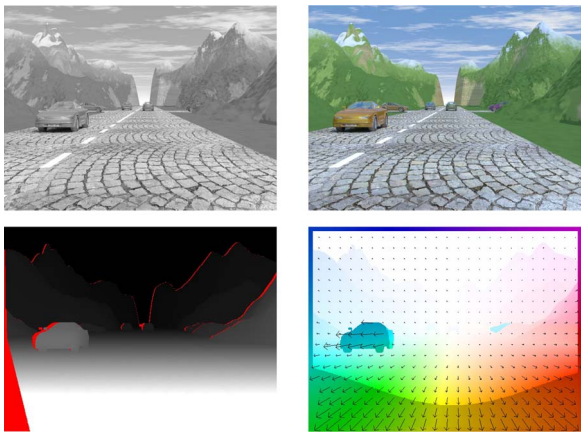


Fig. 5. (Top) Example images (frame 42) from synthetic sequence 1 in grayscale (left) and color (right). (Bottom) Disparity map (left), with light to dark as near to far and red as occlusion, and flow map (right), with the color as direction (see border) and saturation as length. A vector map is overlayed for additional information.

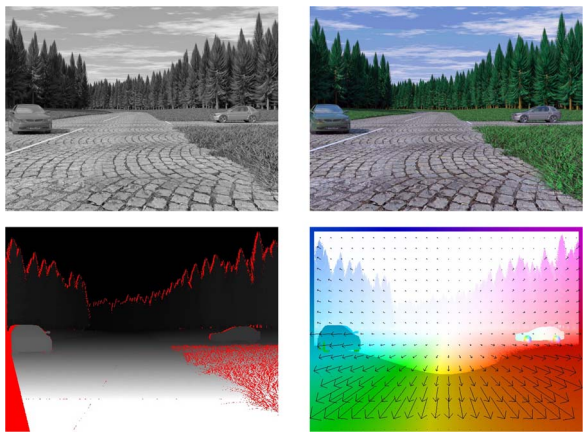


Fig. 6. (Top) Example images (frame 219) from synthetic sequence2 in grayscale (left) and color (right). (Bottom) Disparity map (left) and flow map (right), with color encoding as in Fig. 5.

also very limited with respect to the covered situations or (unpredictable) events in the real world.

Synthesized data are particularly important for motion analysis algorithms; relatively slow recording of video sequences (e.g., at 25 Hz) does not allow us to use prediction error analysis [45] for evaluation, and there still are not many studies on the performance of motion analysis on real-world sequences (e.g., in [52]).

A. *Sequences 1 and 2*

Synthetic sequence 1 [a synthetic Persistence of Vision Raytracer (POV-Ray) sequence of 100 stereo frames] was introduced in [48] and made publicly available (with stereo and motion ground truth) in set 2 in [12]. This sequence was the first long stereo sequence, with ground-truth data for the optical flow (both x- and y-directions), disparity, and disparity rate (change in disparity between frames for scene-flow ground truth). These data are generated with ray tracing and texture mapping, generating a very clean-looking image. Fig. 5 shows one example of the original images and ground-truth interpretations. Furthermore, it is one of the first stereo databases that contain $>$ 8-b dynamic range; the sequence contains 12-b grayscale and 3 $\times$ 12-b color depth. This case is comparable to top-of-the-line commercially available machine vision cameras (e.g., [4] and [41]). This scene has been used to compare stereo [37], optical-flow [38], and scene-flow algorithms [51] in various papers.

There is an advanced sequence available, i.e., synthetic sequence 2 in set 2 [12]. This sequence contains a more realistic driving situation, which also includes trees and grass. This sequence aims at being more challenging for the optical-flow and stereo algorithms. Example frames are shown in Fig. 6. It has all the same qualities as sequence 1 (i.e., high-dynamic-range input, with the ground truth available). Furthermore, the ground-truth ego-motion (i.e., fundamental matrix) is available for every sequential pair of images. This condition gives the ground-truth movement of the cameras from frames $t-1$ to t. This case allows us to use this information to create "biased" algorithms and also test ego-motion/fundamental-matrix algorithms. Ego-motion estimation is a very important aspect for driver assistance, because vibrations and variations in the road cause very large rotational ego-motion between frames.

B. *Results on the Data Provided*

The following results are for synthetic sequence 1. To compute the optical flow, we decided to test PyrHS, BBPW, and TVL1 (see Section I-C). This subset of algorithms highlights the basic and state-of-the-art algorithms for the optical flow. The results for mean endpoint errors (as introduced in Section III) are shown in Fig. 7, left.

Furthermore, we can alter the input data to contain noise that is present in real-world imagery. This approach was done in [38] and highlighted that illumination differences cause the major problems in both stereo and optical-flow algorithms. This condition is obvious, because both types of correspondence algorithms rely on the *intensity consistency assumption*, i.e., that the pixel on an object will look identical between corresponding images.

Sample results are shown in Fig. 7, right. The shape is because the brightness differences are large ($\pm$100 intensity values) at the start of the scene and reduce to zero at midpoint, before increasing back to $\pm$100. Obviously, TVL1 is more sensitive to major illumination differences compared with the other algorithms. Furthermore, with a difference of only $\pm$10 (see around the middle of the sequence), the algorithm rankings are the same as the case for no illumination difference.

This evaluation was only given as an example of what can be done with the provided data with the ground truth. A much more extensive test, varying parameters, and noise properties can be investigated to exploit these data. One major hole in the literature is an extensive evaluation of the importance of having a high-dynamic range for machine vision. From a practical point of view, we have experienced that the stereo and optical-flow results are of a much higher quality when the dynamic range is high. This case is obvious, because the cost functions have an easier discretization between possible matches. This

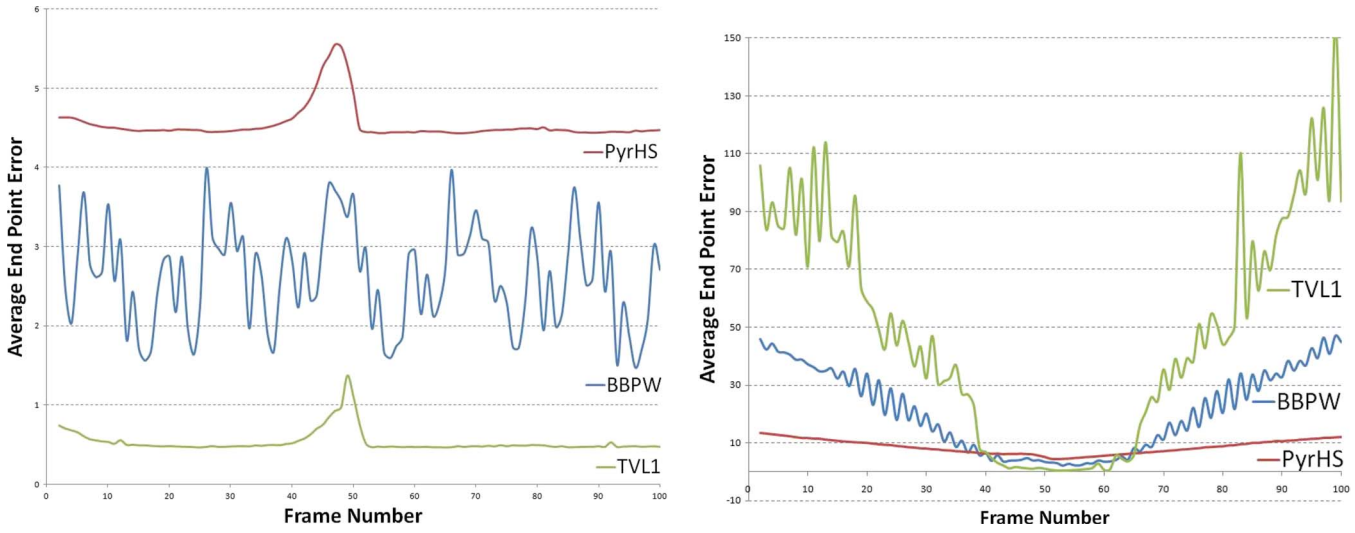


Fig. 7. Mean endpoint error results across image sequence, compared with the ground truth on the original data (left) and on the illumination altered data (right). The total average was 4.5 (PyrHS), 2.6 (BBPW), and 0.53 (TVL1) for the original data and 8.6 (PyrHS), 21.0 (BBPW) and 48.3 (TVL1) for the illumination altered data.

effect needs to be studied in detail, and the data provided here make that condition possible.

C. Possible Future Extension of Set 2

Most studies work on adding artificial noise to generated scenes (as done in [37] and [38]). This noise needs to be more realistic. One way of doing this approach is by introducing the noise generation into the image generation process. We demonstrate opportunities of physics-based rendering for camera modeling, which is planned to be used in further sequences in set 2 of EISATS.

Fig. 8 shows a 3-D model of an urban road intersection, and stereo sequences (path tracing) are rendered either with a simple ray tracer or LuxRender. This approach involves the use of a realistic model of atmosphere and sunlight. A simulation of realistic specular highlights or blooming is of importance, because these events frequently occur in an imagery of outdoor scenes and cause major problems to correspondence algorithms. A scene with specular highlights or reflections cannot be shown by cameras as "perfect," as shown in Fig. 8, upper row (left, or second to the left). An image with moderate bloom and some chromatic aberration simulates some realistic distortion, as shown for common cameras. An image with severe blooming (see Fig. 8, upper row, right) simulates a defective camera (or a camera with overexposure, which often happens in outdoor environments).

We also studied the behavior of BP stereo on such synthetic images. As expected, depth maps that are derived from ray-traced stereo pairs contain only minor errors in image regions that show reflections. However, depth maps from images with moderate blooming and chromatic aberration are not significantly degraded compared to results from undistorted data. Images that are severely degraded by blooming show impaired results. Further studies in this area will identify which noise affects results for stereo and optical flow the most, thus giving the community tools on where they should try to adapt their algorithms.

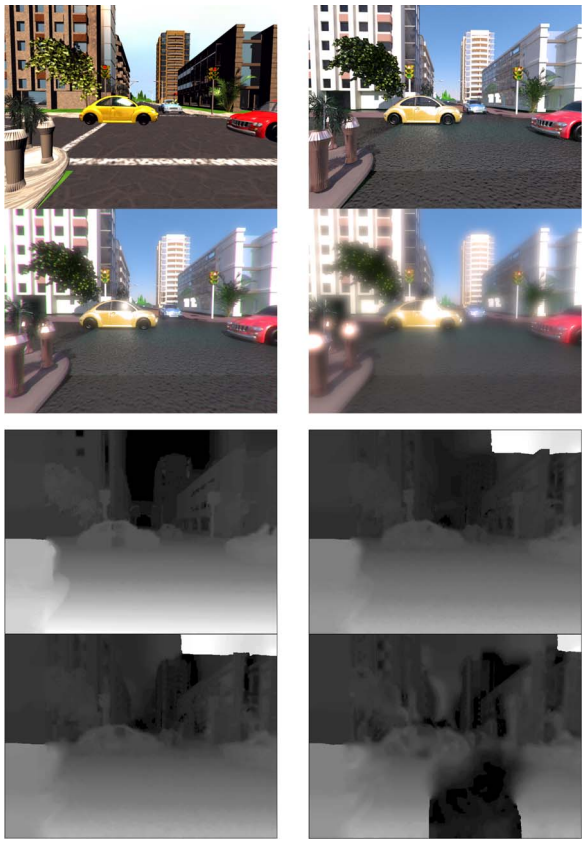


Fig. 8. (Upper row, left to right) 3-D model rendered with simple ray tracing, tone-mapped output of LuxRender (some issues with material support of this render engine are apparent such as missing road marks), also with moderate blooming and chromatic aberration, and severe blooming. (Lower row, left to right) Corresponding BP depth maps.

VII. Independent Moving Objects in Color Stereo Sequences

The sequences in set 3 of EISATS were particularly designed for studying the detection of independently moving objects. This set provides two situations, both with (very) long image

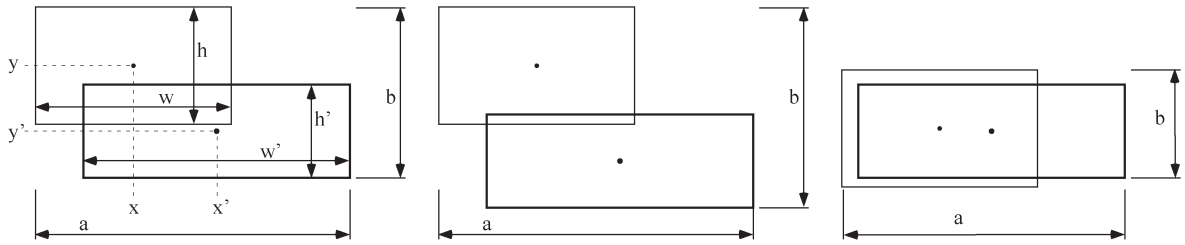


Fig. 9. (Left) Because of $w'h' > wh$ in the case shown, the estimate is equal to $ab/w'h'$. (Middle) ab increases, resulting in an increase of the overlap value $> 0 \cdot 5$. (Right) ab decreases, resulting in an overlap $< 0 \cdot 5$.

sequences. For the detection of independent moving objects (IMOs), we also provide the ground truth in terms of labeled regions with associated information (e.g., the type of IMO, on which lane the car is driving, and additional occlusion properties of the IMOs). Moreover, we define several measurements that allow for the comparative evaluation of IMO detection algorithms. The two sequences show the situations "Suburban Bridge (851 frames)" and "Suburban follow" (1182 frames) with ground-truth information in terms of labeled IMOs.

We discuss the data labeling used. In the following discussion, true IMOs are denoted by $A = (x, y, w, h)$, where the position of the IMO's center is denoted by (x, y), and its width and height is denoted by (w, h). Detected IMOs are denoted by $B = (x', y', w', h')$. The *overlap value* for A and B is an estimate for the distance between a true labeled and a detected IMO and is calculated as follows (see Fig. 9):

$$\frac{(X_+ - X_-)(Y_+ - Y_-)}{\max\{wh, w'h'\}}$$

where $X_+ = \max\{x + (w/2), x' + (w'/2)\}$, $X_- = \min\{x - (w/2), x' - (w'/2)\}$, $Y_+ = \max\{y + (h/2), y' + (h'/2)\}$, and $Y_- = \min\{y - (h/2), y' - (h'/2)\}$.

The defined overlap value is not a metric: it is symmetric $d(A, B) = d(B, A)$, and we have $d(A, A) = 0$, but $d(A, B) = 0$ does not mean that $A = B$, and this measure does not also satisfy the triangularity constraint $d(A, C) \leq d(A, B) + d(B, C)$, for sets of pixels A, B, and C. However, the overlap value is easy to calculate and is proved in our experiments to be a reasonable estimate. On the other hand, the cardinality of the symmetric difference between A and B divided by the cardinality of the union of both sets would be a metric (for a proof see [26]), but this measure is more costly to calculate.

Finally, the counts of hits, misses, or false alarms provide an evaluation of the quality and reliability of the detection. A detected IMO is considered true if the overlap value with a true IMO is less than or equal to $0 \cdot 5$, where equal to zero means that both circumscribing rectangles coincide.

We have developed a two-stage vision system for extracting driving-relevant information from stereo cameras that are mounted in a moving car (for details, see [5], [40], and [42]). For the gaze target identification, we provide the position of the gaze point within the frame and identity the gaze target as the ground truth. The position is given in normalized (x, y)-coordinates, ranging from (0, 0) in the lower left to (1, 1) in the upper right corner of a recorded frame. Gaze target identities were classified by a human observer into one of the 15 active target classes (e.g., lane markings on the right and tangent point on lane markings at the center of the road).

Fig. 10. Frame 875 of the Suburban Bridge sequence (left camera) shows two IMOs, for which hand-labeled data are provided. Furthermore, gaze positions are available, both in coordinates and classified targets. In the ego-vehicle, we observe the number plate of IMO 3 (black and white cross) and further data, also on IMO 4.

IMOs were hand labeled frame by frame and tracked across successive frames. For each IMO, the following parameters are given:

- the identification number;
- type (e.g., car, truck, motorcycle, bike, or pedestrian);
- a flag that indicates if the IMO is partially occluded (1) or not (0);
- the lane on which the IMO travels (i.e., lane 1 to denote the same as the test car, lane 2 for the opposite, lane 3 for side road left, and lane 4 for side road right).

With respect to the image frame, the IMO's center (x, y) and extension (width and height) are given. These data range on both axes from 0 to 1, with the origin (0, 0) being in the left lower corner. Fig. 10 shows one example frame from the Suburban Bridge sequence. On frame 875, IMOs 3 and 4 are shown, where IMO 3 partly occludes IMO 4.

Gaze point targets were classified by hand on a frame-to-frame basis into one of the 15 active and three error classes, i.e., lane markings on the left side, center, or right side of the road, boundary posts to either side of the road, tangent points on these lane markings, where applicable, the road surface ahead (i.e., on the first 20–30 m in front of the car) or farther away, street signs and traffic lights, and IMOs ahead (on the same lane), upcoming (on the opposite lane), or moving on cross roads. Gazes to any

other point (including the dashboard) were classified into the residual class, whereas errors were either associated with the start or the end of the recording or were classified as a general error. The predictive value of eye movements on car-directed actions by the driver has recently been demonstrated [24]. Next to the aforementioned IMO information, Fig. 10 shows the position of the gaze point (number plate of the first car).

Human drivers face a dual task. On the one hand, they need to steer the car through straight and curved sections of the roads, and for this part, they usually direct their gaze to the tangential point or the road surface (i.e., two points that allow drivers to infer the required steering angle by identifying simple geometric means); for example, see [24], [25], and [29]. On the other hand, drivers need to quickly attend to upcoming possible obstacles such as IMOs or points of interest such as crossings. Although there are a number of algorithms for segmenting the scene and identifying *salient* points of interest in a bottom–up manner (as aforementioned), the combination of saliency and relevance (top–down processes) into a priority map [13] is a subject of current research, and there are no databases available for benchmarking so far.

VIII. Conclusion

Rankings of stereo or motion correspondence methods often change along one sequence, and benchmarking on only a few frames of such a sequence is meaningless. Various events in outdoor driving define several challenges for stereo or motion matching, and oftentimes, all the methods experience difficulties with the same event, only at different scales. The importance of testing on such sequences is to identify such events and to aim at improving matching for those particular events. However, we have also imagined that an adaptive strategy may finally be best, selecting stereo or motion matching methods out of a given *toolbox* depending on automatically detected situations. For example, exposure balancing in cameras is already such an adaptation, which needs to be refined and expanded to further layers of data processing.

There are reasonable solutions for stereo analysis in outdoor environments, on sequences from moving cameras, but none of the techniques was superior in all the tested situations. Adaptive selection of techniques (from a toolbox, where available) would require time-efficient higher level mechanisms that identify situations.

A careful evaluation of stereo or motion algorithms (comparable to efforts when performing car crash tests for physical performance) requires testing on very large and representative data sets. Testing on data that represent very different traffic situations goes beyond common test behavior in the current computer vision community. There are more valuable sources for testing algorithms for driver assistance (e.g., the Daimler pedestrian benchmark data set by D. Gavrilla) or other traffic-related application areas (e.g., driver fatigue analysis), which also have their particular needs for test data.

Rendered data may be manipulated to simulate particular events. We have demonstrated that different stereo algorithms degenerate to a different degree in case of brightness differences or lighting artifacts, depending on parameters when generating these events.

A future statistical categorization of situations may be based on distributions of selected features in the Fourier domain of signals, on simple features such as mean intensity or variance in randomly selected windows, or on the density of significant scale-invariant features or locally adaptive regression kernels in randomly selected image rows. We have already tested scale-invariant features for this purpose, and a "sparse feature" approach appears to be reasonable for some clustering of image sequences into different categories.

Acknowledgment

The authors would like to thank the anonymous reviewers for valuable comments.

References

[1] A. Briassouli and I. Kompatsiaris, "Change detection for temporal texture in the Fourier domain," in *Proc. ACCV*, 2010, vol. LNCS, pp. 149–160.
[2] S. Baker, S. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szelisky, "A database and evaluation methodology for optical flow," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–8.
[3] J. L. Barron, D. J. Fleet, and S. S. Beauchemin, "Performance of optical flow techniques," *Int. J. Comput. Vis.*, vol. 12, no. 1, pp. 43–77, Feb. 1994.
[4] Basler Vision Technologies. [Online]. Available: http://www.baslerweb.com/
[5] E. Baseski, L. Baunegaard, N. Pugeault, F. Pilz, K. Pauwels, M. M. Van Hulle, F. Wörgötter, and N. Krüger, "Road interpretation for driver assistance based on an early cognitive vision system," in *Proc. VISAPP*, 2009, vol. 1, pp. 496–505.
[6] A. Bellmann, O. Hellwich, V. Rodehorst, and U. Yilmaz, "A benchmarking dataset for performance evaluation of automatic surface reconstruction algorithms," in *Proc. BenCOS*, 2007, pp. 1–8.
[7] S. Birchfield and C. Tomasi, "Depth discontinuities by pixel-to-pixel stereo," *Int. J. Comput. Vis.*, vol. 35, no. 3, pp. 269–293, Dec. 1999.
[8] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High-accuracy optical flow estimation based on a theory for warping," in *Proc. ECCV*, vol. 3024, *LNCS*, 2004, pp. 25–36.
[9] A. Bruhn, J. Weickert, and C. Schnörr, "Lucas/Kanade meets Horn/Schunck: Combining local and global optic flow methods," *Int. J. Comput. Vis.*, vol. 61, no. 3, pp. 211–231, Feb./Mar. 2005.
[10] Carnegie Mellon Univ. Image Data Base. [Online]. Available: http://vasc.ri.cmu.edu/idb/html/stereo/
[11] R. M. Dudley, "Probabilities and metrics: Convergence of laws on metric spaces, with a view to statistical testing," Matematisk Inst., Aarhus Univ., Aarhus, Denmark, Lecture notes series, Rep. 45, 1976.
[12] *.enpeda..* Image Sequence Analysis Test Site (EISATS). [Online]. Available: http://www.mi.auckland.ac.nz/EISATS/
[13] J. H. Fecteau and D. P. Munoz, "Salience, relevance, and firing: A priority map for target selection," *Trends Cogn. Sci.*, vol. 10, no. 8, pp. 382–390, Aug. 2006.
[14] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient belief propagation for early vision," *Int. J. Comput. Vis.*, vol. 70, no. 1, pp. 41–54, Oct. 2006.
[15] D. Gavrilla, Daimler pedestrian benchmark data set, 2009, follow "Looking at people". [Online]. Available: http://www.gavrila.net/Research/research.html
[16] S. Guan, R. Klette, and Y. W. Woo, "Belief propagation for stereo analysis of night-vision sequences," in *Proc. PSIVT*, vol. 5414, *LNCS*, 2009, pp. 932–943.
[17] P. Handschack and R. Klette, "Quantitative comparisons of differential methods for measuring of image velocity," in *Proc. Aspects Visual Form Process.*, Capri, Italy, 1994, pp. 241–250.
[18] R. Hartley and A. Zisserman, *Multiple-View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ. Press, 2000.
[19] R. Haeusler and R. Klette, "Benchmarking stereo data (not the matching algorithms)," in *Proc. DAGM*, 2010, pp. 383–392.
[20] S. Hermann and R. Klette, "The naked truth about cost functions for stereo matching," Univ. Auckland, Auckland, New Zealand, MI-tech-TR 33, 2009.

[21] H. Hirschmüller, "Accurate and efficient stereo processing by semi-global matching and mutual information," in *Proc. CVPR*, 2005, vol. 2, pp. 807–814.
[22] H. Hirschmüller and D. Scharstein, "Evaluation of stereo matching costs on images with radiometric differences," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 9, pp. 1582–1599, Sep. 2009.
[23] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artif. Intell.*, vol. 17, pp. 185–203, 1981.
[24] F. I. Kandil, A. Rotter, and M. Lappe, "Driving is smoother and more stable when using the tangent point," *J. Vis.*, vol. 9, no. 1, pp. 11:1–11:11, 2009.
[25] F. I. Kandil, A. Rotter, and M. Lappe, "Car drivers attend to different gaze targets when negotiating closed versus open curves," *J. Vis.*, vol. 10, no. 4, pp. 24:1–24:11, 2010.
[26] R. Klette and A. Rosenfeld, *Digital Geometry—Geometric Algorithms for Digital Picture Analysis*. San Francisco, CA: Morgan Kaufmann, 2004.
[27] R. Klette, S. Sandino, T. Vaudrey, J. Morris, C. Rabe, and R. Haeusler, "Stereo and motion analysis of long stereo image sequences for vision-based driver assistance," in *Proc. DAGM*, Jena, Germany, Sep. 9, 2009.
[28] V. Kolmogorov and R. Zabih, "What energy functions can be minimized via graph cuts?" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 2, pp. 147–159, Feb. 2004.
[29] M. F. Land and D. N. Lee, "Where we look when we steer," *Nature*, vol. 369, no. 6483, pp. 742–744, Jun. 1994.
[30] Z. Liu and R. Klette, "Dynamic programming stereo on real-world sequences," in *Proc. ICONIP*, vol. 5506, *LNCS*, 2008, pp. 527–534.
[31] Z. Liu and R. Klette, "Approximated ground truth for stereo and motion analysis on real-world sequences," in *Proc. PSIVT*, vol. 5414, *LNCS*, 2009, pp. 874–885.
[32] B. McCane, K. Novins, D. Crannitch, and B. Galvin, "On benchmarking optical flow," *Comput. Vis. Image Understand.*, vol. 84, no. 1, pp. 126–143, Oct. 2001.
[33] Middlebury Vision Evaluation. [Online]. Available: http://vision.middlebury.edu/
[34] T. Mimuro, Y. Miichi, T. Maemura, and K. Hayafune, "Functions and devices of Mitsubishi active safety ASV," in *Proc. IEEE Intell. Vehicles*, 1996, pp. 248–253.
[35] R. Mohan, G. Medioni, and R. Nevatia, "Stereo error detection, correction, and evaluation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 2, pp. 113–120, Feb. 1989.
[36] S. Morales and R. Klette, "A third eye for performance evaluation in stereo sequence analysis," in *Proc. CAIP*, vol. 5702, *LNCS*, 2009, pp. 1078–1086.
[37] S. Morales, T. Vaudrey, and R. Klette, "Robustness evaluation of stereo algorithms on long stereo sequences," in *Proc. IEEE Intell. Vehicles*, 2009, pp. 347–352.
[38] S. Morales, Y. W. Woo, R. Klette, and T. Vaudrey, "A study on stereo and motion data accuracy for a moving platform," in *Proc. FIRA RoboWorld Congr.*, vol. 5744, *LNCS*, 2009, pp. 292–300.
[39] Y. Ohta and T. Kanade, "Stereo by two-level dynamic programming," in *Proc. IJCAI*, 1985, pp. 1120–1126.
[40] K. Pauwels, "Computational modeling of visual attention: Neuronal response modulation in the Thalamocortical complex and saliency-based detection of independent motion," Ph.D. dissertation, Katholieke Univ. Leuven, Leuven, Belgium, 2008.
[41] Point Grey. [Online]. Available: http://www.ptgrey.com/
[42] N. Pugeault, K. Pauwels, F. Pilz, M. M. Van Hulle, and N. Krüger, "A three-level architecture for model-free detection and tracking of independently moving objects," in *Proc. Int. Conf. Comput. Vis. Theory Appl.*, 2010, pp. 237–244.
[43] Q. Ji, Z. Zhu, and P. Lan, "Real-time nonintrusive monitoring and prediction of driver fatigue," *IEEE Trans. Veh. Technol.*, vol. 53, no. 4, pp. 1052–1068, Jul. 2004.
[44] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vis.*, vol. 47, no. 1–3, pp. 7–42, Apr.–Jun. 2002.
[45] R. Szeliski, "Prediction error as a quality metric for motion and stereo," in *Proc. ICCV*, 1999, vol. 2, pp. 781–788.
[46] H. Takeda, S. Farsiu, and P. Milanfar, "Kernel regression for image processing and reconstruction," *IEEE Trans. Image Process.*, vol. 16, no. 2, pp. 349–366, Feb. 2007.
[47] N. A. Thacker, A. F. Clark, J. L. Barronc, J. R. Beveridged, P. Courtneye, W. R. Crum, V. Ramesh, and C. Clark, "Performance characterization in computer vision: A guide to best practices," *Comput. Vis. Image Understand.*, vol. 109, no. 3, pp. 305–334, Mar. 2008.
[48] T. Vaudrey, C. Rabe, R. Klette, and J. Milburn, "Differences between stereo and motion behavior on synthetic and real-world stereo sequences," in *Proc. Int. Conf. Image Vis. Comput. New Zealand*, 2008, pp. 1–6.
[49] T. Vaudrey, A. Wedel, and R. Klette, "A methodology for evaluating illumination artifact removal for corresponding images," in *Proc. CAIP*, vol. 5702, *LNCS*, 2009, pp. 1113–1121.
[50] A. Wedel, T. Pock, C. Zach, H. Bischof, and D. Cremers, "An improved algorithm for TV-L^1 optical flow," in *Statistical and Geometrical Approaches to Visual Motion Analysis*, D. Cremers, B. Rosenhahn, A. L. Yuille, and F. R. Schmidt, Eds. Berlin, Germany: Springer-Verlag, 2009, pp. 23–45.
[51] 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. Eur. Conf. Comput. Vis.*, 2008, pp. 739–751.
[52] X. Yang and R. Klette, "Approximate ground truth in the real world for testing optical flow algorithms," in *Proc. IEEE ICETCE*, 2011, pp. 6475–6478.
[53] C. Zach, T. Pock, and H. Bischof, "A duality-based approach for real time TV-L^1 optical flow," in *Proc. Pattern Recogn.—DAGM*, 2007, pp. 214–223.

Reinhard Klette is a Professor with the Computer Science Department, University of Auckland, Auckland, New Zealand, and the Head of a research group that works on projects for computer-vision-augmented vehicles (e.g., vision-based driver assistance). He is currently on the editorial board of the *International Journal of Computer Vision*. He has (co)supervised about 20 Ph.D. students and about 100 M.Sc. students, and he has more than 250 publications in peer-reviewed journals or conference proceedings. He is a coauthor of books on computer vision, image processing, geometric algorithms, and panoramic imaging.

Dr. Klette was an Associate Editor for the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE from 2001 to 2008. He was invited to present keynotes at about 20 international conferences worldwide.

Norbert Krüger received the M.Sc. degree from the Ruhr-Universität Bochum, Bochum, Germany, and the Ph.D. degree from the University of Bielefeld, Bielefeld, Germany.

He is currently a Professor with the Mærsk McKinney Møller Institute, University of Southern Denmark, Odense, Denmark, where he also leads the Cognitive Vision Laboratory, which focuses on computer vision and cognitive systems, particularly the learning of object representations in the context of grasping. He has also been working on computational neuroscience and machine learning.

Tobi Vaudrey received the Ph.D. degree from the University of Auckland, Auckland, New Zealand, in 2011. His Ph.D. dissertation was on robust dynamic vision from a moving platform.

He is currently with the University of Auckland. His research interests include vision-based algorithms, particularly optical-flow and stereo matching, focusing on making these approaches real time and robust to real-world issues. His main research focus is on the fusion of stereo and optical flow into one framework (scene flow) and making this approach real time and real-world applicable.

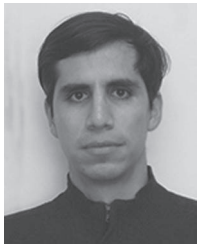
Karl Pauwels received the M.Sc. degree in commercial engineering, the M.Sc. degree in artificial intelligence, and the Ph.D. degree in medical sciences from the Katholieke Universiteit Leuven (KULeuven), Leuven, Belgium.

He currently holds a postdoctoral post with the Laboratorium voor Neuro- en Psychofysiologie, Faculty of Medicine, KULeuven. His research interests include optical flow, stereo and camera motion estimation in the context of real-time computer vision.

Marc van Hulle received the M.Sc. degree in electrotechnical engineering (electronics) and the Ph.D. degree in applied sciences from the Katholieke Universiteit Leuven (KULeuven), Leuven, Belgium, the D.Tech. degree from the Technical University of Denmark in 2002, and the honorary Ph.D. degree from the Brest State University, Brest, Belarus, in 2009.

He is currently a Full Professor with KULeuven, where he heads the Computational Neuroscience Group, Laboratorium voor Neuro- en Psychofysiologie, Faculty of Medicine. His research interests include computational neuroscience, neural networks, computer vision, data mining, and signal processing.

Dr. Hulle received the SWIFT Prize from the King Baudouin Foundation of Belgium in 2009.



Sandino Morales received the B.Sc. degree in mathematics from the Universidad Nacional Autónoma de México, México City, México, in 2004 and the M.C.S. degree in automatic control systems from the Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, México City, in 2007. He is currently working toward the Ph.D. degree with the University of Auckland, Auckland, New Zealand.

His research interests include stereo vision algorithms and the evaluation of their performance.

Farid I. Kandil received the M.Sc. degree in psychology and the Ph.D. degree in biology from the University at Bremen, Bremen, Germany.

He subsequently held a postdoctoral post with the University College London, London, U.K., and the University of Muenster, Germany. For one year, he served as a Clinical Neuropsychologist. He is currently with the University of Muenster, Germany. His research includes mid-level visual processing, e.g., figure-ground segregation and higher order motion perception, and the use of both approaches in human car driving.

Ralf Haeusler received the M.Sc. degree from Jena University, Jena, Germany. He is currently working toward the Ph.D. degree with the University of Auckland, Auckland, New Zealand, with the *.enpeda..* Group, Tamaki Innovation Campus.

Nicolas Pugeault received the M.Sc. degree from the University of Plymouth, Plymouth, U.K., in 2002 and the M.Eng. degree from the Ecole Superieure d'Informatique, Electronique, Automatique, Paris, France, in 2004, and the Ph.D. degree from the University of Göttingen, Göttingen, Germany, in 2008.

For the two following years, he was subsequently a Research Associate with the University of Edinburgh, Edinburgh, U.K., and an Assistant Professor with the University of Southern Denmark, Odense, Denmark. Since 2009, he has been a Research Fellow with the Centre for Vision, Speech, and Signal Processing, University of Surrey, Surrey, U.K. He is the author or a coauthor of more than 30 technical publications, conference proceedings, editorials, and books.

Clemens Rabe was born in Giessen, Germany, in 1979. He received the Diploma degree in computer engineering from the University of Applied Sciences Wuerzburg, Schweinfurt, Germany, in 2005 and the Ph.D. degree in engineering from the University of Kiel, Kiel, Germany, in 2011.

He is now with the Environment Perception Group, Daimler Research, Sindelfingen, Germany. His research interests include stereo vision, motion analysis, and traffic scene understanding.

Markus Lappe received the Ph.D. degree in physics from the University of Tübingen, Tübingen, Germany.

He conducted research work on computational and cognitive neuroscience of vision with the Max Planck Institute for Biological Cybernetics, Tübingen, the National Institutes of Health, Bethesda, MD, and the Department of Biology, Ruhr University Bochum, Bochum, Germany. Since 2001, he has been a Full Professor of experimental psychology and a Member of the Otto Creutzfeldt Center for Cognitive and Behavioral Neuroscience, University of Muenster, Germany.