Towards Benchmarking of Real-World Stereo Data

Ralf Haeusler, Sandino Morales, Simon Hermann, and Reinhard Klette
The *.enpeda..* Project
The University of Auckland, New Zealand

***Abstract*—The paper proposes the prediction of stereo matching performance based on analyzing the given stereo data (and not based on test runs of stereo matching algorithms). For justifying our approach we compare results obtained by prediction error analysis (for different stereo matching algorithms) with three different data evaluation techniques: a count of SIFT matches, a mismatch count between census transform features, and the quality of dense optical flow fields based on a total-variation energy minimization. The paper shows that there are reasonable indications that such measures, quantifying matches of features or image regions, correlate with stereo performance to some degree. This study on data evaluation is initiating a new direction of research, and it concludes with the suggestion of studying further measures or more data for the ultimate goal of supporting an adaptive optimization or selection of stereo matching techniques with respect to given image data.**

I. Introduction

Quantitative evaluation of stereo methods supports progress in improving the performance of stereo matching methods. Often, the set of data used for testing algorithms remains extremely limited, compared to needs in real-world applications such as vision-based driver assistance [6], where optimization of stereo matching needs to be adaptive to the recorded *situations*, defined by various events in traffic scenes such as lighting, density of traffic, road geometry, and so forth. A few image samples cannot represent the diversity of possible stereo recordings, enforced by the fact that used data are often either synthetic or engineered, such of a quality that is not really challenging for the top performing stereo algorithms.

Reported differences in the performance of stereo methods for a few image samples, such as on [8], are not applicable when selecting matching techniques for more diversified sets of stereo data. For example, for one of the most popular images used for testing stereo algorithms, `Teddy`, we observed the following: When comparing generated depth maps to the provided ground truth using normalized cross correlation, the simple sum of absolute differences (SAD) block matching method performed almost as good as computationally expensive global stereo methods. Obviously, this SAD method is not a competitor when it comes to real world data.

The paper introduces quantitative measures that aim at expressing the extent to which stereo matching methods are challenged by a particular set of stereo data. A SIFT-based measure was proposed and studied in [5]. Here we add two more measures to the discussion.

For being absolutely clear here: We do not aim at finding a measure to answer "How good is a particular algorithm?", but we are interested in the question "How difficult are particular stereo data?". Obviously, there is no lower limit for the quality of stereo data. For example, even the powerful human visual system can fail on some data (e.g., in case of major contrast differences between left and right views; we tested limits of human stereo fusion by viewing different sets of recorded stereo data on 3D displays using polarized glasses). The practical relevance of evaluating data lies in the need to adapt stereo matching algorithms to changes in recorded stereo data. Failure due to bad input data (e.g., driving in the night in heavy rain) is not acceptable, for example, for vision-based driver assistance. At least, the system needs to understand that available methods cannot cope with the current data complexity.

Prediction error analysis [15] was applied in [10] for stereo matching on large datasets of trinocularly recorded video sequences. Sparse matching statistics was used in [5] for data evaluation. In this paper we compare the technique of prediction error analysis with sparse matching statistics. The only prerequisite is that we assume rectified stereo input pairs.

For defining measures that predict stereo performance, depending only on the quality of available stereo data and not on implicit assumptions (inherent to particular stereo methods), the approach in [5] quantified performance of sparse SIFT matching [7] between rectified stereo frames without constraining matching by epipolar geometry. In fact, consistency of matches with epipolar geometry was used to evaluate those matches. Matching-based measures in this paper include the previously proposed SIFT-feature technique, but also census transform signatures [14] and the plausibility of dense motion fields, calculated as TV-L_1 optical flow [2]. (This optical flow method performed well in general on traffic-related video sequences [17]; however, its standard sequential implementation is still computationally expensive and therefore, in this implementation, not yet suitable for realtime applications.)

Section II reviews the application of prediction error analysis for stereo evaluation. Section III introduces matching with SIFT and census signatures. Section IV explains how these approaches can be applied to estimate stereo frame quality. Section V presents our experiments and discusses the results. Section VI concludes.

II. Prediction error analysis

For evaluating stereo algorithms on real-world images, an application of a prediction error technique [15] was proposed in [10], using three time-synchronized and calibrated cameras for recording "long" (i.e., 100 to 300 stereo frames) image sequences. Two of those sequences were the input data for

(a) Reference image (b) Match image (c) Control image (d) Disparity map (e) Warped match image

Fig. 1. Illustration for prediction error analysis. The disparity map is used for creating a virtual view for the right (i.e., control) camera. This virtual view is compared with the actually recorded right camera image.

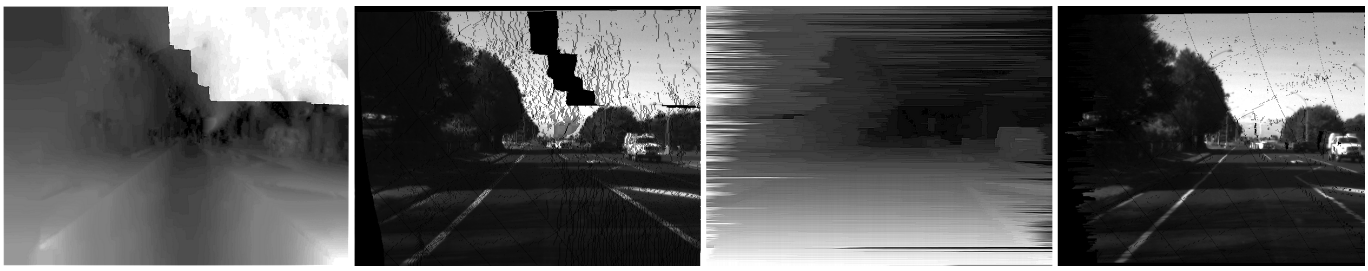


Fig. 2. Disparity maps and corresponding virtual views for the dataset 092110 (BP on the left, and DP on the right).

various stereo algorithms, and the third sequence was used as "ground truth". We briefly recall this approach.

Rectified images from two cameras (the reference camera C_r and the matching camera C_m, respectively) are selected as input data for a given stereo matching algorithm. Images recorded with a third camera (the control camera C_c) are used to evaluate the disparity map D as calculated with a stereo algorithm. Using the disparities in D, and the calibration data between C_r and C_c, images recorded with C_r are warped into a virtual image that predicts what would actually be recorded with C_c. See Figs. 1 and 2.

The evaluation is carried out by comparing the virtual image with the corresponding image of C_c using normalized cross correlation (NCC), what is a common choice for a *similarity measure*. Let I_c be an image recorded with C_c and I_v the corresponding virtual image. The NCC between them is defined as follows:

$$S(I_c, I_v) = \frac{1}{|\Omega|} \sum_{(x,y)\in\Omega} \frac{[I_c(x,y) - \mu_c][I_v(x,y) - \mu_v]}{\sigma_c \sigma_v} \quad (1)$$

where μ_c and μ_c denote the means, and σ_c and σ_v the standard deviations of I_c and I_v, respectively. Ω is the set of pixels showing non-occluded points (i.e., seen by the three cameras); $|.|$ is the cardinality.

The main advantage of this evaluation approach is that performance evaluation of stereo algorithms can be done objectively using a broader class of data sets, as no ground truth of disparity data is required.

In this work, we analyze the performance of some stereo algorithms using this approach of prediction error analysis, while analyzing the proposed measures of data complexity at the same time on the trinocular sequences.

III. Matching techniques

We apply SIFT and census transform features. Interest points of SIFT features [7] are defined by extrema in a difference-of-Gaussians scale-space, also applying sub-pixel accuracy and rejection of poorly defined locations. Together with a "well-defined" descriptor, these are called *distinctive features*.

A SIFT descriptor is defined by image gradients at different locations around the interest point. Neighboring pixels are summarized in orientation histograms. The original implementation of SIFT uses 16×16 descriptor arrays. These are summarized in 4×4 histogram bins, each containing magnitudes for eight different orientations. Thus, a descriptor vector is of dimension 128.

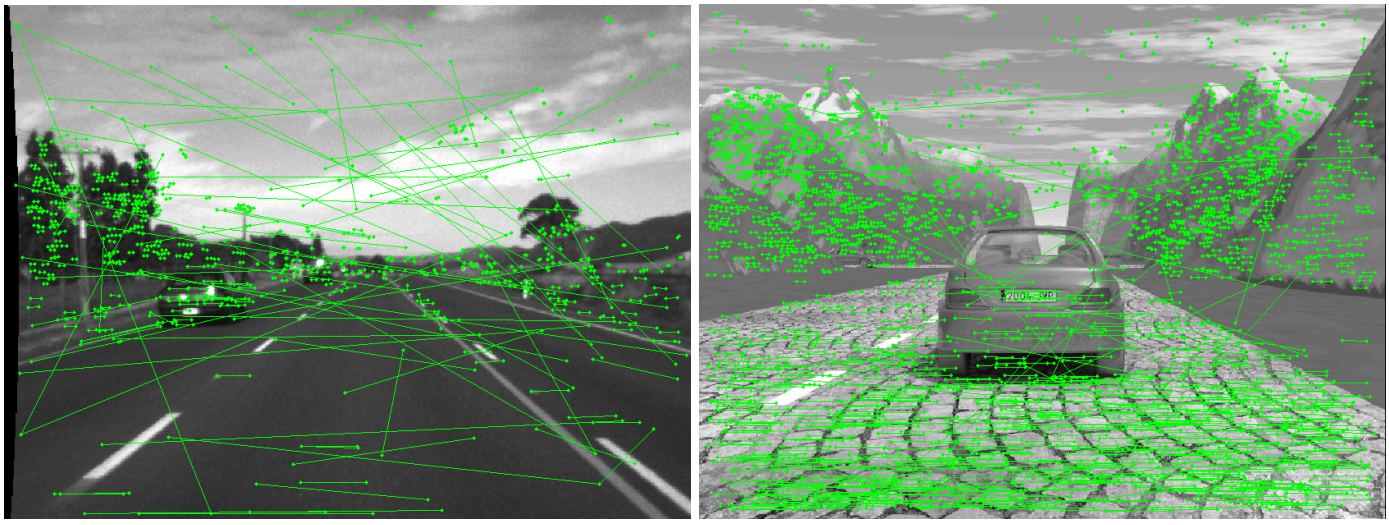
Matching can be done by choosing corresponding vectors according to their Euclidean distance, usually involving some heuristics to exclude unreliable matches. Best SIFT matches in stereo frames are illustrated in Fig. 3.

This sketched SIFT process, and the dimension of the descriptor vector make this method computationally expensive for deriving matching statistics and less suitable for realtime applications than, for example, the following census transform.

The census transform is successfully used in correspondence analysis [18], for example, for calculating sparse optical flow in realtime [14]. For defining the census transform descriptor (also called *signature*), consider a (e.g., square) neighborhood $\mathcal{N}(p)$, for pixel p. For $q \in \mathcal{N}(p)$, let

$$\xi(p,q) = \begin{cases} 1 & I(q) < I(p) \\ 0 & I(p) \leq I(q) \end{cases}$$

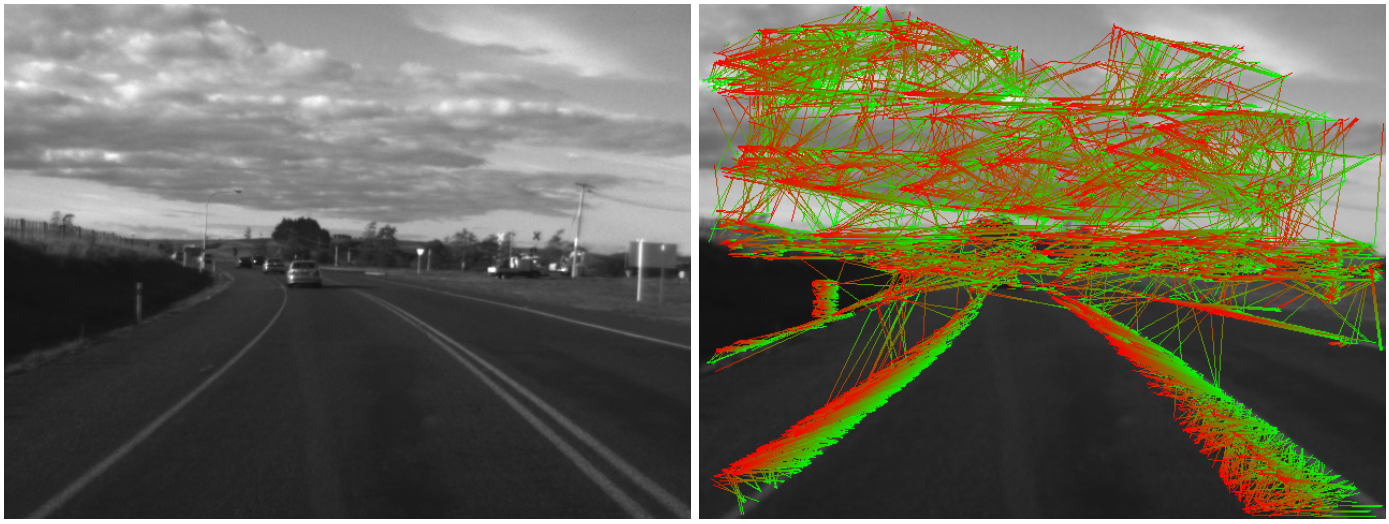
A bitstring (i.e., the signature) is generated by concatenation of all $\xi(p,q)$ values when traversing $\mathcal{N}(p)$ in a specified order.



(a) Best SIFT matches on a recorded stereo frame

(b) Best SIFT matches on a synthetic stereo frame

Fig. 3. Both images show the left image of a stereo frame, and best matches in the right image are projected into the left image and connected by a line segment. Best matches which are inconsistent with epipolar geometry are not "on the same image row". Synthetic images tend to have very few of those inconsistent matches.



(a) Real world input image

(b) Census flow

Fig. 4. Census-based matches may be numerous in some of the image regions; correct matches are here superimposed by many incorrect ones.

The matching measure between two pixels is the Hamming distance of their corresponding bitstrings.

For optical flow calculation, ternary signatures are used in [18], defined as follows for some small ε:

$$\xi(p,q) = \begin{cases} 0 & I(p) - I(q) > \varepsilon \\ 1 & |I(q) - I(p)| \leq \varepsilon \\ 2 & I(q) - I(p) > \varepsilon \end{cases}$$

The distance between ternary signatures can be computed by the number of positions with differing values at that position.

Matching is dense when applying those signatures (i.e., correspondences are sought for every pixel), but without having an option for subpixel accuracy. Initial matches are calculated by a cascade of filters that remove (e.g.) "poorly defined" matches [18]; this results in sparse matching. An example for such a matching process (using ternary signatures) is shown in Fig. 4.

IV. Matching Statistics and Data Quality

At this point it is just our hypothesis that sparse matching of distinctive features may provide a measure that is strongly correlated to the outcome dense stereo matching processes (say, "scaled by the quality of the actually applied matching algorithm"). However, this hypothesis (i.e., the existence of such a correlation) was already supported by results in [5].

To check the correctness of a match in a rectified stereo frame, we do the following: A match between a feature location (i_l, j_l) in the left image and a feature location (i_r, j_r)

(a) Reference image (b) Match image (c) TV-L1 flow for reference image

Fig. 5. Applying dense flow to a stereo frame of suboptimal quality. All areas that are not light blue or not colored contribute to our error measure.

in the right image is considered to be *correct up to known constraints* if

$$|j_l - j_r| < \varepsilon \ \wedge \ i_r \in [i_l, i_l - d_{max}] \tag{2}$$

for a chosen small $\varepsilon > 0$ (we choose $\varepsilon = 1$); d_{max} is the maximum disparity between both stereo views.

Equation (2) appears to be very much "forgiving". However, the probability of a misclassification is at most $(2\varepsilon d_{max} - 1)/(M \cdot N)$ in an image of size $M \times N$. This assumption may be violated, for example, for images with repetitive textures in some areas.

When applying SIFT features, we use the counts of correct (up to known constraints) matches and the ratio between detected and matched features for given stereo pairs. If the feature detector identifies n features in the base image that lead to m matches in the match image, and that from those m matches, o are classified as being incorrect (up to known constraints), then we define that

$$x = n/m \qquad \text{and} \qquad y = 100 \times o/m \tag{3}$$

where x is the *matching rate* and y the *mismatch rate*. Thus, $x \geq 1$ and $y \leq 100\%$.

The matching rate x expresses how many features on average lead to one match (no matter whether correct or not), while the mismatch rate y identifies the percentage of incorrect (up to known constraints) matches.

For the much faster census transform matching process, we only count the mismatch rate y, also referring to Eq. (2).

Estimating stereo performance based on dense (TV-L_1) optical flow maps is done as follows: We assume that a motion vector has direction a, with a value between 0 and 2π, and length $b \in [0, 1]$. Ideally, the direction a should always be π (i.e., horizontal). However, for vectors with length b close to zero, the direction is meaningless and should not contribute (much) to an error measure. Thus, for each pixel $p = (i, j)$ we calculate the magnitude of an error at p as follows:

$$e_{ij} = b^2 \, |a - \pi| \tag{4}$$

The total error E for an image equals

$$E = \sum_{(i,j)} e_{ij} \tag{5}$$

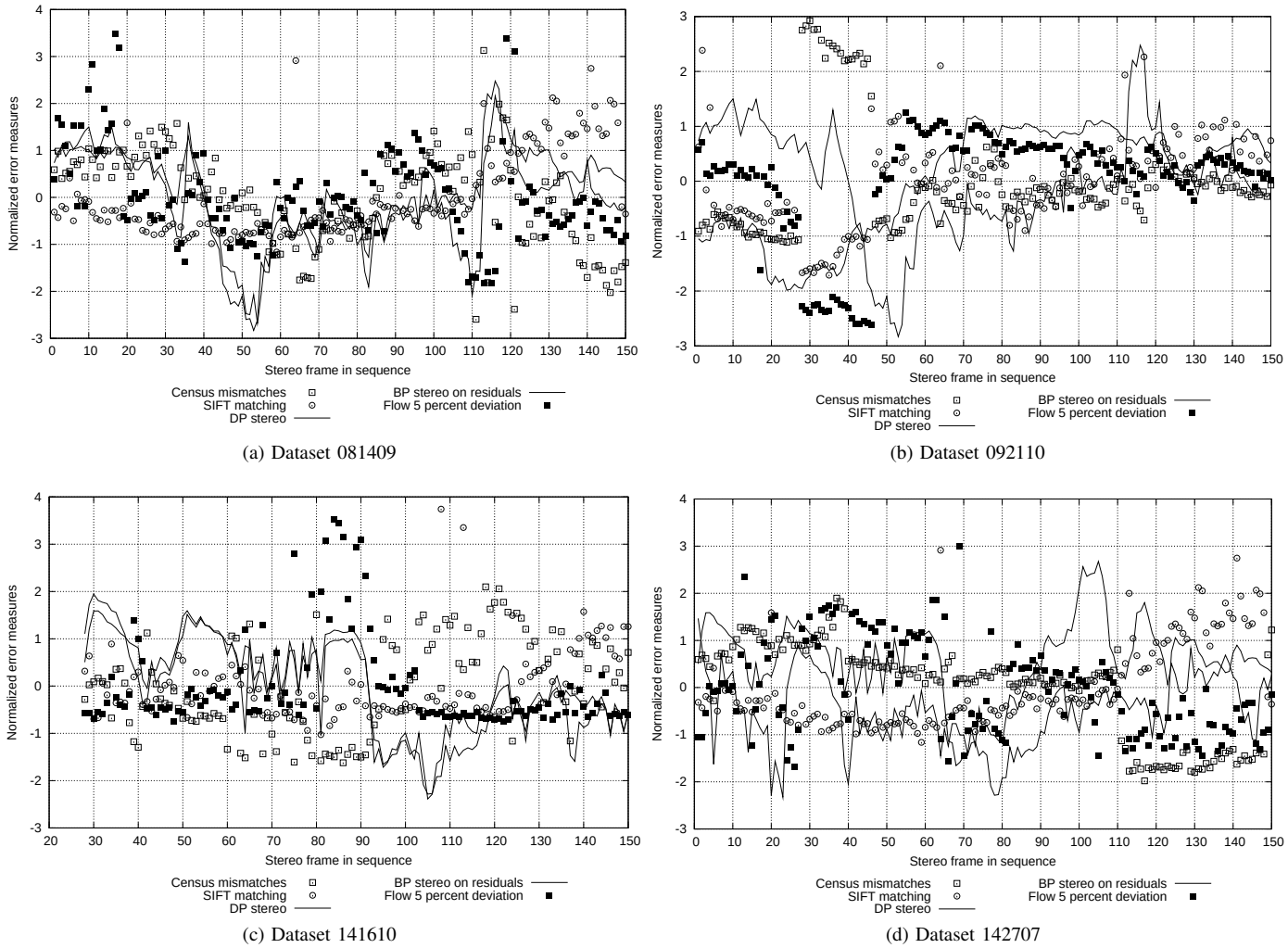
V. Experiments

We compare stereo performances on four different datasets, each consisting of up to 150 trinocular frames with known intrinsic and extrinsic calibration parameters. For each of these trinocular frames, we compute the prediction error as outlined in Section II using dynamic programming stereo (DP) [12], and belief propagation stereo (BP) [4]; BP is on residual images [1]. Furthermore, we compute SIFT and census based correspondences and apply the matching count as outlined in Section IV. For SIFT, we use the value $x + y$ as a resulting scalar measure. Finally, we also calculate dense flow using TV-L_1 optical flow [2] (illustrated in Fig. 5) and use the error measure described at the end of the last section. Each of these values is plotted in Fig. 6 for the four datasets.

To enhance compatibility between different measures, we plot normalized values: Let T be the number of frames, $E(t)$ be the error measure for frame at time t. For $1 \leq t \leq T$, we

TABLE I
Summarizing cross correlation values.

	Method	SIFT	DP	BP	TV-L1
Dataset 081409	Census	-0.25	0.34	0.17	0.25
	SIFT		0.30	0.42	-0.12
	DP			0.85	0.33
	BP				0.38
Dataset 092110	Method	SIFT	DP	BP	TV-L1
	Census	-0.50	-0.01	-0.27	-0.78
	SIFT		-0.17	0.49	0.57
	DP			-0.14	-0.12
	BP				0.66
Dataset 141610	Method	SIFT	DP	BP	TV-L1
	Census	0.02	-0.36	-0.54	-0.67
	SIFT		-0.16	-0.14	-0.22
	DP			0.93	0.20
	BP				0.36
Dataset 142707	Method	SIFT	DP	BP	TV-L1
	Census	-0.79	-0.27	0.08	0.50
	SIFT		0.13	0.12	-0.52
	DP			-0.01	-0.21
	BP				0.12



(a) Dataset 081409

(b) Dataset 092110

(c) Dataset 141610

(d) Dataset 142707

Fig. 6. Four datasets, each with five different error measures. For better visual comparison, we display normalized values for all error measures.

display $E_n(t)$ in our charts, with

$$E_n(t) = (E(t) - \mu_T)/\sigma_T \tag{6}$$

for mean μ_T and variance σ_T^2 of $E(t)$ on those T frames.

We also compute cross correlations between these series of values for each dataset. For example, consider normalized values $E_{nBP}(t)$ for BP prediction errors, and normalized values $E_{nDP}(t)$ for DP prediction; then the cross correlation equals $1/T \sum_{t=1}^{T} E_{nBP}(t)E_{nDP}(t)$.

Table I provides these cross correlation values for all combinations of error measures E, and for all four datasets. The complete graphs of error measures are shown in Fig. 6.

The tables of cross correlation values indicate that there is only a minor correlation between summarized errors on whole sequences, except stronger correlations between DP and BP prediction error results on the first two datasets (i.e., 081409 and 092110).

We discuss the obtained results qualitatively, for each dataset separately.

1) Dataset 081409: There is an obvious divergence of stereo and matching results for Frames 45 – 56. Matching or flow results do not indicate an expected drop in performance, although prediction error analysis does for the actual stereo results. In fact, visual inspection of these frames did not tell us about a particular issue with image quality in those frames. Frames 110 – 114 are the only ones that appeared problematic in this sequence. (The car is passing below a highway, leading to strong changes in exposure.) Prediction errors detect problems only on residual images. Optical flow and census matching detected this "critical" situation, while SIFT-matching did not.

2) Dataset 092110: The sequence contains major brightness differences between reference and match images, due to shadows caused by trees along the road. Visual inspection verifies good stereo image quality only for Frames 28-46, what is noticed by the census measure, but not by any of the others. Even the results of prediction error analysis are contradictory for the whole sequence.

3) Dataset 141610: The whole sequence is prone to severe reflections on the windscreen of the test vehicle, except for Frames 49 – 51 and Frames 120 – 125. For all of the proposed methods, the good (i.e., by visual inspection) quality of these few frames goes unnoticed.

4) Dataset 142707: In this sequence, major brightness differences are starting at Frame 110, and all previous frames appear (i.e., by visual inspection) to be of satisfactory quality. Changes in the curves in Fig. 6 (for this dataset) are not related to variations in image quality.

VI. Conclusion

The undertaken studies confirm the initial findings in [5] that measures for evaluating stereo data sets verify different stereo matching complexities for different data sets (in short: evaluations of stereo matchers on one data set may not be of relevance for estimating performance on another dataset). However, correspondences between the performance of stereo matching techniques and proposed data complexity measures still needs to be understood for challenging real-world sequences, as used as examples in this paper.

In [5] it was shown that *there is* a correspondence between challenges in stereo matching when using either synthetic or engineered stereo pairs, or real world sequences. In this paper we attempted to go one step further: leaving synthetic or engineered data aside, how about having some convincing measures for the huge diversity of real-world stereo sequences? We have proposed measures in this paper, and understand that any single measure of those is still insufficient to cope with the complexity of real-world data. This "one step further" proved to be a step into a new dimension of the problem.

Not surprisingly, the evaluation of stereo methods without having ground truth about disparities is an extremely challenging task. The prediction error analysis technique [9] is a promising way to evaluate stereo techniques. This paper is on evaluating stereo data for adding another option for evaluating stereo techniques, and the proposed measures appear to be "somehow" useful for indicating "critical" frames, but do not succeed in general, and also do not "behave" in a way that visual checks of data are consistent with the calculated measures.

The general intention is to design an automatic system to isolate potential sources of stereo matching errors (i.e., to identify "interesting" scenarios for further improvements of stereo matching techniques). The decoupling of disruptive influences (also called "events" at the beginning of this paper) in recorded scenes is probably impossible without having first a comprehensive analysis of the recorded scene. But studies on more advanced stereo data complexity measures (e.g., by combining the proposed measures) are certainly needed to obtain further clarity on this issue.

References

[1] Aujol, J.F., Gilboa, G., Chan, T., Osher, S.: Structure-texture image decomposition - modeling, algorithms, and parameter selection. *Int. J. Computer Vision*, **67**:111–136 (2006)

[2] Brox, T., Bruhn, A., Papenberg, N., Weickert, J.: High accuracy optical flow estimation based on a theory for warping. In *ECCV*, volume 4, LNCS 3024, pages 25–36 (2004)

[3] EISATS (*.enpeda..* Image Sequence Analysis Test Site): www.mi.auckland.ac.nz/EISATS.

[4] Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient belief propagation for early vision. *Int. J. Computer Vision*, **70**:41–54 (2006)

[5] Haeusler, R. , Klette, R.: Benchmarking stereo data (not the matching algorithms). In *DAGM*, to appear (2010)

[6] Klette, R., Sandino, S., Vaudrey, T., Morris, J., Rabe, C., Haeusler, R.: Stereo and motion analysis of long stereo image sequences for vision-based driver assistance. Keynote, DAGM, Jena/Germany (2009)

[7] Lowe, D. G.: Distinctive image features from scale-invariant keypoints. *Int. J. Computer Vision*, **60**:91–110 (2004)

[8] Middlebury College Stereo Vision Page. vision.middlebury.edu/stereo/

[9] Morales, S., Klette, R.: A third eye for performance evaluation in stereo sequence analysis. In *CAIP*, LNCS 5702, pages 1078–1086 (2009)

[10] Morales, S., Vaudrey, T., Klette, R.: Robustness evaluation of stereo algorithms on long stereo sequences. In *IEEE Intelligent Vehicles Symp.* pages 347 – 352, (2009)

[11] Morales, S., Woo, Y., Klette, R., Vaudrey, T.: A study on stereo and motion data accuracy for a moving platform. In *FIRA RoboWorld Congress Advances Robotics*, pages 292 – 300 (2009)

[12] Ohta, Y., Kanade, T.: Stereo by two-level dynamic programming, in *Int. Joint Conf. Artificial Intelligence*, pages 1120–1126 (1985)

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

[14] Stein, F.: *Efficient computation of optical flow using the census transform.* In *DAGM*, LNCS 3175, pages 79–86 (2004)

[15] Szeliski, R.: Prediction error as a quality metric for motion and stereo, In *IEEE Int. Conf. Computer Vision*, pages 781–788 (1999)

[16] Vedaldi, A., Fulkerson, B.: VLFeat: An open and portable library of computer vision algorithms. www.vlfeat.org (2008)

[17] Yang, X., Klette, R.: Evaluation of motion analysis on synthetic and real-world image sequences. MItech-TR 58, Multimedia Imaging, The University of Auckland (2010)

[18] Zabih, R, Woodfill, J.: Non-parametric local transforms for computing visual correspondence. In *ECCV*, LNCS 801, pages 151–158 (1994)