Third-Eye Stereo Analysis Evaluation Enhanced by Data Measures

Verónica Suaste[1], Diego Caudillo[1], Bok-Suk Shin[2], and Reinhard Klette[2]

[1] CIMAT and the University of Guanajuato, Mexico
[2] The *.enpeda..* Project, The University of Auckland, New Zealand

Abstract. Third-eye stereo analysis evaluation compares a virtual image, derived from results obtained by binocular stereo analysis, with a recorded image at the same pose. This technique is applied for evaluating stereo matchers on long (or continuous) stereo input sequences where no ground truth is available. The paper provides a critical and constructive discussion of this method. The paper also introduces data measures on input video sequences as an additional tool for analyzing issues of stereo matchers occurring for particular scenarios. The paper also reports on extensive experiments using two top-rated stereo matchers.

1 Introduction

Modern applications of stereo analysis require that stereo matchers work accurately on long or continuous binocular input video data. For example, in vision-based driver assistance, those data are recorded for any possible traffic scenario [9]. Robust matchers need to work accurately for various scenarios. In general it is expected that there is no single best matcher; an adaptive selection of a matcher (within a given 'toolbox') appears to be a possible solution.

The third-eye method of [11] provides stereo analysis performance evaluation for long or continuously recorded stereo sequences. For a current application of this method, see [12]. We provide in this paper a critical and constructive discussion of this method, pointing to weaknesses and also outlining ways how to overcome those. Video data measures are used to discuss solutions and to propose ways for a detailed analysis of situations where a stereo matcher fails (and should be improved accordingly), extending our initial discussion of data measures in [10].

For testing, the eight long trinocular stereo sequences of Set 9 on EISATS [4] have been used (each 400 stereo frames long, except the 'People' sequence which is only 234 frames long); see Fig. 1. The tested stereo matchers are *iterative semi-global matching* (iSGM) [7] and *linear belief propagation* (linBP) [10]. Both apply the census transform as the data cost function, and linBP uses a truncated linear smoothness constraint [5]. Both stereo matchers, iSGM and linBP, rank high on the KITTI stereo benchmark suite (`www.cvlibs.net/datasets/kitti/`).

The paper is structured as follows: Section 2 provides used notations and definitions. Section 3 illustrates interesting cases when using the third-eye approach. Section 4 discusses the use of data measures for solving critical cases and for discussing stereo performance more in detail. Section 5 concludes.

Fig. 1. Examples of frames of the eight sequences in Set 9. Upper row, left to right: sequences called 'Barriers', 'Bridge', 'Dusk', and 'Midday'. Lower row, left to right: 'Night', 'People', 'Queen', and 'Wiper'.

2 Fundamentals

The third-eye method [11] requires that three calibrated cameras record time-synchronized the same scene. In case of Set 9 on EISATS [4], the cameras are placed on a bar on the left, center, and right position behind the windscreen of the *ego-vehicle* (i.e. the car the stereo-matcher is operating in). Two of the images, the center or reference image, and the right or match image, are used to calculate a disparity map by the chosen stereo matching algorithm. The disparity map is used to map all the pixels in the reference image into that position in the left or control image where the pixel value would be visible from the pose of the left camera. This calculated *virtual image* V is then compared with the control image C, for example by using the normalized cross correlation (NCC) index used as a quality measure:

$$M_{NCC}(V,C) = \frac{1}{|\Omega|}\Sigma_{(x,y)\in\Omega}\frac{[V(x,y)-\mu_V][C(x,y)-\mu_C]}{\sigma_V\sigma_C} \tag{1}$$

The domain Ω contains only those pixels which are successfully mapped from the reference image into the domain of the virtual image (i.e. non-occluded pixels); μ and σ represent mean and standard deviation of the corresponding images.

Due to possibilities of a misleading influence of homogeneous intensity regions, [6] suggested to use a further restricted set Ω which only contains pixels locations which are in distance of 10 pixels or less to an edge pixel in the reference image. We use this modified NCC-mask measure, called M_{NCC_Mask}, as our standard measure for the third-eye approach.

Having stereo sequences of length 400 (or 234 in one case) in our test data, the measure M_{NCC_Mask} produces a real-valued function for each used stereo matcher on such a sequence. We also define data measures on input data of one of the cameras (e.g. variance of intensities or of Sobel edge values), or by comparing images recorded by two of the cameras (e.g. M_{NCC} between left and center image). Those data measures also define real-valued functions, and they are motivated as follows:

Homogeneous images are a difficult case for stereo matching, thus we considered the variance of intensity values in reference and match image. Typically those variances of reference and match image are about the same, so we only use the standard deviation $Sigma\_left$ of the reference image.

Stereo matching is supported by having image features such as edges or corners. There are complex edge detectors such as that of [2], or very simple edge detectors such as the Sobel operator. For avoiding a bias introduced by the edge detector we use the simple Sobel operator and measure the standard deviation $Sigma\_Sobel$ of operator values $|C_x| + |C_y|$ on control image C.

An important assumption for stereo matching is that both images are captured in the same environment, with just small changes due to a minor variation in view point or viewing direction. For example, see [3] where this is discussed for stereo vision in astrophysics. We decided to use the above NCC measure between reference and match image as our third data measure $NCC\_leftright$.

We intend to compare two of such real-valued functions having the same domain, not in a rigorous sense of direct analysis, but with respect to the curves 'behavior', such as the distribution of local minima or maxima. For discussions about special kinds of, or comparisons between functions, see, for example, [1,3,8,14]. One option is to study or compare the derivatives of the functions. But, working in discrete domains, that implies a need to choose a neighborhood of some size and a method for approximating derivatives.

Thus we decided to keep one function f fixed as a reference, and to apply a transformation to the other functions g which allows us to define an analytical distance between the new function g_{new} and f, thus defining an alternative relationship between functions. The defined distance will not be a metric in the mathematical sense because we do not aim at symmetry, and the distance between two different functions (e.g. two constant functions with different values) can be zero when applying our distance measure.

Let μ_f and σ_f be the mean and standard deviation of function f. Given are two real-valued functions f and g with the same discrete domain and non-zero variances. Let

$$\alpha = \frac{\sigma_g}{\sigma_f}\mu_f - \mu_g \quad \text{and} \quad \beta = \frac{\sigma_f}{\sigma_g}$$
$$g_{new}(x) = \beta(g(x) + \alpha) \tag{2}$$

As a result, g_{new} has the same mean and the same variance as function f. Now we define our distance function in the common L_1 way, as, for example, described by [1]:

$$d_1(f_1, f_2) = \int |f_1(x) - f_2(x)| \, \mathrm{d}x \tag{3}$$

Our distance is then defined by $d(f, g) = d_1(f, g_{new})$. Indeed, this distance measure is not symmetric. However, we identify the value of $d(f, g)$ with the *structural similarity* between both functions: lower values for $d(f, g)$ mean that g is *structurally close* to f.

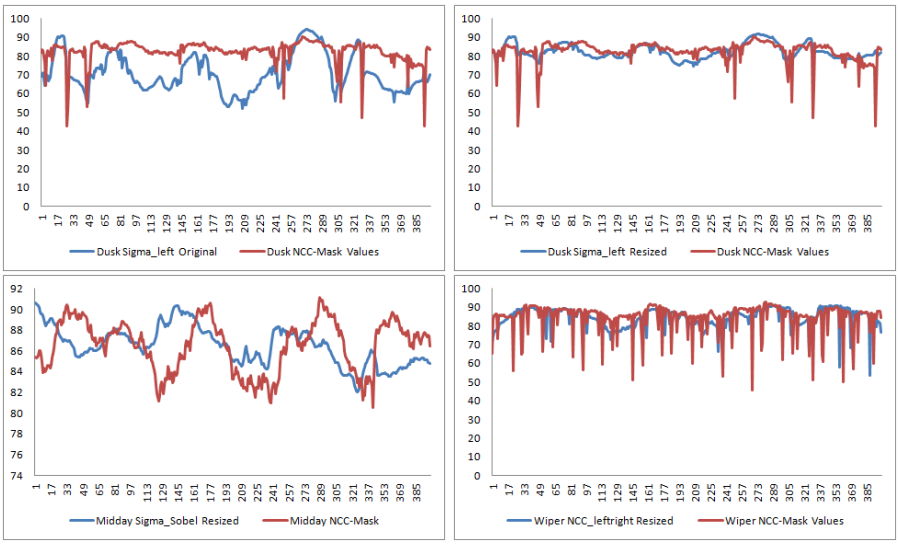

Fig. 2. Illustration of four pairings of functions.

Figure 2 shows pairs of real valued functions on a discrete domain. Function *Sigma_left* represents the standard deviation of intensity values in the reference image, and function *Sigma_Sobel* represents the standard deviation of Sobel edge values for the reference image. Function *NCC_leftright* represents the NCC-measure when comparing the reference and the match image of the sequence. Function *NCC_Mask* is the defined standard measure for the third-eye approach.

The two functions in the upper left image of Fig. 2 have different means and different variances. The upper right shows both functions after Sigma_left was scaled to have the same mean and the same variance as NCC_Mask. The lower row shows two already scaled pairings of functions. Subjective inspection shows inconsistencies between both functions in the lower left, but 'fairly good structural similarity' for both functions in the lower right (when zooming into the figure). Obviously, those subjective inspections can also be replaced by an analytical analysis by calculating the sum of absolute differences in function values, as specified by Equ. (3) and by our distance measure $d(f,g) = d_1(f, g_{new})$. In general, if a distance value $d(f,g)$ is less than half of the standard deviation σ_f used for scaling then both functions are considered to be structurally similar.

3 Discussion of Third-Eye Results

Concluding from visual inspections of calculated disparity maps for all the eight test sequences (and many sequences for earlier projects) we see a very close correspondence between calculated values of the NCC-Mask measure and the actual performance of studied stereo matchers. So far the assumption was that a value of the NCC-Mask measure below the 70% mark is an indication for a

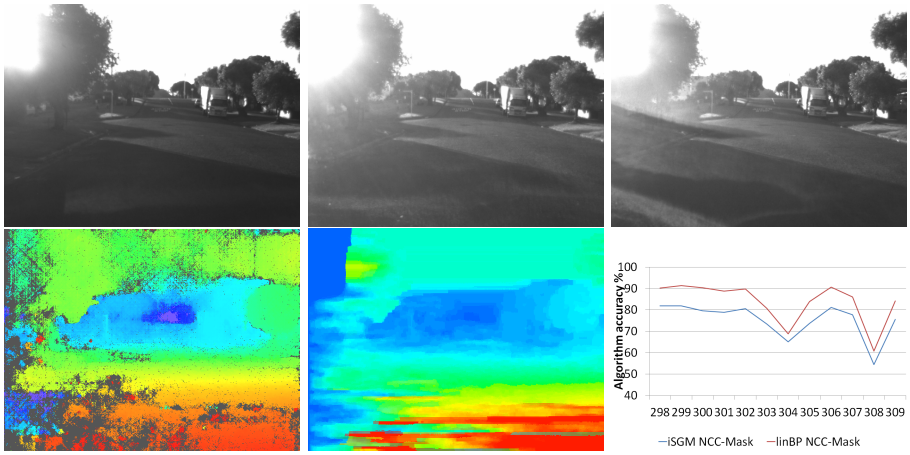

Fig. 3. Dusk sequence. Upper row, left to right: Control image, reference image, and match image. Lower row, left to right: Disparity maps for iSGM, linBP, and NCC-Mask plot for eleven frames from 298 to 309.

failure. However, in the current study we refined this threshold: we recommend to define it in dependency of data measure values for the given stereo frame, for example on the standard deviation of the gradient in the reference image.

As a first case we show a situation where a failure is properly detected. We present Frame 304 from the 'Dusk' sequence; see Fig. 3. Temporarily around this frame, sun strike creates a difficult lighting situation. There are no considerable changes in occluded pixels between the three camera views. Therefore, the third-eye evaluation performance is not considerably affected by occluded pixels. Figure 3 shows the color-encoded disparity maps of iSGM and linBP for this frame. Visually we observe that the performance of both matching algorithms is not good. Both NCC-Mask measures have local minima of about 70% at Frame 304. The third-eye approach works fine in this case.

In the same sense, the NCC-Mask measure also indicates good or bad performances as illustrated in Fig. 4. The upper row illustrates depth maps resulting from iSGM. The left map is for Frame 176 in the sequence 'Queen' showing an excellent result; the right map is for Frame 382 in the sequence 'Dusk' showing a failure. The lower row shows depth maps for linBP for those two frames. However, here, linBP performs not well for Frame 176 in 'Queen', but better than iSGM for Frame 382 in 'Dusk'. The third-eye approach also works fine in general for indicating the 'current winner' (of all participating stereo matchers) for a given situation. There is no all-time winner so far for the tested stereo matchers.

The third-eye approach provides a summarizing single value for each frame, and these summarizing values may not correspond to subjective visual evaluations. From the appearance of the depth maps, iSGM performs better in detecting depth discontinuity edges, thin vertical structures, or other rapid changes in

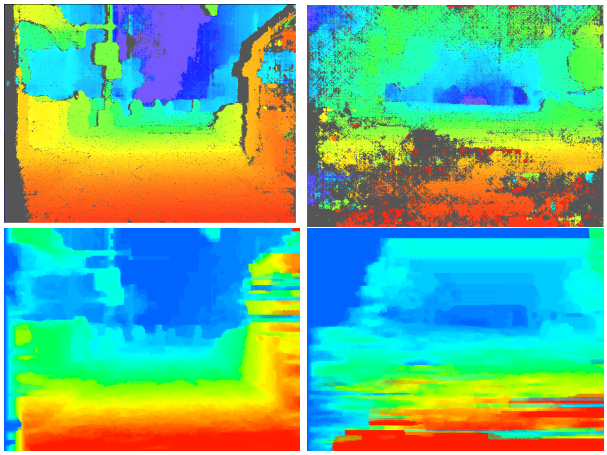

Fig. 4. Depth maps provided by iSGM (upper row) and linBP (lower row), for Frame 176 of 'Queen' (left column) and for Frame 382 of 'Dusk' (right column).

depth. However, linBP is often performing better on large homogeneous areas. See Fig. 5 for plots of NCC-mask measures for sequences 'Barriers', 'People', 'Queen', and 'Night'. In nearly all of the shown 1,434 frames, the value of linBP is above that of iSGM. (For plots for the remaining four sequences of Set 9 of

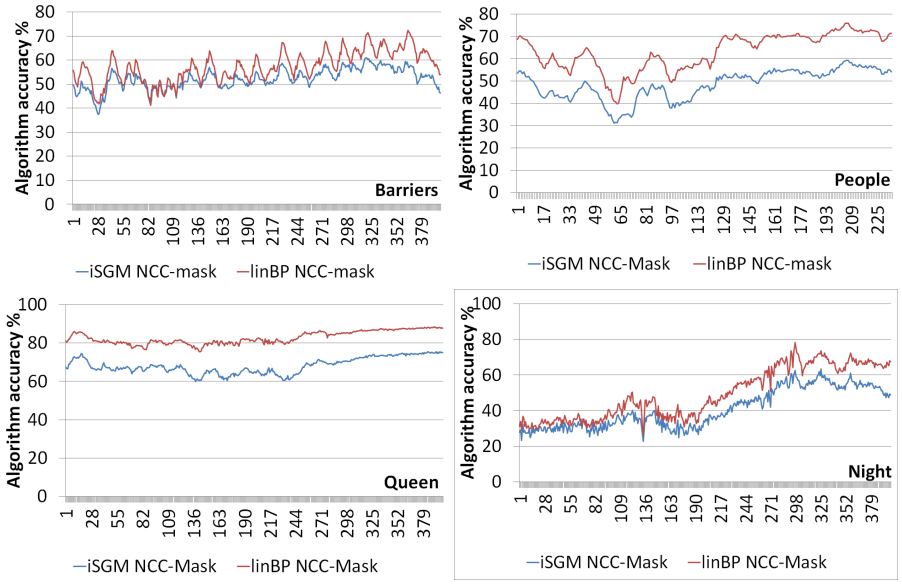

Fig. 5. Plots of the NCC-Mask measure for iSGM and linBP for four of the eight sequences of Set 9 of EISATS.

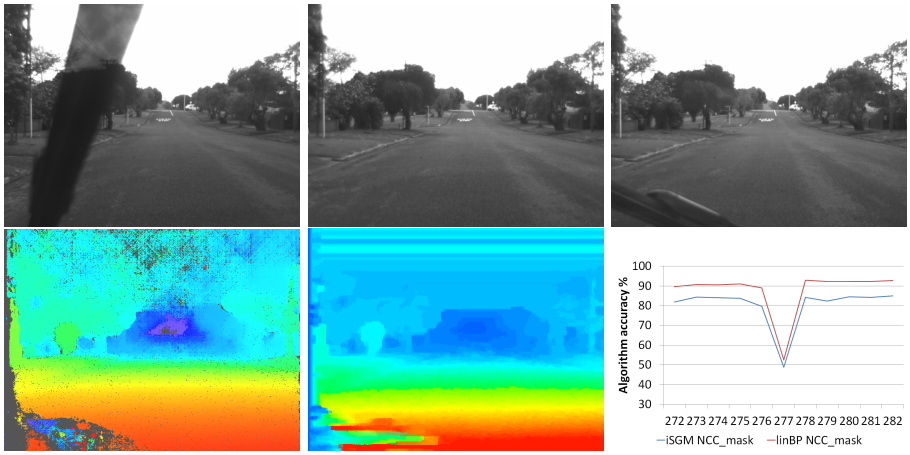

Fig. 6. Wiper sequence. Upper row, left to right: Control image, reference image, and match image. Lower row, left to right: Disparity maps for iSGM, linBP, and NCC-mask plot for eleven frames from 272 to 282.

EISATS, see [10].) The standard deviations of NCC-Mask for linBP, and the distance to the scaled NCC-Mask for iSGM are (5.22, 2.58) for 'Barriers' (2.58 is about 49% of 5.22), (0.87, 0.77) for 'Bridge' (88%), (3.77, 1.91) for 'Dusk' (50%), (1.61, 1.16) for 'Midday' (72%), (13.03, 3.15) for 'Night' (24%), (7.14, 4.27) for 'People' (59%), (2.96, 2.24) for 'Queen' (75%), and (4.86, 1.55) for 'Wiper' (32%). According to our 50% rule defined at the beginning of the next section, we consider both NCC-Mask curves as being structurally similar for 'Barriers', 'Dusk', 'Night', and 'Wiper', on the other four sequences both stereo matchers behave 'qualitatively different'. This analysis is our first important contribution to the application of the third-eye approach.

A second important comment about the third-eye approach: summarizing numbers do have limitations when interpreting. The more accurate detection of 3D details by iSGM compared to linBP is not (!) expressed in the obtained number, but the NCC-mask curves express in general accurately the ups and downs in a matcher's performance.

Finally, as a third important comment, there are cases where the measure provided by the third-eye approach does not coincide with what we see in disparity maps of a stereo matcher due to differences between control and reference image. In order to illustrate this phenomena, we show as an example a trinocular frame of the 'Wiper' sequence. In Fig. 6 we have the disparity maps given by iSGM and linBP for Frame 277 of this sequence and the local plot of the NCC_Mask measures. From the shown disparity maps we would expect a high NCC_Mask value, but there are local minima in the functions. If we observe the complete information of this particular frame given by all the three cameras (see Fig. 6, upper row) we notice that the control image differs significantly from reference and match images, not by different lighting (as for the example in Fig. 3)

but due the wiper position in the control image. Therefore, NCC_Mask gives us a low value. In the 'Wiper' sequence we observe the same effect for Frames 31, 185, and 216, where a low NCC_Mask is incorrectly indicating a failure of the stereo matcher.

Regarding the third comment, a simple solution is to use the NCC measure for quantifying similarity between the control and the reference image; if the similarity value goes below a defined threshold then the third-eye NCC_Mask value is insignificant. Typically this is only true for a very short time (as for the moving wiper).

4 Analysis using Data Measures

We use the NCC_Mask measure for iSGM as the reference function and compare data measures with this function using our distance definition. Table 1 shows those distances, also listing the standard deviation $\text{Sigma}_{NCCMask}$ of the NCC_Mask measure for comparison. *Our 50% rule*: If a distance is below 50% of this standard deviation then we call the functions *structurally similar*.

Table 1. Distance values between NCC_Mask (for iSGM) and data measures.

	Barriers	Bridge	Dusk	Midday	Night	People	Queen	Wiper
Sigma_left	2.28	1.91	3.69	1.52	9.27	6.90	3.78	5.16
Sigma_Sobel	2.25	1.81	4.24	2.50	9.96	9.03	3.49	6.53
NCC_leftright	2.75	2.62	4.24	**1.26**	10.69	5.21	3.32	4.59
$\text{Sigma}_{NCCMask}$	2.85	2.24	6.09	2.47	11.72	7.44	5.35	7.79

The closest to 50% is NCC_leftright for 'Midday'. This shows that structural similarity is low between NCC_Mask and the used three data measures in general. It appears that the complexity of video data for stereo matching cannot simply be estimated by just using a summarizing distance value for one of those three global data measures for a whole sequence, showing (in our case) 400 frames of one particular situation.

A more refined approach is to study the graphs of the functions of the data measures, as we already did for NCC_Mask in the third-eye approach. Scaled functions are shown in Fig. 7, together with NCC_Mask (for iSGM) which was kept constant for scaling. Sequences 'Midday' and 'Queen' have a very low variance, and 'Dusk' and 'Wiper' represent special challenges for stereo matching apparent by rapid drops in NCC-Mask values. Values of these curves show locally some kind of correspondence with NCC_Mask, and sometimes differences. Correspondences may explain a fail of the stereo matcher, and differences may provide hints how to improve the stereo matcher at this particular place of the stereo sequence.

A more refined approach is to use the local variance of data measures (e.g. for six frames backward, the current frame, and six frames forward). We demonstrate this for a case where iSGM failed, and discuss the data measures only in the local context of 13 frames. See Fig. 8.

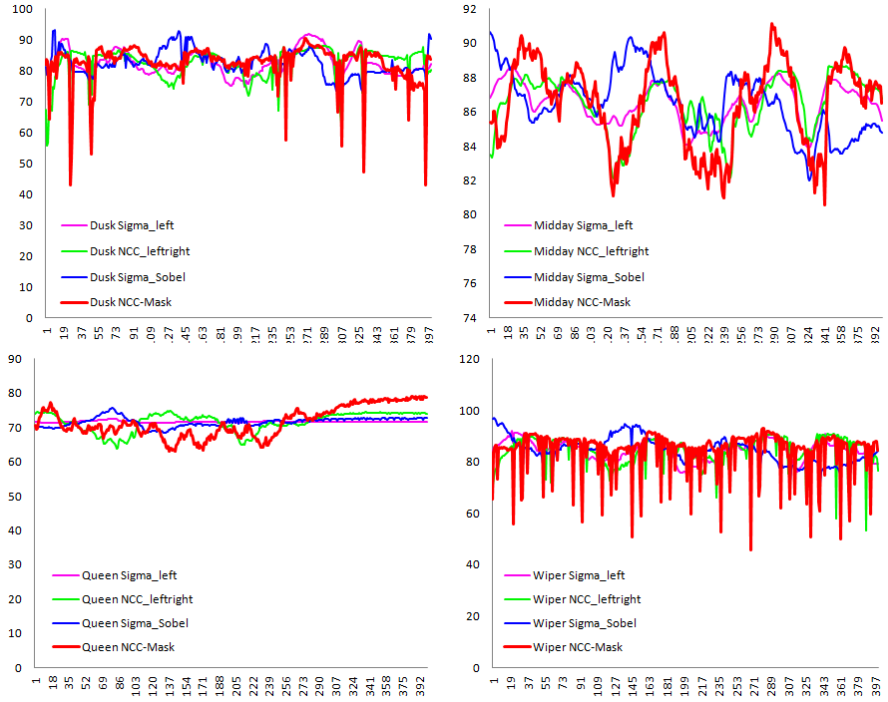

Fig. 7. Scaled curves for direct visual comparison.

We analyze the situation around Frame 330 of the 'Dusk'. Not only the NCC_Mask values (see figure on the left), but also the appearance of iSGM disparity map for this frame indicates a fail in stereo matching. The figure shows on the right the local variances of the data measures and of NCC_Mask. They all go up around Frame 330, but are nearly constant before and after.

This illustrates a general observation: at places where a fail in stereo matching occurred, typically also one, two or all three of the data measures showed a large local variance.

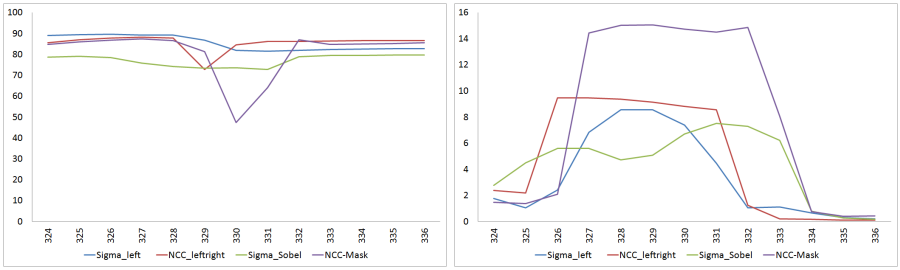

Fig. 8. Functions on the 'Dusk' sequence between Frames 324 to 336. Left: comparison between scaled data measures and iSGM NCC_Mask. Right: comparison of the local variance of data measures and iSGM's NCC_Mask.

5 Conclusions

The third-eye approach is a valuable tool for analyzing stereo matchers on long sequences or for continuous recording. Detections of a 'fail' are important for implemented systems, and this paper provided a critical discussion how to detect such 'fails', and how to use data measures for a more detailed analysis, especially at places where a 'fail' was detected, and where a data analysis might lead to some ideas how to improve the given stereo matcher at this place, for the shown particular situation.

Acknowledgement: The authors thank Simon Hermann and Waqar Khan for providing the executables for iSGM and linBP, respectively, and Sandino Morales for providing the sources for the third-eye approach and comments on the paper.

References

1. Baeza-Yates, R., Valiente, G.: An image similarity measure based on graph matching. Proc. String Processing Information Retrieval (2000)
2. Canny, J.: A computational approach to edge detection. IEEE Trans. PAMI **8** (1986) 679–698
3. Drayer, B.: Image comparison on the base of a combinatorial matching algorithm. LNCS 6835 (2011) 426-431
4. EISATS benchmark data base. The University of Auckland, www.mi.auckland.ac.nz/EISATS (2013)
5. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient belief propagation for early vision. Int. J. Computer Vision **70** (2006) 41–54
6. Hermann, S., Morales, S., Klette, R.: Half-resolution semi-global stereo matching. Proc. IEEE Intelligent Vehicles Symposium (2011) 201–206
7. Hermann, S., Klette, R.: Iterative semi-global matching for robust driver assistance systems. Proc. Asian Conf. Computer Vision, LNCS (2013)
8. Kalinic, H., Loncaric, S., Bijnens, B.: A novel image similarity measure for image registration. Proc. Image Signal Processing Analysis (2011) 195–199
9. Klette, R., Krüger, N., Vaudrey, T., Pauwels, K., Hulle, M., Morales, S., Kandil, F., Haeusler, R., Pugeault, N., Rabe, C., Lappe, M.: Performance of correspondence algorithms in vision-based driver assistance using an online image sequence database. IEEE Trans. Vehicular Technology **60** (2011) 2012–2026
10. Khan, W., Suaste, V., Caudillo, D., Klette, R.: Belief propagation stereo matching compared to iSGM on binocular or trinocular video data. MItech-TR-82 (2013) The University of Auckland
11. Morales, S., Klette, R.: A third eye for performance evaluation in stereo sequence analysis. Proc. Computer Analysis Images Patterns (2009) 1078–1086
12. Ranftl, R., Gehrig, S., Pock, T., Bischof, H.: Pushing the limits of stereo using variational stereo estimation. Proc. IEEE Intelligent Vehicles Symposium (2012) 401–407
13. Rosin, P., Ellis, T.: Image difference threshold strategies and shadow detection. Proc. British Machine Vision Conf. (1995) 347–356
14. Thiyagarajan, M., Samundeeswari, S.: A new similarity measure for image segmentation. Int. J. Computer Science Engineering **02** (2010) 831–835