# Prediction Error Evaluation of Various Stereo Matching Algorithms on Long Stereo Sequences

Sandino Morales and Reinhard Klette

The *.enpeda..* Project, The University of Auckland
Auckland, New Zealand

**Abstract.** Prediction errors are commonly used when analyzing the performance of a multi-camera stereo system using at least three cameras. This paper discusses this methodology for performance evaluation on long stereo sequences (in the context of vision-based driver assistance systems). Three cameras are calibrated in an ego-vehicle, and prediction error analysis is performed on recorded stereo sequences. They are evaluated using various common stereo matching algorithms, such as belief propagation, dynamic programming, semi-global matching, or graph cut. This performance evaluation is demonstrated on synthetic and real data.

## 1 Introduction

Assume a rectified stereo pair of images and a disparity map obtained by applying some stereo matching algorithm. One way to evaluate the performance of this matching algorithm, in absence of ground truth data, is to calculate – from both, the stereo input images and the calculated disparity map – a new (third) image, as it would appear for a *virtual camera*, assumed to be at a defined pose, and compare this with an image actually recorded at that pose. At pixels of the virtual camera we assign either visible surface textures, value 'black' for pixels occluded in the left image, and 'white' for pixels occluded in the right image. The comparison with the recorded image has to take those uncertainties into account.

This performance analysis is known as *prediction error evaluation* [15], and it is applied when at least three images of the same scene are available; see, for example, [1]. The third image is used as ground truth data, and statistical analysis is performed to analyze the matching algorithms.

We are recording video data with a three-camera system, and the described evaluation is not only done for one triple of images but for three (long) image sequences, and for one long synthetic stereo image sequence where a third camera may be simulated based on available ground truth (see Figure 1).

The outline of this paper is as follows. We first briefly recall the geometric approach that is commonly used to generate the novel view from a previously calculated disparity map and a pair of rectified images. Then, we specify the used stereo algorithms and the quality metrics used to perform the analysis.

We finalize the discussion by presenting experimental results of our research on stereo sequences, and derive some conclusions.

(a) Left synthetic image.

(b) Right synthetic image.

(c) Left real world image.

(d) Right real world image.

(e) Middle real world image.

**Fig. 1.** Examples for used sequences. (a) and (b): Left and right frames no. 22 of the synthetic sequence. (c), (d) and (e): Left, right and middle frames no. 95 of the real-world sequence.

## 2 Geometry of the Third View

Assume that three cameras are given, two of them are set up (rectified) in such a way that their image planes satisfy the *standard stereo geometry* (e.g., see [9], and [6] for rectification). We denote the left camera of the stereo system and its respective image as the *reference camera* and *reference image*; the right camera is denoted as the *matching camera*, recording the *matching image*. The third camera is potentially at an arbitrary pose and provides the *third image*. The calculated image is the *novel image*.

Assume that the coordinate system of the reference camera is also the world coordinate system. In order to keep a simple notation, the coordinates of a point in the different image planes will be written in terms of the coordinate system defined by the respective camera; transforms into the image plane are obvious.

Following central projection geometry [6], the calculation of a novel view is straightforward. Suppose that the reference, matching and novel cameras are in a geometric position as sketched in Figure 2. The camera center of the reference camera lies at the origin $O = (0, 0, 0)$, the camera center of the matching camera is at point $O_M = (b, 0, 0)$; and the focal point of the novel camera is at point $O_N = (b_1, b_2, b_3)$.

Suppose that the disparity values of the reference and the matching images have been calculated by some stereo matching algorithm. Let $P = (X, Y, Z)$ be a scene point visible for all the three cameras and $p = (x, y)$, $p_M = (x_M, y_M)$, and $p_N = (x_N, y_N)$ its projections on the reference, matching, and novel image planes, respectively.

For the assumed case of standard stereo geometry between reference and matching image, we provide a formula below to obtain the coordinates of $p_N$ in terms of the coordinates of $p$, the base-line distance $b$, the focal length $f$ of the reference and matching cameras (due to the rectification of left and right

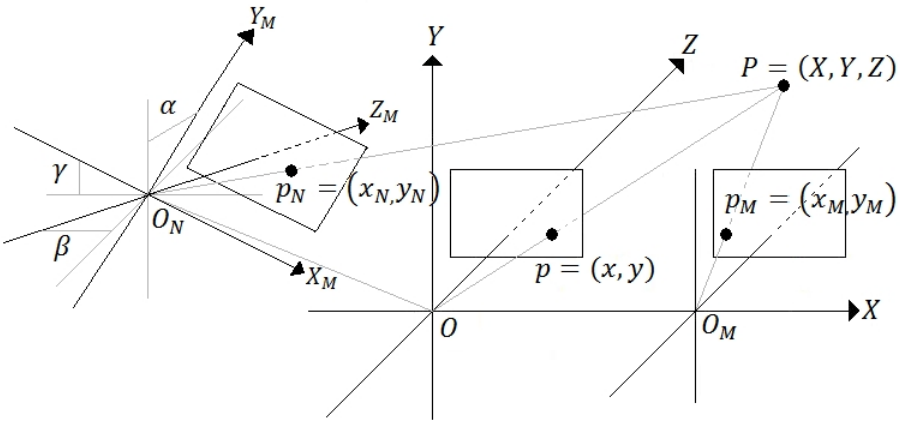



**Fig. 2.** Notation as used for describing a three-camera configuration.

cameras), and the already calculated disparity $d$ between $p$ and $p_M$. Since $P$ is visible from the reference and matching cameras, by triangulation, it is possible to recover the coordinates of $P$ (in terms of the coordinate system of the reference camera):

$$X = \frac{x \cdot b}{d}, \quad Y = \frac{y \cdot b}{b} \quad \text{and} \quad Z = \frac{f \cdot b}{d}$$

Now, let $(X_N, Y_N, Z_N)$ be the coordinates of $P$ with respect to $O_N$. Using homogenous coordinates and letting $\mathbf{C}$ and $\mathbf{S}$ be used for denoting the *cosine* and *sine* functions, respectively, the matrix

$$M = \begin{pmatrix} \mathbf{C}\gamma\mathbf{C}\beta & -\mathbf{C}\gamma\mathbf{S}\beta\mathbf{S}\alpha - \mathbf{S}\gamma\mathbf{C}\alpha & \mathbf{S}\gamma\mathbf{S}\alpha - \mathbf{C}\gamma\mathbf{S}\beta\mathbf{C}\alpha & -u_1 \\ \mathbf{S}\gamma\mathbf{C}\beta & \mathbf{C}\gamma\mathbf{C}\alpha - \mathbf{S}\gamma\mathbf{S}\beta\mathbf{S}\alpha & -\mathbf{S}\gamma\mathbf{S}\beta\mathbf{C}\alpha - \mathbf{C}\gamma\mathbf{S}\alpha & -u_2 \\ \mathbf{S}\beta & \mathbf{C}\beta\mathbf{S}\alpha & \mathbf{C}\beta\mathbf{C}\alpha & -u_3 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

where (Note: angles as in Figure 2.)

$$\begin{aligned} u_1 &= b_1\mathbf{C}\gamma\mathbf{C}\beta + b_2(-\mathbf{C}\gamma\mathbf{S}\beta\mathbf{S}\alpha - \mathbf{S}\gamma\mathbf{C}\alpha) + b_3(\mathbf{S}\gamma\mathbf{S}\alpha - \mathbf{C}\gamma\mathbf{S}\beta\mathbf{C}\alpha) \\ u_2 &= b_1\mathbf{S}\gamma\mathbf{C}\beta + b_2(\mathbf{C}\gamma\mathbf{C}\alpha - \mathbf{S}\gamma\mathbf{S}\beta\mathbf{S}\alpha) + b_3(-\mathbf{S}\gamma\mathbf{S}\beta\mathbf{C}\alpha - \mathbf{C}\gamma\mathbf{S}\alpha) \\ u_3 &= b_1\mathbf{S}\beta + b_2\mathbf{C}\beta\mathbf{S}\alpha + b_3\mathbf{C}\beta\mathbf{C}\alpha \end{aligned}$$

specifies the following mapping:

$$(X_N, Y_N, Z_N, 1) = \boldsymbol{M} \cdot (X, Y, Z, 1)^T$$

Let $m_{ij}$ be the element at position $i, j$ in $\boldsymbol{M}$, for $1 \leq i, j \leq 3$, and $f_N$ the focal length of the novel camera. Thus, using the equations defined by central projection [6], we have that

$$x_N = f_N \cdot \frac{m_{11}(bx - db_1) + m_{12}(by - db_2) + m_{13}(bf - db_3)}{m_{31}(bx - db_1) + m_{32}(by - db_2) + m_{33}(bf - db_3)} \tag{1}$$

$$y_N = f_N \cdot \frac{m_{21}(bx - db_1) + m_{22}(by - db_2) + m_{23}(bf - db_3)}{m_{31}(bx - db_1) + m_{32}(by - db_2) + m_{33}(bf - db_3)} \tag{2}$$

where $d$ and $b$ were defined above as the disparity between points $p$ and $p_M$, and as the base line distance between the reference and matching cameras, respectively. With these two *forward equations* (e.g., see [8]) it is possible to map any pixel location $(x, y)$ in the reference image into a point $(x_N, y_N)$ in the image plane of the novel image, which is then possibly visible in the novel image. We denote this transform by $\boldsymbol{N}$.

## 3 Poses of the Third Camera

In this section we discuss possible poses of the third camera. Occluded points may cause a bias when evaluating the performance of an algorithm. We illustrate this by examples generated for the synthetic stereo sequence, in Set 2 on [3]. The

occluded points vary depending on the pose of the third camera, defining the novel view.

By increasing the difference between the pose of the third view from the pose of the reference image, more occluded areas occur in the novel view. Note that points can not be reconstructed due to occlusions between the reference and matching images, even that such points may be visible in the reference and in the third view. In such cases it is impossible to assign a texture value to those pixels in the novel image as there is no depth information available that allows us to apply Equations (1) and (2). Such areas can be reduced if the third view is to the right of the reference image, since some other (non occluded) points will be mapped into those occluded positions. However, when trying to exclude such occlusions, other points (i.e., the ones that are not visible in the reference image) may become occluded.

Figure 3 shows three different occlusion cases. We calculate the reference and matching images from the depth ground truth of a stereo pair of Set 2 from [3]. Figure 3(b) is the reconstruction of the reference image; occluded points (between reference and matching image) are shown as white pixels; they can not be reconstructed with this approach since there is no disparity information available, even if, as in this case, they are visible in the reference image. Figure 3(c) is the reconstruction of the matching image; occluded pixels are represented as black pixels; they are visible in the matching image but not in the reference image. Figure 3(a) is an example of a novel view in which both kinds of occlusions occur (white and black).

**The Symmetric Pose.** The symmetric pose of the third camera (half-way between reference and matching camera) is expected to be the one which minimizes impacts of occlusions (i.e., the total number of either black or white pixels). In evaluations it would be ideal to separate the impact of occlusions from those of incorrect matching. Thus, the symmetric case seems to be a good choice. However, errors of incorrect mismatches do not have such an obvious impact compared to cases where the third pose differs (much) from the symmetric case.

**Collinear Poses.** In this case, the focal point of the third camera is on the base line of the first two cameras. If the third view is on the left of the reference camera, both kinds of occlusions (black and white) are present in the novel view. In this research we decided for the collinear case, having the third camera approximately 40 cm to the left of the reference camera. (Rectified reference and matching camera are about 30 cm apart.) Thus we have to deal with both kinds of occlusions in the third image sequence.

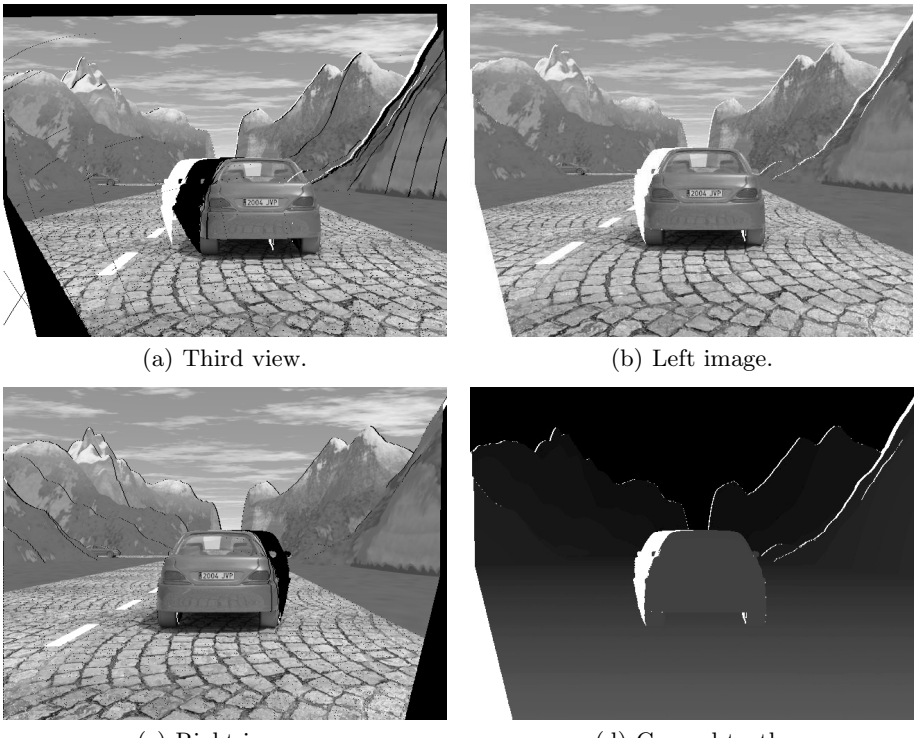
## 4 Evaluations using the Third Sequence

This section further explains the set up of the experiments, we present the used data set, and the quality metrics. We also introduce briefly the tested stereo algorithms.

**Outline of Experiments**. For each rectified pair of frames of a given sequence and its calculated depth map, we generate an image as it would be seen by a virtual camera in exactly the same collinear pose of our third camera (left of the reference camera). This allows to compare the intensity values of the novel image with those of the available third view.

**Data Set**. We used two sequences for this paper (see Figure 1). The gray-value synthetic sequence from Set 2 of [3] consists of 100 stereo pairs with available ground truth [16]. We generate the third view sequence (with occlusions) as being about 40 cm to the left of the reference camera. The usage of this sequence allows us to integrate results from a previous study [12] obtained for the same data set.

We compare the evaluation results for this synthetic example with those obtained for a trinocular real world sequence of 150×3 images taken with three



(a) Third view. (b) Left image.

(c) Right image. (d) Ground truth.

**Fig. 3.** (a) Calculated (from ground truth) third image of a virtual camera positioned on the left of the reference camera. (b) and (c) are calculations of the left (reference) and right (matching) images from the stereo sequence in Set 2 of [3] with the approach as presented here. (d) Ground truth for the left image, with disparity code: light = close, dark = far, white = occlusion.

calibrated cameras mounted in the research vehicle of the *.enpeda..* project, see Figure 1. We selected the middle and right camera to be the reference and matching camera, respectively. The camera center of the reference camera is considered to be the origin of the world coordinate system, and the other two cameras are calibrated with respect to this coordinate system.

**Stereo Algorithms**. We aimed at testing a representative collection of various stereo algorithms, and they are as follows:
*Dynamic programming stereo.* We compare a standard algorithm [13] (DP), against one with temporal (DPt), spatial (DPs), or temporal and spatial (DPts) propagation; see [11].
*Belief propagation stereo.* We use a coarse-to-fine algorithm BP [4] with quadratic cost function, with parameter settings as reported in [5].
*Semi-global matching.* An SGM strategy [7] allows us to use different cost functions; we use mutual information (SGM MI) and Birchfield-Tomasi (SGM BT).
*Graph Cut.* For a detailed discussion of the GC method, see [2] and [10].

**Quality Metrics**. We use the following two quality metrics.
*Root mean squared error.* Let $(x, y)$ be a pixel in the reference image $I_R$ with intensity $I_R(x, y)$ and $(x_N, y_N) = \boldsymbol{N}(x, y)$ a pixel in the novel image $I_N$ with intensity $I_N(x_N, y_N) = I_R(x, y)$. Thus, we compute, for frame $t$ of the respective sequence, the RMS between the third image $I_T$ and the novel image as follows:

$$R(t) = \frac{1}{|\Omega_t|}\sqrt{\sum_{\Omega_t}\left(I_T(x_N, y_N) - I_N(x_N, y_N)\right)^2}$$

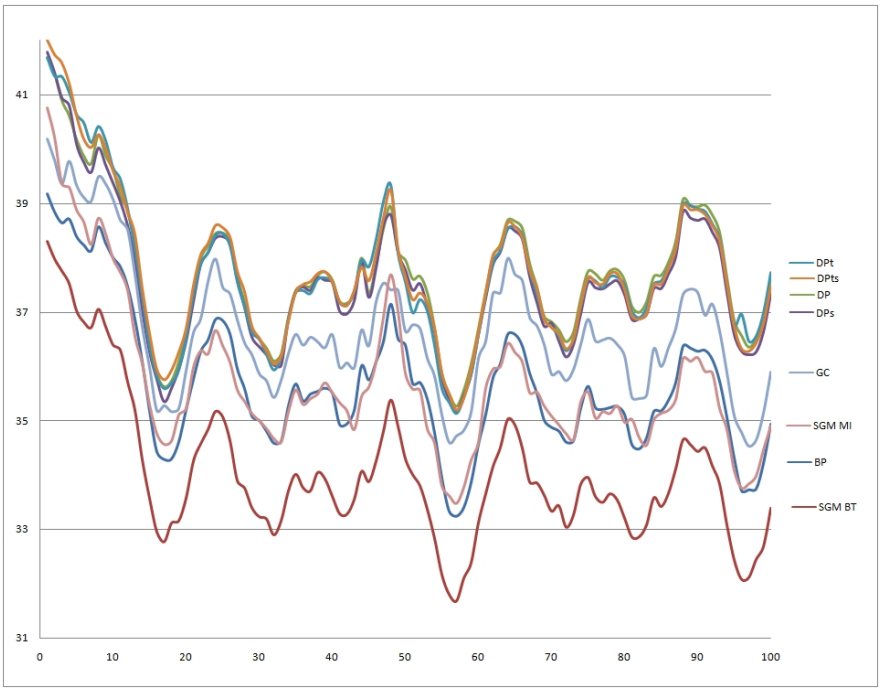
where $|\Omega_t|$ denotes the cardinality of the discrete domain $\Omega_t$ of non occluded pixels.
*Normalized cross correlation.* The normalized cross correlation (CC) is used to compare the third camera image against the novel image. Using the same notation as above, the normalized cross correlation at time $t$ is defined as

$$CC(t) = \frac{1}{|\Omega_t|}\sum_{\Omega_t}\frac{(I_T(x_N, y_N) - \mu_T)(I_N(x_N, y_N) - \mu_N)}{\sigma_T \sigma_N}$$

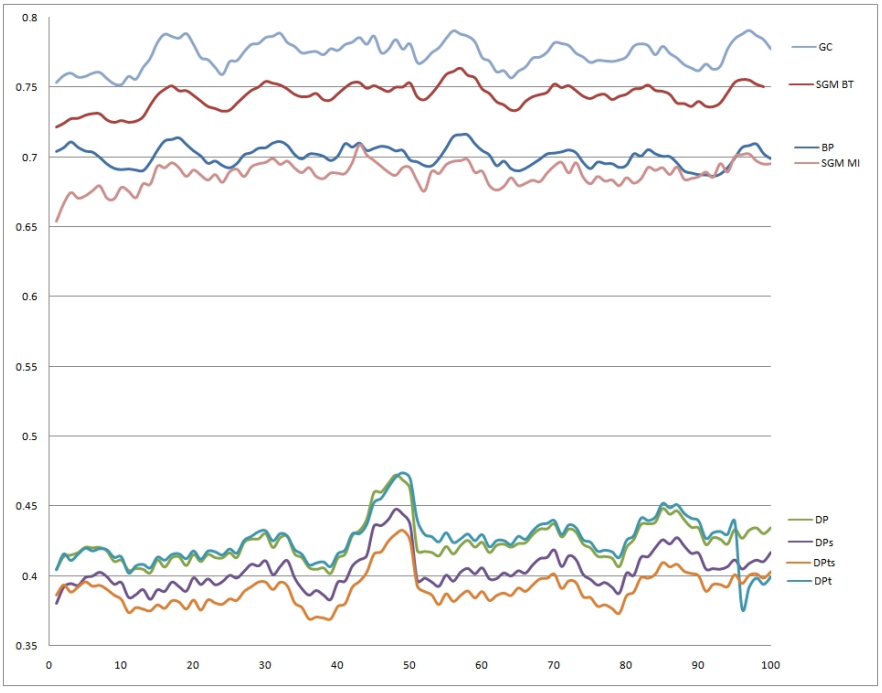
$\mu_N$ and $\mu_T$ denote the means, and $\sigma_N$ and $\sigma_T$ the standard deviations of $I_N$ and $I_T$.

**Results for the synthetic sequence.** For the RMS results (see Figure 4(a) and Table 1), the algorithm with the best overall performance was SGM BT, followed by BP and SGM MI; GC ranks fourth followed by the dynamic programming algorithms. Note that a ranking for just a single stereo pair may look different; this is a summary.

The large error occurring in the first 10 frames the error is a result of big occluded areas caused by a close object - a car. The local maximum around frame 45 is caused by a similar situation. However, we can conclude that the summarized ranking of the algorithms is not affected by those situations. The




(a) RMS




(b) CC

**Fig. 4.** Frame-by-frame results for the synthetic sequence. For (b) note that closer to 1.0 means "better" (i.e., the correlation between the two images is higher).

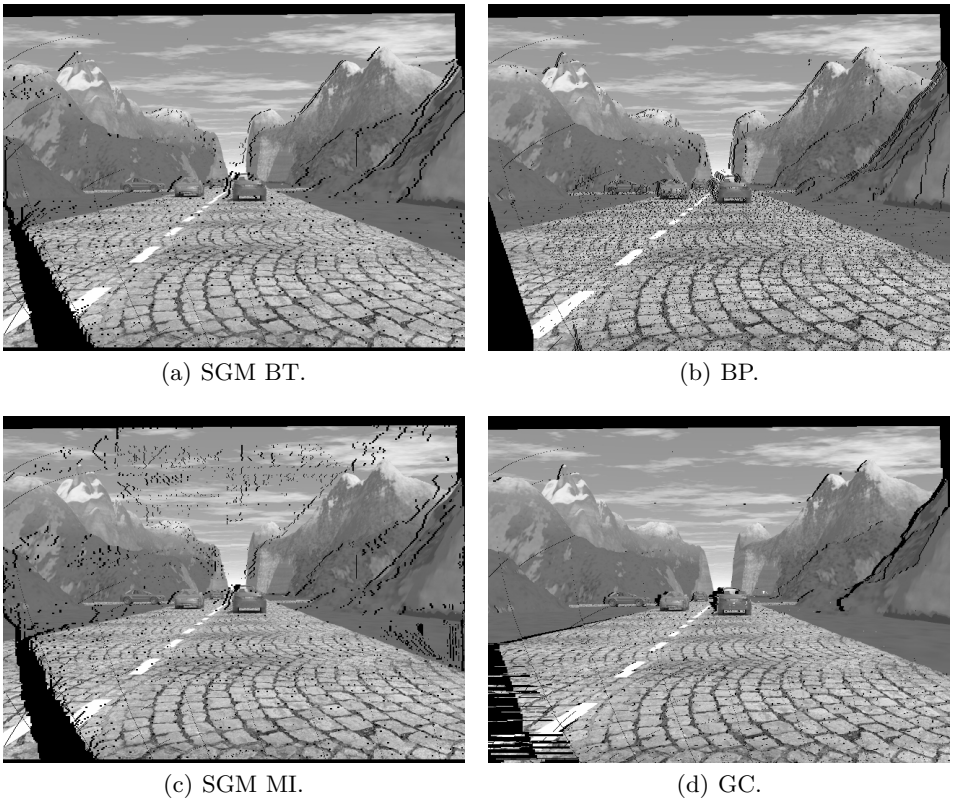
(a) RMS results.

| Algorithm | Mean | Min | Max |
|---|---|---|---|
| SGM BT | 34.05 | 13.67 | 30.68 |
| BP | 35.69 | 14.72 | 31.59 |
| SGM MI | 35.72 | 14.24 | 29.85 |
| GC | 36.67 | 17.30 | 34.19 |
| DPs | 37.55 | 13.75 | 32.43 |
| DPt | 37.68 | 12.99 | 32.43 |
| DP | 37.70 | 12.95 | 32.53 |
| DPts | 37.70 | 13.03 | 28.98 |

(b) CC results.

| Algorithm | Mean | Min | Max |
|---|---|---|---|
| GC | 0.77 | 0.75 | 0.79 |
| SGM BT | 0.74 | 0.72 | 0.76 |
| BP | 0.70 | 0.69 | 0.72 |
| SGM MI | 0.69 | 0.65 | 0.71 |
| DPt | 0.43 | 0.38 | 0.45 |
| DP | 0.42 | 0.40 | 0.47 |
| DPs | 0.40 | 0.38 | 0.45 |
| DPts | 0.39 | 0.37 | 0.43 |

**Table 1.** Summarizing results for the synthetic sequence.



(a) SGM BT. (b) BP.

(c) SGM MI. (d) GC.

**Fig. 5.** Examples of novel views for the best four performing algorithms. The black strip on the top is due to a minor tilt in the pose of the third view.

other local maxima, around frames 15, 20 and 60, are due to errors in the calculated disparity maps, as there are no obvious changes in occluded areas in those frames.

The ranking of the algorithms resemble the one obtained in [12], where a different evaluation methodology was used (not a third camera but just a comparison with ground truth). There it was also stated that SGM BT performed best for this sequence. However, BP and SGM MI swapped their positions in those two different evaluations. Another difference is that DPs ranked third among the four dynamic programming algorithms in that previous study, but shows now best performance out of those four.

The CC measure ranking is different from the one derived from RMS. The GC algorithm performs best overall, followed by SGM BT, BP and SGM MI. Figure 5 shows the novel views of frame 22 for the top four performing algorithms. The four dynamic programming algorithms were the worst again; with DPt performing best for most of the frames, and DPts being the worst. For the top four algorithms it is evident that the performance on the first 10 frames, is again, impacted by occlusions; the four dynamic programming algorithms do not show this change in performance.

**Results for the Real World Sequence.** For RMS, all the eight algorithms behave pretty much the same! The difference in magnitude is not evident at all as the function graphs are highly overlapping; see Figure 6. However, Table 2 shows that DPts appears to be the best algorithm, followed closely by DPt. The worst algorithm by far is GC in this case. The local maxima correspond, also for this sequence, to frames where there are closer objects to the ego-vehicle, causing more occluded areas. – The CC results showed a totally different ranking. For this metric, BP performed the best, followed by DP and DPt; DPts was the worst algorithm for this metric, which tells us that it calculates inaccurate values at many pixels, but errors are fairly small. Note that SGM MI performs better than SMG BT, which confirms the ranking in [12], where SMG BT proved to be more sensitive to common real-world noise than SGM-MI.

## 5 Conclusions

This paper evaluates the performance of several stereo algorithms, using the generation of a novel view from the disparity map. We conclude that this prediction error analysis is a valuable tool to test the performance of stereo algorithms; when no real-world ground truth is available.

We notice a good correlation with RMS evaluations as previously obtained for the used synthetic sequence in [12], where the methodology was characterized by using the ground truth. Also, we could confirm again that SGM BT has more difficulties to deal with noisy (outdoor) images than SGM MI. Occlusions seem to have an influence on the magnitudes of the errors, but do not seem to affect the ranking of the algorithms very much. It is also evident that testing algorithms on real world sequences is necessary, as the ranking of one algorithm may vary totally if used on a synthetic or a real world sequence.

(a) RMS results.

| Algorithm | Mean | Min | Max |
|---|---|---|---|
| DPts | 19.74 | 13.03 | 28.98 |
| DPt | 19.75 | 12.99 | 32.43 |
| SGM MI | 20.09 | 14.24 | 29.85 |
| SGM BT | 20.09 | 13.67 | 30.68 |
| DP | 21.19 | 12.95 | 32.53 |
| BP | 21.91 | 14.72 | 31.59 |
| DPs | 22.23 | 13.75 | 32.43 |
| GC | 24.79 | 17.30 | 34.19 |

(b) CC results.

| Algorithm | Mean | Min | Max |
|---|---|---|---|
| BP | 0.85 | 0.79 | 0.89 |
| DP | 0.84 | 0.76 | 0.89 |
| DPt | 0.83 | 0.75 | 0.72 |
| GC | 0.80 | 0.74 | 0.84 |
| SGM MI | 0.73 | 0.63 | 0.81 |
| SGM BT | 0.69 | 0.62 | 0.73 |
| DPs | 0.62 | 0.54 | 0.72 |
| DPts | 0.50 | 0.42 | 0.61 |

**Table 2.** Results for the real world sequence.

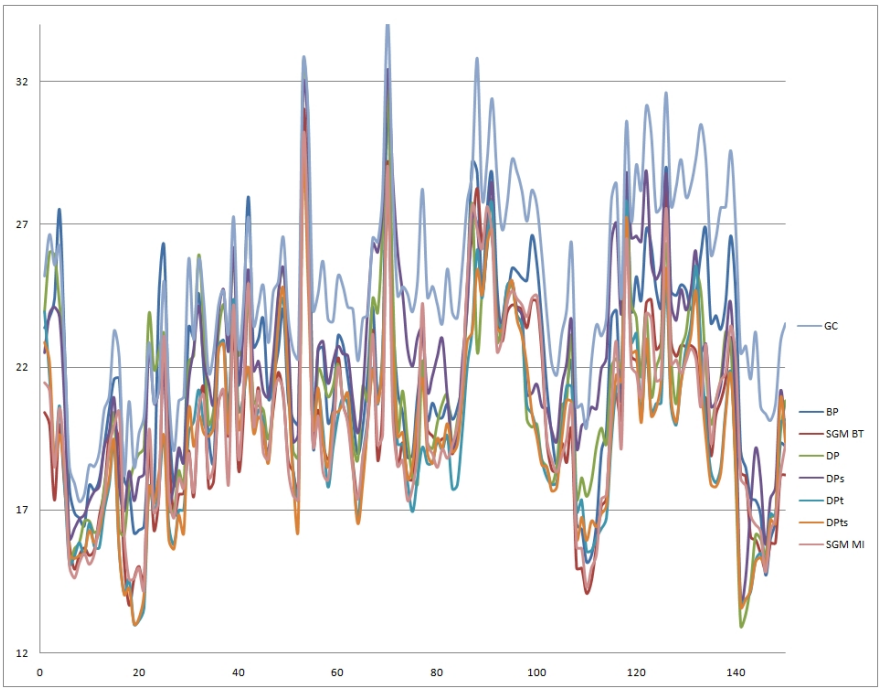
Future work may use different positions for the third camera and different metrics in order to widen the study about relationships between occlusions and accuracy.

**Acknowledgement**. The authors thank Stefan Gehrig for his implementations of SGM, Shushi Guan for his specification of the [4] implementation of BP, Joachim Penc for his GC implementation and Tobi Vaudrey for his valuable comments.

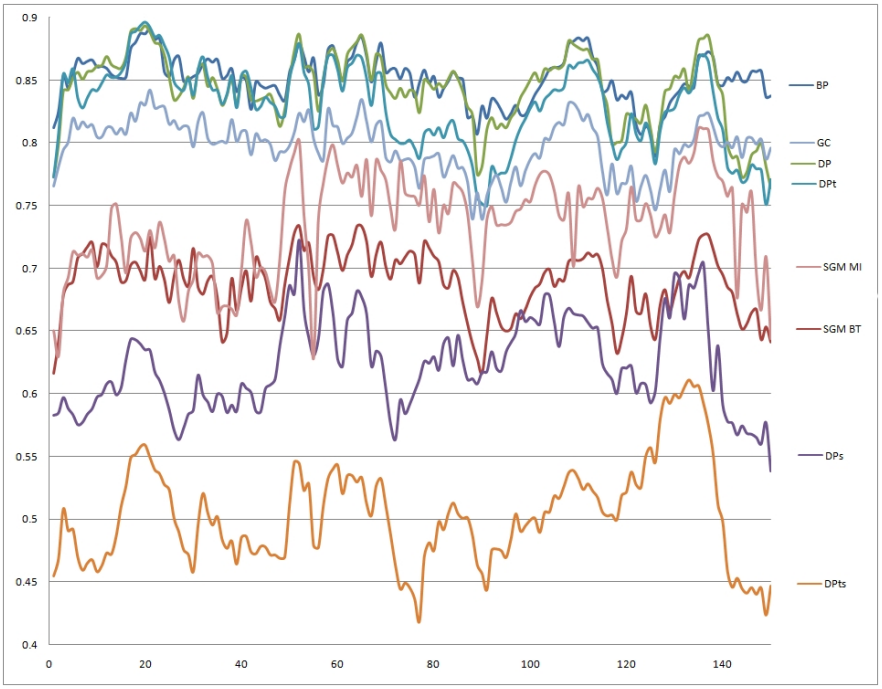
## References

1. 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. Computer Vision*, CD, 2007.
2. Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Analysis Machine Intelligence*, **23**:1222–1239, 2001.
3. *.enpeda..* image sequence analysis test site (EISATS). `http://www.mi.auckland.ac.nz/EISATS/`
4. P.F. Felzenszwalb and D.P. Huttenlocher. Efficient belief propagation for early vision. *Int. J. Computer Vision*, **70**:261–268, 2006.
5. S. Guan, R. Klette, and Y.W. Woo. Belief propagation for stereo analysis of night-vision sequences. In Proc. *PSIVT*, LNCS 5414, pages 932–943, 2009.
6. R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. 2nd edition, Cambridge University Press, Cambridge, 2004.
7. H. Hirschmüller. Accurate and efficient stereo processing by semi-global matching and mutual information. In Proc. *Computer Vision Pattern Recognition*, volume 2, pages 807–814, 2005.
8. R. Klette and P. Zamperoni. *Handbook of Image Processing Operators*. Wiley, Chichester, 1996.

9. R. Klette, K. Schlüns and A. Koschan. *Computer Vision. Three-Dimensional Data from Images*. Springer, Singapore, 1998.
10. V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts? *IEEE Trans. Pattern Analysis Machine Intelligence*, **26**:65–81, 2004.
11. Z. Liu and R. Klette, Dynamic programming stereo on real-world sequences. In Proc. *ICONIP*, LNCS, 2009 (to appear).
12. S. Morales, T. Vaudrey, and R. Klette. An in depth robustness evaluation of stereo algorithms on long stereo sequences. In Proc. *Intelligent Vehicles*, 2009 (to appear).
13. Y. Ohta and T. Kanade. Stereo by two-level dynamic programming. In Proc. *IJCAI*, pages 1120–1126, 1985.
14. D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Computer Vision*, **47**:7–42, 2002.
15. R. Szeliski. Prediction error as a quality metric for motion and stereo. In Proc. *Int. Conf. Computer Vision*, volume 2, pages 781–788, 1999.
16. T. Vaudrey, C. Rabe, R. Klette, and J. Milburn. Differences between stereo and motion behaviour on synthetic and real-world stereo sequences. In Proc. *Image Vision Computing New Zealand*, IEEE online, 2008.

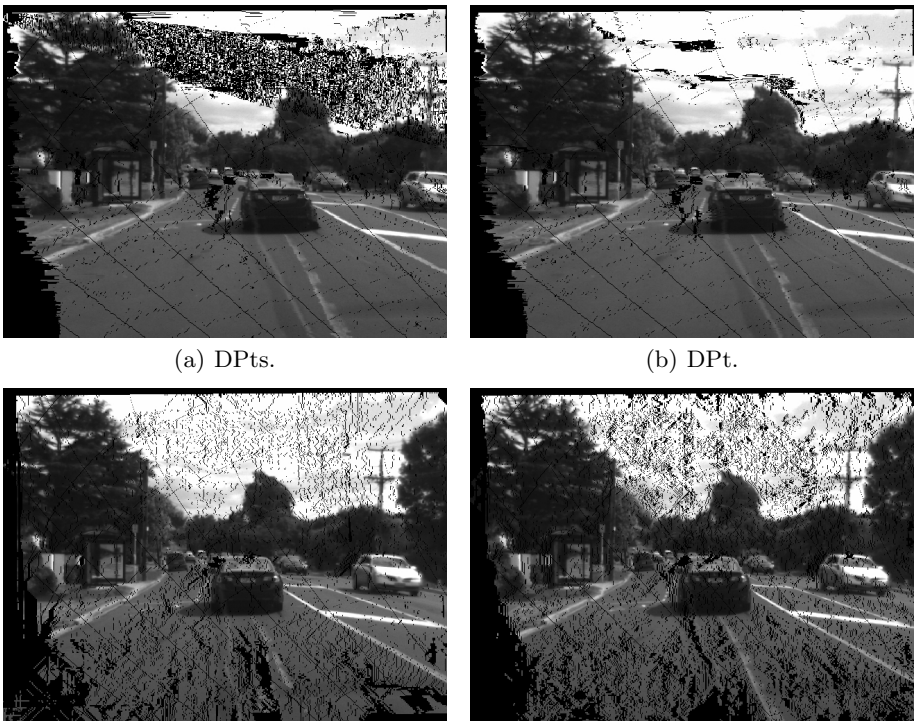



(a) RMS.




(b) CC.

**Fig. 6.** Frame-by-frame results for the used real world sequence.



(a) DPts. (b) DPt.

(c) SGM MI. (d) SGM BT.

**Fig. 7.** Examples of novel views for the top four performing algorithms on the real-world sequence. The black strip on the top is due to a minor tilt in the pose of the third view