Real-World Stereo-Analysis Evaluation

Sandino Morales, Simon Hermann, and Reinhard Klette

.enpeda.. Project, The University of Auckland, New Zealand
pmor085@aucklanduni.ac.nz
www.mi.auckland.ac.nz

Abstract. Evaluation of stereo-analysis algorithms is usually done by analysing the performance of stereo matchers on data sets with available ground truth. The trade-off between precise results, obtained with this sort of evaluation, and the limited amount (in both, quantity and diversity) of data sets, needs to be considered if the algorithms are intended to be used for the analysis of real-world environments. The quality of the computed disparity maps is affected by common and unavoidable noise present in real-world images. In this chapter we discus an approach to objectively evaluate the performance of stereo-analysis algorithms using real-world image sequences. The lack of ground truth is tackled by incorporating an extra camera into a multi-view stereo camera system. The proposed, relatively simple hardware set-up can easily be reproduced such that test-data sets can be generated for specific applications.

1 Introduction

Vision-based driver-assistance systems are designed to detect dangerous driving scenarios by understanding the 3-dimensional environment around the *ego-vehicle* (i.e. the mobile platform carrying the recording cameras). All the objects present in a given scene (e.g., other vehicles, pedestrians, road signs, the road itself, or the available free driving space) need to be detected and segmented; so it can later be decided wether they would represent a possible danger to the ego-vehicle.

In this chapter, we are particularly interested in *low-level* image processing (i.e., the first steps in a rather long process). Specifically, we are concerned about the evaluation of depth values detected by using *binocular stereo-matching algorithms*.

Stereo-vision algorithms generate 3-dimensional information from a given scene by matching corresponding pixels in (at least) one pair of images. Depth calculated via stereo-analysis algorithms is commonly incorporated into algorithmic pipelines for a wide variety of applications. This goes from navigation tasks for diversity of ego-vehicles, such as cars [36, 57], robots [47], forklifts [53], or wheelchairs [49], to industrial safety equipment [54] or real-time video conferencing [46].

In particular, in driver-assistance systems, stereo-analysis algorithms are included into different processes, such as *object segmentation* (e.g., objects are pedestrians or other vehicles) [30], *road modelling* [52], or *free space detection* [2].

Despite widespread acceptance of stereo-analysis algorithms as a 'fairly reliable' source of 3-dimensional data, there is still a need to develop an objective evaluation scheme that can evaluate their performance when using *real-world* images as input data. The lack of "true" measurements (i.e. for comparing with *ground truth*), represents a hard obstacle in this area, as exact camera pose detection, together with the generation of precise 3-dimensional models of uncontrolled environments, is extremely difficult.[1]

Evaluation of stereo-analysis algorithms can currently be divided into two mayor groups. *Accuracy* is measured using data with available ground truth. *Confidence* is estimated for data recorded in uncontrolled environments (without having ground truth available), for example by comparing stereo results of left-right and right-left matching.

Evaluation using data with available ground truth allows a precise comparison between true values and those obtained with the algorithms. This sort of evaluation is limited by the quantity and diversity of available data sets. Test images, along with ground truth, are generated either in laboratories under highly controlled conditions (*engineered* images) [50], or by rendering 3-dimensional modelled scenes (*synthetic* images) [60].

Engineered images challenge algorithms with real-world objects that might be known as being problematic for stereo-analysis algorithms (e.g., textureless areas, slanted planes, and so forth). But, they are limited to a few images, showing close range scenarios that are almost free of real-world effects such as multiple light sources, non-Lambertian surfaces, unexpected shadows (lighting artefacts), camera misalignment or blurring, and so forth. Scenes corresponding to common driving conditions (e.g. rainy days, busy pedestrian crossings, or different objects moving 'randomly' and at multiple distances) cannot be recorded in a laboratory.

[50] presented an evaluation and classification scheme for stereo-analysis algorithms that has been widely followed by the computer-vision community. A main contribution of this work was a data set of several stereo pairs with available ground truth.

Synthetic data sets with available ground truth have also been made available online for some years; for more recent examples, see [8, 9, 59, 60]). These computer-generated data sets allow us to test algorithms in simulated environments in which the algorithms are expected to work (e.g. see Set 2 from [13] and [59] for sequences related to driver-assistance systems). These data are usually 'multi-second driving sequences' (e.g. of more than 50 stereo frames) with simulating movement of both the virtual camera and some objects present in the scene. However, they are limited by the models followed to generate the images, the surrounding environment and the motion of the objects. They also lack of 'issues' found in images recorded in uncontrolled real-world environments with

[1] The words *true* or *truth* are used in this chapter for a particular measuring approach (e.g. manual measurements, or high-end laser-range data) considered to be "highly reliable", but with being aware that measuring always involves errors.

real cameras. Synthetic scenes are typically not yet aiming at a comprehensive physics-based modelling of cameras, lighting, or surfaces [32].

Ground truth-based evaluation is a good option for debugging, tuning of algorithms' parameters, or for exploring new matching approaches. For some applications, highly selective evaluations might be sufficient (i.e., for stereo vision in controlled environments such as automated factories or warehouses). But this cannot be expected for applications such as driver-assistance systems where stereo-vision programs have to provide reliable data on every road, under all kinds of weather conditions, and in any traffic context. According to [22], available data sets of engineered or synthetic images do only represent a very selective challenge for algorithms, with different characteristics (formally defined in [22]) compared to real-world data.

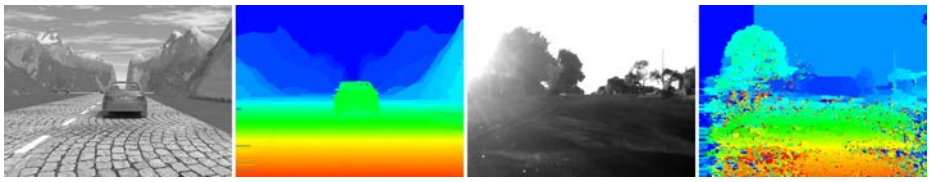

Fig. 1. Disparity maps computed with the same stereo matcher (namely BPM-CEN, to be defined further below). *Left*: Reference image of a synthetic scene from [13], and computed disparity map. *Right*: Reference image of a real-world sequence and computed disparity map. Both disparity maps are encoded from red (maximum disparity) to blue (minimum disparity). BPM-CEN shows good performance on the synthetic sequence, but fails 'totally' on the shown real-world sequence.

Figure 1 shows reference images and disparity maps for a synthetic (left) and a real-world (right) sequence. Both disparity maps were generated with the same stereo algorithm (graph-cut stereo with census as cost function; see Section 3) using exactly the same parameters. For the 'synthetic' disparity map it is easy to recognize all the objects present in the scene. For the real-world case, a lot of details is missing (e.g. the two trees on the left are merged into a single object) and a lot of noisy measurements is introduced.

The question is, how to ensure that evaluation, done for real-world sequences without available ground truth about scene geometries and *ego-motion* (i.e trajectory of the ego-vehicle), still can use some kind of objective testing? Ego-motion may be understood to some degree using visual odometry [28]. What other information can be provided to ensure objective testing?

One of the first evaluation schemes, by setting-up a testbed (data set and evaluation criteria) with real-world stereo pairs, is reported in [21]. The author provided twelve pairs of images to selected 60+ research groups worldwide. Those images were aerial views of real-world locations. For evaluating the calculated stereo measurements, manual checking ("truth on the ground", also known as

ground truth) was performed using an *analytical plotter* (for around 50% of provided measurements).

A similar test bed was proposed in [6], the *JISCT* data set of 49 images (still available on [29]). Most of them are real-world images, but there are also engineered and synthetic stereo pairs. However, none of them came with ground truth. The initial study involved only five research groups. The evaluation was based on a "reported value and unreported value" approach, i.e. whether the algorithm reported a value in (a manually) selected region where a measurement was feasible to be calculated.

Some other methods have been proposed to evaluate stereo-analysis algorithms in the absence of ground truth. In [14], the authors calculated (manually) true depth values at 200 randomly selected points.

In [4], the evaluation was done by measuring the number of "successfully" matched pixels using a left-right consistency check [27]. Confidence measures are another example of evaluation in the absence of ground truth [20, 45]. The idea is to measure the reliability of the calculated values for each pixel using heuristic or probabilistic approaches. Some techniques, specifically designed for driver-assistance systems, were proposed in [40, 55]. Both evaluation schemes can only be applied if certain conditions are satisfied during the recording of the scenes. Recently, a conference [10] provided its own evaluation test-data set. However, there was no provided ground truth or an objective evaluation scheme for comparing results.

Generating ground truth for outdoor environments has also been investigated in [5, 44] by providing "true" distance measurements using a high-end *laser-range finder*. Despite the accuracy of the measurements, the reduced resolution of the laser-range finder (compared to the camera) and the elevated prices of the laser-range finder is still limiting the applicability of this option.

Intermediate extra images have been used in [56] for defining a *prediction error* for optic flow algorithms: based on optic flow calculated for images t and $t+2$, a virtual view is interpolated for image $t+1$ and compared with the actually provided image $t+1$.

Similarly, [42] proposed to used a third camera (which records a *control image*) for analysing stereo-analysis accuracy for recorded image sequences: depth data calculated at time t for the "left" and "right" camera are used to warp the image (say) of the left camera into the pose of the third camera. This virtual view is compared with the actually recorded view in the third camera at time t. (Of course, such a 3-camera set-up can also be generalised; important is to have one additional view for comparison. It can be extended to trinocular stereo-analysis, with a fourth control camera for comparison, and so forth.)

Following [56], this kind of performance analysis, characterized by the use of additional input data, is called *prediction error evaluation*. See also [3] for a discussion of using at least three images of the same scene.

In this chapter we discuss this approach as defined in [42], provide more detailed ways to compare virtual and recorded view, discuss particular experiments, and summarise altogether our experience with this approach since 2009.

In short, it offers an objective way to evaluate stereo-analysis algorithms on real-world image sequences. We limit our discussion for the use of three cameras only, called *reference*, *match*, and *control camera*. The third image of the control camera can be see as being the ground truth in this case. Due to the sequence character of the proposed approach, statistical analysis is performed to analyse the matching algorithms.

The warping of the *reference image* of a given stereo pair into a *virtual image* is defined by the calibrated camera geometry, and important is to note that we are not aiming at producing a "nice" warped image; we are just mapping intensity data of the reference image onto the nearest pixel in the image plane of the third camera (possibly overwriting previously mapped values). The control camera should also not be in a pose which supports similarities between virtual and control image (e.g. as it would be the case if the control camera would be positioned between reference and match camera).

A key issue of the method is the selection of the measure for the prediction error for comparing the sequence of virtual images with the sequence of control images (i.e. not only for a few triples of images but for long trinocular image sequences). The use of long sequences allows us to investigate the influence of changes in conditions when recording the stereo image data on the performance of the algorithms (e.g., local brightness variations between reference and match image, changes in scene geometry, camera issues, or lighting variations).

The main advantage of the proposed stereo-sequence analysis approach is that the required hardware setup can be easily reproduced. Based on today's time efficiency of stereo matchers, it can be used for real-time evaluations, and thus also as a basic module for designing an adaptive computer vision system for vision-based driver assistance (as discussed in [32]).

For experiments, to be reported in this chapter, we selected eight sequences of 400 trinocular frames each, recorded in different scenarios (for a characterization of the selected sequences, see [32]). We aim at illustrating to use of the proposed method for understanding behaviours of stereo-analysis algorithms depending on given input image sequences.

This chapter is structured as follows: In Section 2 we describe the generation of the virtual image and discuss the position of the control camera. This section also discusses the selected prediction-error measure. In Section 3 we briefly identify the stereo-analysis algorithms that are used for the presented experiments. For selected trinocular sequences and a discussion about obtained evaluation results, see Section 4. Section 5 concludes.

2 Approach

Consider time-synchronised recording of a scene by three video cameras. Video data captured by *reference* and *match* camera are rectified in such a way that each stereo pair I_r and I_m satisfies the *standard stereo geometry* (SSG) (e.g., as defined in [33]). The third camera acts as *control camera* and is potentially in arbitrary pose "towards the scene recorded by reference and match camera".

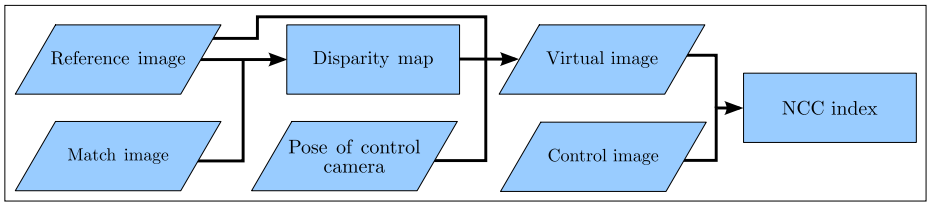

Fig. 2. Sketch of the followed approach. The NCC index is calculated between the generated virtual image and the recorded control image.

The objective is to generate a *virtual image* I_v from a disparity map calculated by a stereo-analysis algorithm (using the stereo frames from the reference and match camera), and to compare I_v with the *control image* I_c recorded with the control camera.

We generate I_v by mapping (warping) the pixels of the reference image I_r into the locations where they would have been recorded in I_c. Then, I_c and I_v are compared using *normalized cross-correlation* (NCC) as a measure; see Section 2.3 for its specification. Figure 2 summarizes the followed approach.

2.1 Common Forward Equations

Assume that the coordinate system of the reference camera is identified with the world coordinate system. Image coordinates are defined by each camera individually. Locations of reference, match and control camera are sketched in Figure 3. In world coordinates, the optical centre of the reference camera lies at the origin $O_r = (0, 0, 0)^T$, that of the match camera at $O_m = (b, 0, 0)^T$, and that of the control camera at $O_c = (b_1, b_2, b_3)^T$.

Let $P = (X, Y, Z)^T$ be a scene point in the shared field of view of all the three cameras; and $p_r = (x, y)^T \in I_r$, $p_m = (x_m, y_m)^T \in I_m$, and $p_c = (x_c, y_c)^T \in I_c$ be the projections of P onto the rectified image planes of the three cameras. The corresponding image point in the virtual image is denoted by $p_v = (x_v, y_v)^T$.

For the assumed case of standard stereo geometry between reference and match image, we provide a formula below to obtain the coordinates of p_v in terms of the coordinates of p_r, and the internal parameters of the stereo camera defined by the reference and match cameras (i.e., base-line distance b and unified focal length f) and the corresponding disparity value d (computed by some stereo-analysis algorithm) between p_r and p_m. Since P is visible from reference and match camera, by triangulation, it is possible to write the coordinates of P with respect to the coordinate system of I_r as follows:

$$(X, Y, Z)^T = \frac{b}{d}(x, y, f)^T \tag{1}$$

Now, let $(X_c, Y_c, Z_c)^T$ be the coordinates of P with respect to O_c. Using homogeneous coordinates and (for abbreviation) letting **C** and **S** be short for *cosine*

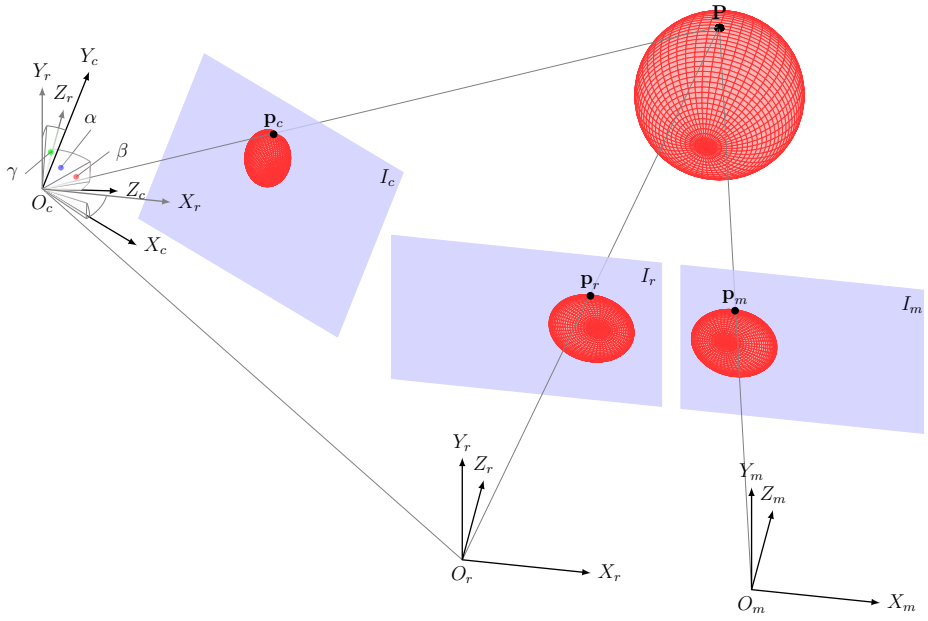

Fig. 3. A general trinocular camera configuration. The two cameras (represented by their coordinate systems) on the right are assumed to satisfy the standard stereo geometry, the third camera is rectified with respect to internal camera parameters only (i.e. thus representing ideal central projection).

and *sine* functions, respectively, the matrix

$$M = \begin{pmatrix} \mathbf{C}\gamma\mathbf{C}\beta & -\mathbf{C}\gamma\mathbf{S}\beta\mathbf{S}\alpha - \mathbf{S}\gamma\mathbf{C}\alpha & \mathbf{S}\gamma\mathbf{S}\alpha - \mathbf{C}\gamma\mathbf{S}\beta\mathbf{C}\alpha & -u_1 \\ \mathbf{S}\gamma\mathbf{C}\beta & \mathbf{C}\gamma\mathbf{C}\alpha - \mathbf{S}\gamma\mathbf{S}\beta\mathbf{S}\alpha & -\mathbf{S}\gamma\mathbf{S}\beta\mathbf{C}\alpha - \mathbf{C}\gamma\mathbf{S}\alpha & -u_2 \\ \mathbf{S}\beta & \mathbf{C}\beta\mathbf{S}\alpha & \mathbf{C}\beta\mathbf{C}\alpha & -u_3 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{2}$$

specifies the mapping

$$(X_T, Y_T, Z_T, 1)^T = \boldsymbol{M} \cdot (X, Y, Z, 1)^T \tag{3}$$

where angles α, β, and γ are as in Figure 3, and

$$u_1 = b_1\mathbf{C}\gamma\mathbf{C}\beta + b_2(-\mathbf{C}\gamma\mathbf{S}\beta\mathbf{S}\alpha - \mathbf{S}\gamma\mathbf{C}\alpha) + b_3(\mathbf{S}\gamma\mathbf{S}\alpha - \mathbf{C}\gamma\mathbf{S}\beta\mathbf{C}\alpha) \tag{4}$$

$$u_2 = b_1\mathbf{S}\gamma\mathbf{C}\beta + b_2(\mathbf{C}\gamma\mathbf{C}\alpha - \mathbf{S}\gamma\mathbf{S}\beta\mathbf{S}\alpha) + b_3(-\mathbf{S}\gamma\mathbf{S}\beta\mathbf{C}\alpha - \mathbf{C}\gamma\mathbf{S}\alpha) \tag{5}$$

$$u_3 = b_1\mathbf{S}\beta + b_2\mathbf{C}\beta\mathbf{S}\alpha + b_3\mathbf{C}\beta\mathbf{C}\alpha \tag{6}$$

Let m_{ij} be the element at position (i, j) in matrix $\boldsymbol{M}$, for $1 \leq i, j \leq 3$. Let f_c be the focal length of the control camera. Thus, using the equations defined

by central projection, we have that

$$x_v = f_c \cdot \frac{m_{11}(bx - db_1) + m_{12}(by - db_2) + m_{13}(bf - db_3)}{m_{31}(bx - db_1) + m_{32}(by - db_2) + m_{33}(bf - db_3)} \tag{7}$$

$$y_v = f_c \cdot \frac{m_{21}(bx - db_1) + m_{22}(by - db_2) + m_{23}(bf - db_3)}{m_{31}(bx - db_1) + m_{32}(by - db_2) + m_{33}(bf - db_3)} \tag{8}$$

where d and b were defined above as being the disparity between pixels p_r and p_m and the length of the baseline between reference and match camera, respectively.

With these two *forward equations* (e.g., see [35]) it is possible to map any pixel location $(x, y)^T$ in the reference image into a pixel $(x_v, y_v)^T$ in the image plane of the third camera. We select the nearest pixel position in this virtual image (i.e. in the pose of the third camera) because we do not aim at any visual improvement of this mapping (e.g. by interpolation of pixel values).

2.2 Poses of the Third Camera

In this section we discuss possible poses of the control camera. Note that the pose of the control camera defines the final appearance of the generated virtual image. The three cameras can be in an arbitrary position, but constrained by the fact that reference and match images need to satisfy SSG after rectification. In the following we denote the reference camera also as being the *left camera* of this pair of two rectified cameras.

In order to reduce the number of occluded points between reference and control camera, we aim at having the focal point of the control camera collinear with the focal points of the two other cameras. We discuss possible poses of the control camera in such a *horizontal configuration*.

Occluded points may cause areas with no texture in I_v, or pixels from I_r being mapped onto the wrong position due to having erroneous disparity results for occluded pixels in the stereo pair. We illustrate this by examples generated using available ground truth for the synthetic sequence No. 1 from Set 2 of [13], see [60].

By increasing the translational distance between the poses of the control and the stereo-camera system, more occluded areas occur on I_v. Occlusions could be reduced (in general) by having the control camera positioned between reference and match camera. Figure 4 shows three different occlusion cases. For this figure we vary the poses of an imaginary third camera with respect to the used poses of reference and match camera when rendering this sequence No. 1 from Set 2. The disparity map I_d is the available ground truth.

On the left, the figure shows the virtual view corresponding to the pose of the reference image (i.e. the third camera was assumed to coincide with the reference camera). White pixels represent occluded pixels between reference and match image. Obviously, no disparity information is available for those. They are already occluded with respect to both stereo cameras. For the centre image of the figure, the third camera moved into the pose of the match camera. Occlusions are now shown in black, and correspond to occluded pixels between reference

Fig. 4. Different types of occlusions for a horizontal configuration. *Left*: white pixels indicate occlusion between reference and match camera. *Centre*: black pixels indicate occlusion between third and reference camera (here: third camera is at position of match camera). *Right*: combined visualization where third camera is now left of the reference camera.

and control camera. The virtual view generated for a pose to the right of the reference image (in a horizontal configuration) would tend to "cover" also such occluded pixels that are visible for the reference camera but not for the match camera. On the right, the figure shows a virtual view based on the pose of the third camera located to the left of the reference camera. It is an example of a virtual view in which both kinds of occlusions occur (white and black).

For the first configuration there are no occlusions between reference and control camera. This configuration is actually known in self-consistency studies [39]. However, we are interested into using an additional image for the evaluation, not yet involved in the given stereo analysis, thus allowing us to obtain additional insights into the performance.

A symmetric pose of the control camera (focal point half-way on baseline between reference and match camera, and perpendicular bisector incident with optical axis of control axis) is also expected to minimize the impacts of both types of occlusions (i.e., the total number of either black or white pixels). In performance evaluation, it would be ideal to separate the impact of occlusions from those of incorrect stereo matching. Thus, the symmetric case seems to offer the possibility to focus on disparity errors. However, errors due to mismatches are actually often not as "obvious" for the symmetric case compared to a third-camera pose which differs (much) from the symmetric case.

Thus, altogether, an in-depth statistics about error distributions for different third-camera poses in a horizontal configuration (e.g. depending on scene geometry) might be of interest. However, in our practical tests we realized quickly that having the third camera in a "different pose" compared to the stereo-camera pair, but still "reasonably close" to this pair for not having to many occlusion issues, provides a better "challenge" than having a symmetric camera set-up.

The experiments reported in this chapter had the control camera approximately 50 cm to the left of the reference camera; reference and match camera are about 30 cm apart. This translational distance between control and reference camera appeared to be large enough for detecting miscalculated disparity values

(even if disparities are small), but is still not yet exaggerating the influence of occluded pixels.

2.3 Evaluation Index

As evaluation index we calculate the normalized cross correlation index between the virtual image I_v and the control image I_c, for each trinocular stereo frame at time t in a given image sequence. The NCC index is given by

$$\mathrm{NCC}(I_c, I_v) = \frac{1}{|\Omega|} \sum_{(x,y)\in\Omega} \frac{\big[I_c(x,y) - \mu_c\big]\big[I_v(x,y) - \mu_v\big]}{\sigma_c \sigma_v} \tag{9}$$

where μ_c and μ_v denote the means, and σ_c and σ_v the standard deviations of the control and virtual images, respectively.

The set Ω is a subset of all pixel locations. It needs to be selected for defining a "meaningful measure". The default approach is that Ω is simply defined by pixels having a valid disparity.[2] $|\Omega|$ denotes the cardinality of this set.

The NCC index appears to be convenient for the presented evaluation approach (rather than, e.g., just a sum of absolute intensity differences), as it handles photometric differences between reference and control image to some degree, and brightness variations (e.g. non-uniform in a recorded image) are actually very typical for recorded outdoor videos.

2.4 Alternative Approach for Defining Set Ω

Images recorded in the context of driver-assistance systems typically contain two large nearly textureless areas (i.e., featureless regions), namely the sky and the road. State-of-the-art stereo-analysis algorithms often have difficulties for calculating correct disparities in such uniform regions – and the results are actually not so important. Values for the sky are not important at all, and invalid values on the road (if properly detected) can be interpolated for identifying the road manifold.

We notice that the defined evaluation approach might report a good performance in such homogeneous regions even if this is not the case. In such regions it is very likely to occur that a pixel in the reference image with a corresponding miscalculated disparity value is mapped into a pixel in the virtual image that is in the same textureless region (i.e., a region with insignificant intensity differences between its pixels). Thus, values in this region may incorrectly influence the final evaluation index.

Figure 5 shows two virtual images and corresponding disparity maps when using the BPM-CEN stereo matcher (defined later in Section 3) for two consecutive stereo frames (Frames 326 and 327 in the *barrier sequence*). A rectangular region is selected in the middle of the road; it shows differences in miscalculated

[2] Our stereo-analysis algorithms assign a non-positive value to any pixel having no valid disparity.

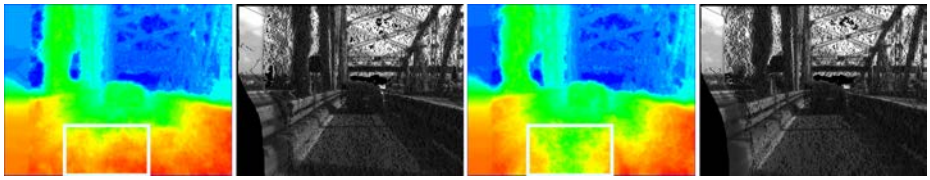

Fig. 5. Samples of disparity maps and corresponding virtual images from consecutive frames from the *barrier sequence*. Both disparity maps show difference in disparity values in the indicated rectangular region, but the corresponding regions in the virtual images look almost the same. Thus, the NCC measure is expected to lead to about the same value in those regions. See Table 1 for the NCC values of the whole images, and of the rectangular regions only.

	Frame #326	Frame #327
NCC value for rectangular window	94.5	90.0
NCC for the whole image	85.7	85.6

Table 1. Evaluation results for a selected rectangular road region compared to the NCC index for the whole image, for two frames illustrated in Figure 5.

disparities in both frames. However, the corresponding regions in the virtual images appear to be almost identical. For frame #326, disparity values in the rectangle are between 28 to 56, and between 21 to 41 for frame #327. For road surface points, this implies an average distance difference of about 5 meters. The evaluation index, restricted to the rectangle, does not show this defect, and it is considerable high compared to the NCC index calculated for the whole image (see Table 5).

The following modified definition of set Ω aims at restricting the performance evaluation to areas being "rich in texture". The basic idea is as follows. Miscalculated disparities at, or within a small distance to pixels with a significant *intensity gradient* (used as a simple texture criterion) should affect the NCC index more than miscalculated disparities in textureless regions. One option is to simply discard the homogeneous regions completely when calculating the NCC index.

Given an image I, we generate a mask I_k that will shrink the domain Ω by eliminating textureless regions. The image I_k is produced in three steps. First, a *binarized gradient image* ∇I is defined as

$$\nabla I(x,y) = \begin{cases} 0, & \text{if } \left|(\partial_x I(x,y), \partial_y I(x,y)\right|_2 > T_1 \\ 1 & \text{otherwise} \end{cases} \tag{10}$$

where ∂_x (or ∂_y) denote the partial derivative in the lateral (or vertical) direction.[3] The sign $|\cdot|_2$ denotes the L_2-norm and T_1 is an adjustable threshold. With ∇I we aim to identify regions with some changes in intensity values.

[3] We use central differences [35] to approximate the partial derivatives.

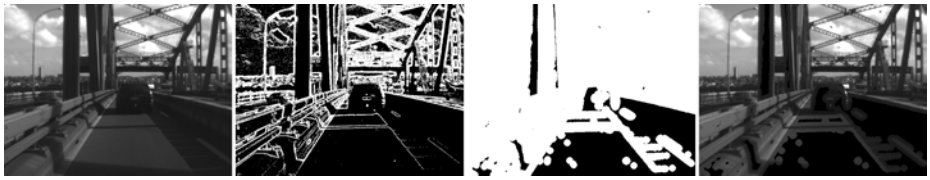

Fig. 6. Illustration of mask generation. From *left* to *right*: original image I, gradient image ∇I, distance mask I_e, and identified textured zones in I.

The second step uses Euclidean distance transformation for generating an image I_e that labels pixels by their L_2-distance to edge pixels identified by ∇I. Finally, we define I_k as

$$I_k(x,y) = \begin{cases} 0, & \text{if } I_e(x,y) > T_2 \\ 1 & \text{otherwise} \end{cases} \tag{11}$$

where T_2 is again a predefined threshold. For the experiments reported in this chapter, we use the control image to define the mask I_k, with $T_1 = 5$ and $T_2 = 10$.

Figure 6 illustrates the process of generating the distance image I_e. The leftmost image is a control image I. The next image shows ∇I with $T_1 = 5$, followed by the distance image I_e using $T_2 = 10$. The resultant "masked" control image is shown in the rightmost position.

Alternatively, the distance values in I_e could be used as weights when defining the NCC index. However, experiments showed that using the defined mask I_k helps to calculate NCC indices which correspond, in general, with subjective visual evaluations of calculated depth accuracies.

3 Tested Stereo-Analysis Algorithms

We are interested in stereo-analysis algorithms for outdoor scenes in the context of driver-assistance systems and related applications; see, for example, [32, 49]. The diversity of recording situations (e.g. in the night, in rain, with lighting artefacts) basically implies that one particular algorithm or parametrization cannot be the all-time winner; and some kind of adaptation needs to be supported.

3.1 Three Matchers

For the experiments to be reported in this chapter, we selected three dense stereo-analysis algorithms based on techniques that have shown a good performance in previous studies [42, 44]. We test them for three different cost functions.

Belief-Propagation Matching (BPM)

We use a max-product iterative *belief propagation* algorithm as presented on [15]. This algorithm uses a truncation parameter for both, the cost function and the smoothness term. The smoothness term is a truncated quadratic

	BPM					GCM				SGM	
	dMax	sMax	λ_d	iteration	level	λ_1	λ_2	threshold	K	c_1	c_2
SAD	100	500	0.3			4.2	1.4	1	7		
CEN	75	600	0.6	7	6	3	1	1	5	30	150
EPE	33	200	0.225			2.6	0.86	16	4.33		

Table 2. Parametrization of the used stereo-analysis algorithms. BPM (SGM) uses identical values for number of iterations (c_1) and level of tree (c_2) for the three different cost functions SAD (sum of absolute differences), CEN (census function), or EPE (end-point error).

function, which allows to obtain a smooth disparity map but without penalizing depth discontinuities to much. Message passing is based on 4-adjacency. The original source code on [15] was modified to allow the use of different cost functions and of 10 bit input images as stereo frames.

To speed up the matching process, a hierarchical approach (i.e. a coarse-to-fine approach) is considered such that the passing of messages is more efficient when staying with a reduced number of iterations. The truncation parameters for the data (dMax) and the smoothness (sMax) terms, the weighting factor for the data term (λ_d), the number of iterations (iteration), and number of levels (level) of the followed hierarchical approach are shown in Table 2.

Graph-Cut Matching (GCM)

We use a modification of the *graph cut*-based algorithm presented in [7]. For minimizing the energy function, a randomly initialized disparity map is considered as a weighted graph. The optimum disparity map is then calculated using the *α-expansion method* [37]. The implementation of this algorithm uses as smoothness term the binary Potts model to assure that a global minimum can be reached [37]. The three parameters required for defining the Potts model ($\lambda 1$, $\lambda 2$, and the threshold) and the weighting factor for the cost function (K) are summarized on Table 2.

As for BPM, this algorithm was also modified such that a wider range of cost functions could be used.

Semi-Global Matching (SGM)

We also use a semi-global matching algorithm as introduced in [27]. The matching strategy followed by BPM or GCM can be characterized as being potentially *global* (but practically limited by the number of iterations). In contrast, SGM limits its search space to a predefined set of *paths* to obtain an optimum disparity value only with respect to this selected search space. The used SGM implementation has been reported in [25]. We are interested in evaluate two configurations that have been recently proposed; all of them using the census transform as cost function (see Section 3.2). We compare a 4- and (the "usual") 8-path configurations as suggested in [25]. We also examine the performance of

the hierarchical approach presented in [26]. Using four instead of eight paths reduces the computational time, and the quality of the disparity maps should not be drastically affected. The hierarchical approach increase the quality in the matching in areas that are known to be complicated (non-textured areas like the road in DAS) for SGM. The selected values for the two fixed penalties for the smoothness term (c_1 and c_2) are summarized in Table 2.

3.2 Three Cost Functions

Three cost functions are considered for our experiments. Each of them analyses different "characteristics" of the stereo input images when calculating costs for assigning a disparity value to a given pixel. Two of them, the *census* and a *gradient-based* (such as EPE), have been previously identified as being robust in outdoor environments; see [25, 23]. The impact of the third function, the common *sum of absolute differences*, depends on photometric consistency between both images in a given stereo pair.

Census Transform (CEN)

The census transform [62] is defined by the *Hamming* distance between two *signature vectors*. Its use supports robustness of a stereo matcher against common types of noise found in real-world images. Given an arbitrary image I and a neighbourhood $\mathcal{N}$ of a pixel at location $(x, y)^T$ [denoted as $\mathcal{N}(x, y)$], the ith coordinate of the signature vector (of dimension $|\mathcal{N}| - 1$) of $I(x, y)$ is defined as follows:

$$\mathrm{sig}\big(I(x,y)\big)_i = \delta\big(I(x',y')\big) \tag{12}$$

where $(x', y')^T \in \mathcal{N}(x, y) \setminus \{(x, y)^T\}$, and

$$\delta\big(I(x',y')\big) = \begin{cases} 0, \text{ if } I(x,y) \neq I(x',y') \\ 1 \text{ otherwise} \end{cases} \tag{13}$$

The order in which the coordinates of the signature vectors are arranged is irrelevant, but needs to be consistent. Following [25], we use a 9×3 neighbourhood as it favours a stronger data contribution along the epipolar line.

The comparison of the signature vectors is made coordinate-wise using the Hamming distance. For reference image I_r and match image I_m, the cost of associating a disparity d to pixel $I_r(x, y)$ is given by

$$\mathrm{CEN}\big(x,y,d\big) = \sum_{i=1}^{|\mathcal{N}|-1} \begin{cases} 0, \text{ if } \mathrm{sig}\big(I_r(x,y)\big)_i = \mathrm{sig}\big(I_m(x-d,y)\big)_i \\ 1 \text{ otherwise} \end{cases} \tag{14}$$

where $\mathrm{sig}(\cdot)_i$ denotes the ith coordinate of the signature vector sig.

A Gradient-Based Cost Function (EPE)

The selected gradient-based cost function [31] analyses the L_1-distance between the end-points of the gradient vectors. This distance is expected to have

a good performance when using real-world data [23]. To calculate the discrete partial derivatives that define the gradient vector, again we use central differences.

The cost of assigning disparity d to the pixel $I_r(x,y)$, using the end-point error (EPE) cost function, is given by

$$\mathrm{EPE}\big(x,y,d\big) = \Big|\big(\partial_x I_r(x,y), \partial_y I_r(x,y)\big) - \big(\partial_x I_m(x-d,y), \partial_y I_m(x-d,y)\big)\Big|_1 \quad (15)$$

The symbol $|\cdot|_1$ denotes the L_1 norm; we use this norm to compare the estimated end-points of the gradients.

Sum of Absolute Differences (SAD)

The sum-of-absolute-differences (SAD) cost function is an intensity-based similarity measure. It is known for its poor performance when it comes to real-world stereo sequences, as the photometric consistency assumption is commonly violated in those data. We are interested in reconsidering this commonly used statement. The SAD cost of assigning the disparity d to $I_r(x,y)$ on the reference image is given by

$$\mathrm{SAD}\big(x,y,d\big) = \frac{1}{|\mathcal{N}|} \sum_{(x',y')^T \in \mathcal{N}(x,y)} \big|I_r(x',y') - I_m(x'-d,y')\big| \quad (16)$$

where $|\mathcal{N}|$ denotes the cardinality of the used neighbourhood. In the experiments, we use the $8-$neighbourhood for $\mathcal{N}$.

4 Experiments

We evaluate the performance of the three selected stereo-matchers using the three specified cost functions for BPM and GCM; the three presented configurations for SGM use just CEN as a cost function. We use the abbreviations BPM-$*$ or GCM-$*$, where $*$ denotes CEN, EPE, or SAD, and SGM-4, SGM-8, and SGM-HIER for the configurations of the semi-global matcher.

4.1 Evaluation Domains

The *full approach* refers to the method introduced at the beginning of Section 2; the *masked approach* denotes the method discussed in Section 2.4.

BPM and GCM algorithms generate usually a valid disparity (no matter whether correct or incorrect) for almost every pixel in the reference image. Thus, we compare the whole virtual and control image (except for the obviously occluded regions at the left margin of both images). As we are using the same evaluation domain, it is fair to compare the evaluation indices of those two algorithms (i.e. the *boosting effect* from the non-textured regions described in Section 2.4 should affect equally to indices of both algorithms).

Fig. 7. Sample frames from the used 400-frame long trinocular sequences. *Top row*, from left to right: *midday sequence*, *wipers sequence*, *dusk sequence*, and *night sequence*. *Bottom row*, from left to right: *queen-street sequence*, *people sequence*, *harbour-bridge sequence*, and *barrier sequence*.

For SGM, the evaluation domain is defined by the pixels in the disparity map detected as being valid (usually around 60% of the whole image domain). Thus, we only compare results between the three SGM configurations as their disparity maps have a similar amount of valid pixels.

4.2 Data Sets

Regarding the experimental data set, we use eight long (400 trinocular frames each) sequences recorded on real-world environments with test vehicle *HAKA1* (see [32]), thus 9,600 test images in total, each of 640×480 resolution at 10 bit per pixel.

All the three cameras were firmly mounted on a metal bar (behind the windshield) about at the same height as the rear-view mirror. The reference and match cameras were placed on the driver's side of the vehicle. The length of the baseline is about 30 cm, thus, we are able to calculate distances to objects located from just less than 5 meters to the cameras, up to around 310 meters away (i.e., for a disparity value of 1). The control camera was fixed to the left of the rear-view mirror, at around 50 centimetres away from the reference camera. With this set-up we tried to keep the common field of view as large as possible. By keeping a considerable distance between reference and control cameras we support that appearing errors become more evident in the calculated NCC evaluation indices along a test sequence.

The cameras used for recording the sequences were all of the same brand and model (Point Gray Firefly MV[4]); with identical (micro) lenses with a fixed focal length of 6 mm.

Four of our sequences were recorded on the same street (*the reference street*) under different environmental conditions. The street is surrounded by trees such

[4] `http://www.ptgrey.com/products/fireflymv/index.asp`

that illumination artefacts [61] are present in the images, especially if the sun is low on the horizon. There are some road signs and power poles. The surface of the road has actually sufficient texture so it is expected that the road will not be a source of noise during the matching process. The road has several up- and down-hill segments, which makes the sequences also a challenging scenario for other analysis algorithms (e.g., road modelling, see [52]).

The other four sequences were recorded in more dynamic environments. They were recorded on busy roads, where pedestrians and vehicles are part of the scenery. Two of the sequences were recording while driving at about 80 km/h, to test algorithms also for highly dynamic environments. The sequences are available for download in Set 9 from [13].

A brief description of the sequences is as follows:

Midday: This sequence was recorded in the reference street under 'ideal' conditions. The sun was close to its zenith, so there are not many of the undesired illumination artefacts. There is no incoming or oncoming traffic. The idea of recording such a simple sequence is to have a reference sequence, where the algorithms should perform best. Thin structures around the road (e.g., poles, trees with branches, road signs) still make it a challenging test sequence.

Wiper: In order to gain experience on the influence varying occlusions of some regions in one (or both) camera(s) of the stereo system, we recorded a sequence while the wipers have been switched on (but no rain). This sequence was recorded within just a few minutes past the midday sequence on the same default road, expecting that the only "differing" factor for the matching process is the moving wipers.

Dusk: This sequence was recorded wile having the sun in a position close to the horizon. The idea was to try to simulate the very common situation of having large saturated areas in one or in both cameras. As the road is surrounded by trees, there are intervals in the sequence with or without the sun striking directly into the cameras. The shadows of the trees are projected onto the road, and this offers another kind of illumination artefacts.

Night: This sequence was recorded at night. Almost only light provided by the headlamps of HAKA1. The trees around the road covered almost all the light from the lamp posts, which are very sparse in this particular road. The intention of having such a dark night scene was to simulate driving conditions as faced on second-order highways or rural roads.

Queen: This sequence was not recorded on our reference road, but on a main road of Auckland city. It has both, moving and static cars and pedestrians. It was recorded while driving towards a set of traffic lights, with a stop there. There are moving pedestrians at different distances. A bus stopped on the right hand side and has interesting reflections in its windows.

People: This sequence was recorded while HAKA1 was standing still in front of a pedestrian crossing. The sequence has varying numbers of pedestrians in the scene, between 1 up to around 20 at a time. The pedestrians walk only in two directions.

Harbour: This sequence was recorded while driving across Auckland's harbour bridge. The road on the bridge its surrounded by a structure of thin metal poles.

Barriers: This sequence was also recorded while driving across this harbour bridge. In this case the recording vehicle was driving in a lane that it is enclosed by medium-height concrete bars, and also the metal structure of the bridge is further up.

4.3 Results and Discussion

The discussion in this chapter is focused on the most remarkable details of obtained results (e.g., when severe changes in the NCC index were detected, or when results between algorithms are particularly different). The average NNC indices for full or masked approaches, for all sequences and algorithms, are presented in Tables 3 or 4, respectively.

Column 'Win' in the tables shows the total number of frames on which a certain configuration outperformed the others configurations of the same matcher. In the discussion below we may compare the results of all the BPM and GCM configurations directly; but consider separately the results of SGM because the image domain Σ used for BPM and GCM is different to that used for SGM.

Midday: All the algorithms performed 'fairly well' (as expected); the indices reported for this sequence are the highest among all the sequences used in this chapter. See the left image in Figure 4.3 for the SGM results. Interestingly, all the algorithms had local minima at about the same frames within that sequence. Sudden drops in indices are caused by thin structures that surround the road in

		NCC Average								Win
		Barriers	Dusk	Harbour	Midday	Night	People	Queen	Wiper	
BPM	CEN	62	74	63	73	41	61	66	69	17
	EPE	**66**	50	**70**	**91**	**64**	**68**	**80**	**87**	**1449**
	SAD	56	**87**	59	91	63	66	79	86	82
GCM	CEN	**59**	**87**	**62**	**93**	41	**66**	**82**	**89**	**1479**
	EPE	37	82	38	88	21	42	67	83	1
	SAD	40	82	40	60	**62**	62	78	85	172
SGM	CEN4	76	92	80	95	86	79	88	92	154
	CEN8	76	92	80	95	87	79	**89**	92	433
	HIER	**76**	**95**	**81**	**96**	**90**	**79**	89	**94**	**2613**

Table 3. Mean values of NCC indices, rounded to nearest integer, for full analysis. For each sequence, that cost is in bold where the algorithm performed best. The last column (Win) shows for how many frames the specific configuration performed best. We may compare BPM and GCM results directly, but separately from SGM, as image domains used for evaluation are different.

		NCC Average Mask								Win
		Barriers	Dusk	Harbour	Midday	Night	People	Queen	Wiper	
BPM	CEN	57	66	56	69	42	59	65	63	24
	EPE	**61**	**75**	**64**	**79**	**64**	**66**	**75**	**73**	**1184**
	SAD	51	69	53	77	64	64	73	70	83
GCM	CEN	**54**	**76**	**54**	**82**	43	**66**	**78**	**77**	**1743**
	EPE	33	64	32	67	28	39	57	60	0
	SAD	35	66	33	50	**63**	61	71	67	166
SGM	CEN4	31	26	19	33	**34**	51	47	27	353
	CEN8	**34**	33	**25**	38	31	**53**	**53**	33	**1434**
	HIER	31	**41**	24	**45**	26	48	47	**49**	1413

Table 4. Mean values of NCC indices, rounded to nearest integer, for masked analysis. For the meaning of bold numbers, and of the last column (Win), see caption of Table 3.

those frames; miscalculated disparity values for power poles or road signs had a particularly bad effect on NCC indices at those frames. This became even more obvious by using the masked analysis.

For GCM, the leading configuration was GCM-CEN. For the other two configurations, there are several regions with obvious (i.e. visual inspection) disparity miscalculations, and they were correctly penalized with both the full and the masked analysis.

BPM-SAD and BPM-CEN reported the best (and very close to each other) NCCC indices for BPM. For this matcher, the CEN cost function introduced a kind of a "salt and pepper" noise into the disparity maps that was also clearly identified by our evaluation.

SGM-HIER showed a slightly better performance than the other two configuration for full analysis. This could be due to a better performance of SGM-HIER

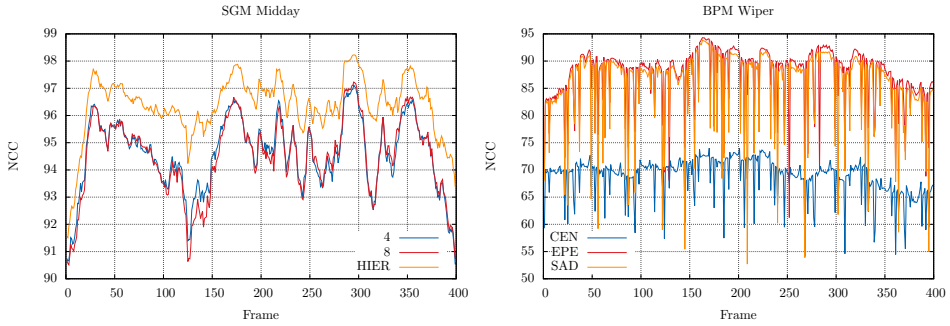

Fig. 8. *Left*: Midday sequence results for SGM. *Right:* Wiper sequence results for BPM.

on road regions (see the identified boosting effect of trinocular analysis). However, the same rank was observed when using the masked analysis.

Wiper: This sequence represents a particular challenge for trinocular analysis. The wipers might not be visible in the stereo results, but they might be still visible in the control image. Thus, in this sequence, low NCC indices might not only be caused by miscalculated disparity values, but also due to having different objects (wipers) present in the virtual and the match image.

All the algorithms (with all the configurations) show a repetitive pattern of local minima, as expected. When the wipers are not present in any of the images, the algorithms performed just as in the midday sequence. Lowest local minima correspond to frames where the wipers were just visible in the control image. For cases where there is a wiper in the stereo-image pair, but not in the control image, the algorithms handled the wipers as invalid pixels (SGM) or by propagating estimated disparity values of surrounding areas (BPM and GCM); this was more evident with the CEN and EPE functions.

Masked and full analysis led to similar results. When considering the masked approach, the local minima are not as low as in the full analysis. Miscalculations introduced by the wipers affected more the sky and road areas (we recall: both regions are ignored when using the masked approach).

For BPM and GCM, the most negatively affected cost function is SAD; drops in magnitudes of indices were the largest compared to other cost functions. BPM-CEN was noticeably robust against the presence of wipers. Figure 4.3, right, shows the performance of the three BPM-configurations. There are large drops of indices for BPM-SAD. For SGM, 8- and 4-path configuration show an almost identical performance. SGM-HIER outperforms both of them. Local minima are not as low for SGM-HIER as for the other two configurations.

Dusk: The results obtained with this sequences are as expected: the performance of all algorithms decreases when the sun strikes directly into the cameras. It can be noticed that there are two intervals with particular low indices, namely at the beginning and at the end of the sequence (see Figure 4.3, left). In those time-intervals the sun is striking straight into the three cameras.

For all the algorithms and cost functions, there are several scattered frames (e.g. around frames 250 and 300) with an extreme low NCC index. They are due to the fact that in those frames the sun struck only the control camera. Thus, there is an analogous effect as with the wiper sequence when there was a wiper only on the control camera. Ignoring those outliers, the shape of the plot increases and decreases depending on whether the sun strikes directly into the three cameras, or not.

With the masked analysis it is more evident that there is a major decay of the performance for BPM-SAD and BPM-SAD at the beginning and end of the sequence where the disparity maps are extremely noisy. In the full analysis, the boosting effect covers the miscalculated disparities, and might lead to a wrong evaluation for a few problematic frames.

We stress the robustness of SGM (in particular for SGM-HIER) when comparing the three algorithms for this sequence. The indices of SGM-HIER, SGM-8

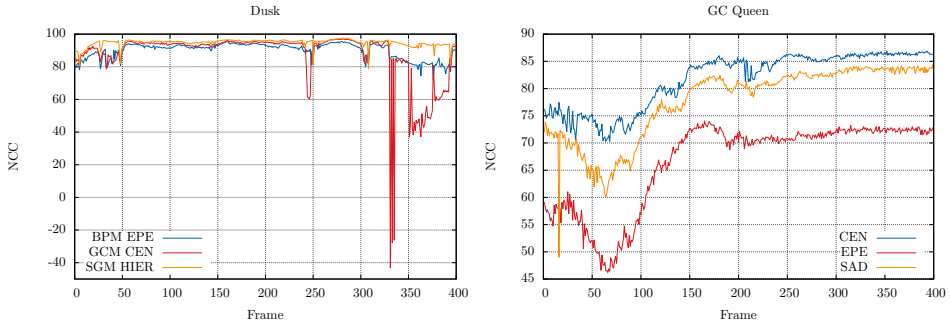

Fig. 9. *Left*: Dusk sequence results comparing best performing configurations of all three matchers. *Right:* Queen sequence results for GCM-configurations.

and SGM-4 are quite similar for most of the frames in the sequence, but SGM-HIER keeps a more "stable" performance in 'complicated intervals' of the sequence. Figure 4.3, left, shows the results for BPM-EPE, GCM-CEN, and SGM-HIER (the best performing configuration of each matcher on this sequence). Note that the low peaks are less intense for SGM. The rank suggested by this plot should be taken carefully, as image domains used for algorithm evaluation are different.

Queen: The results obtained with the full and masked analysis, for all the algorithms, showed a common tendency, with no sudden 'jumps' in the NCC index. There are only a few peaks that represent a miscalculation in a specific frame for a specific configuration.

For example, see Figure 4.3, right, in frame 16, BPM-SAD presented a local minimum for this frame. In this particular case, the algorithm miscalculate the edges of one of the background buildings.

For this sequence, miscalculated disparities are properly identified with full analysis for GCM and BPM. The influence of non-textured regions can be neglected. The sky is covered by surrounding buildings, and the road area is mainly occupied by other vehicles. SGM-HIER and SGM-8 have different rankings when using the full or the masked approaches. It looks like SGM-HIER is taking advantage of the boosting effect from the full analysis for this sequence.

People: Results for all the algorithms show a common pattern for masked and full analysis. Between frames 50 and 100, full and masked analysis report low indices for all the algorithms. This part of the sequence is the most busy one, with many pedestrians present in the scene. The following ups and downs correspond to a single (or two) pedestrian(s) entering or crossing the common field of view. See Figure 10, left, for BPM-results.

As the evaluation approach uses images from three different cameras, and all the pedestrians are fairly close to the recording vehicle, we might conclude that low indices (between frames 50 and 100) are due to occlusions between the cameras. But, because pedestrians are 'fairly slim' structures, even a minor

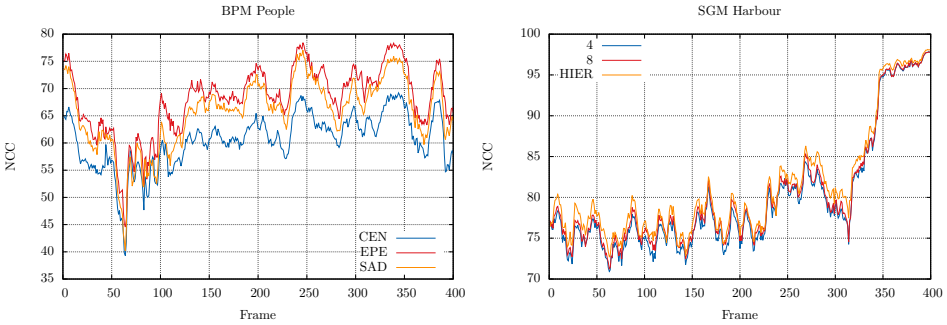

Fig. 10. *Left*: People sequence results for BPM. *Right:* Harbour sequence results for SGM.

miscalculation implies a wrong reconstruction of the whole pedestrian in the virtual image (usually a misplaced body part).

For this sequence, the three SGM configurations show almost indistinguishable performances; it is difficult to detect any difference in NCC mean values. BPM also shows a different behaviour compared to the previous sequences; for this one, BPM-EPE was the configuration with the best performance, and differences between indices of BPM-CEN and the other two configurations were smaller.

The masked analysis shows an almost identical behaviour for BPM and GCM. For SGM, the ranks were totally different for the two types of analysis. It appears that SGM-HIER has more difficulties matching pixels near disparity discontinuities, but performs better on non-textured regions.

Harbour: This sequence presented an interesting difference between the full and the masked approaches for SGM. In all the sequences analysed so far in this chapter, no matter which algorithm, the masked analysis follows the same trend as the full analysis For this sequence, for the three SGM-configurations, full or masked analysis reported a different behaviour in each case. In the full analysis, there is an increasing trend of the index along the sequence; this tendency of the index is particularly strong for the last 100 frames. The masked analysis reported an opposite tendency; the indices decrease along the sequence. As for the full analysis, indices decrease more significantly for the last 100 frames. See Figure 10, right, for the results of the full analysis for SGM.

A possible explanation for this behaviour is that, at the end of the sequence, the metal structure of the bridge disappears from the scene. What is depicted in the images is now mostly sky and road surface, with a large number of skinny poles and small buildings in the background. Increasing indices for the full analysis might be due to the boosting effect on the large non-textured areas in the image. The decreasing tendency of the masked approach could be explained as even the smallest disparity miscalculation would imply a wrong warping of the skinny structures (i.e. the poles and buildings) in the scene. This irregular behaviour needs to be further analysed.

The masked and full approaches for BPM and GCM both show the same trend for all the three cost functions.

Barriers: For this sequence, the algorithms showed common patters, with differences in the magnitude of the index but still a common behaviour. The ups and downs in the indices were dictated by the appearance and disappearance of patches of the sky. The sky region decreases indices and the metal structure of the bridge increases indices; shadows created by the covering structure also contribute to increases. GCM showed an interesting behaviour for this sequence. The road and the barriers are large textured areas whose disparities were better reconstructed in the virtual image generated with BPM-SAD than with the one generated with BPM-CEN. But, the bridge structure was better estimated with BPM-CEN which had the highest NCC values in the evaluation; see Figure 11, left. For this sequence, this behaviour is (already) correctly evaluated with the full analysis. With the masked analysis, it was even more evident.

Night: The matching of corresponding pixels and the evaluation of the matching process (using the trinocular approach) are both challenging due to the limited dynamic range (of about 50 different intensity values only for some of the frames) of the input images from this sequence.

All the algorithms reported very low NCC indices. It is hard to identify visually the 3D structure of the scene in the disparity maps. The only exception was SGM-HIER. In its disparity maps, it is possible to visually identify the road area (illuminated by HAKA1 headlamps) and even some of the objects that surround the road. This SGM-configuration reported the highest evaluation indices for the full analysis. However, it could not be identified as a better performer when using the masked analysis. The boosting effect of the correctly estimated road area seemed to help SGN-HIER in the full analysis.

The results for BPM and GCM show an increasing trend of indices as the sequence progresses. There is more light available in the shown scene in the second half of the sequence (an incoming vehicle with headlamps on is approaching, and the trees around the road are less dense), and more disparity values are cor-

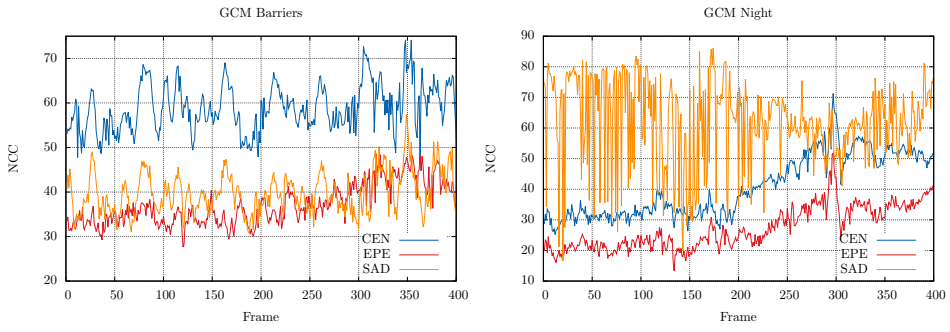

Fig. 11. *Left*: Barriers sequence results for GCM. *Right:* Night sequence result for GCM.

rectly calculated. Figure 11, right, depicts the results for GCM. EPE and SAD cost functions show an increasing trend. High indices reported for some frames by BPM-SAD are due the assignation of a unique (i.e. but incorrect) very low disparity value to most of the pixels in the upper half of the frame. Due to the lack of texture in the image, the trinocular approach fails to assign a low NCC index for the full approach.

4.4 Overall Resume

The BPM algorithm showed an unexpected result when BPM-SAD outperformed BPM-CEN in several sequences (and in some frames even BPM-EPE, the best overall BPM-performer). The census cost function seems to have introduced a kind of some "salt and pepper" noise into the disparity calculations (i.e. non-homogeneous results were homogeneity is expected), which is correctly detected with the trinocular analysis. However, BPM-CEN had a more robust performance, its evaluation index is lower than the one of BPM-SAD, but it showed a more steady behaviour. For some problematic frames (e.g. in the dusk sequence), BPM-SAD generated "useless" data which was not the case for BPM-CEN.

Regarding GCM, the outperforming configuration was GCM-CEN. Even that it generates noisy disparity measurements in non-textured regions (i.e. on the road), it managed to reconstruct better the other structures present in the scenes. The "salt-and-pepper" kind of noise was not introduced when using CEN with GCM. GCM-SAD had an opposite behaviour; the road was uniformly estimated but it introduced "a lot" of incorrect measurements everywhere else; this was better detected when using the masked analysis. GCM-EPE had the poorest performance; the disparity maps have considerable amounts of random values, which degraded significantly the generation of the virtual image.

Among the SGM configurations, SGM-HIER shows the best overall performance with the full analysis. SGM-8 and SGM-4 show very similar evaluation indices; in a direct comparison, even if differences are minor, SGM-8 performed better in a larger number of frames. The most noticeable difference between the three configurations is for the estimation of the road surface. SGM-HIER generates more uniform surfaces; it performed in almost all the sequences better in the full analysis. However, the masked analysis suggests that SGM-HIER (it was just the best in less than half of the sequences) has more difficulties matching regions close to disparity discontinuities, and that its NCC-indices were "helped" by the boosting effect of the full analysis. See Tables 3 and 4 again for direct quantitative comparisons.

5 Conclusions

In this chapter we reported about a testing approach, illustrated by evaluating the performance of three different stereo matching algorithms using long real-world trinocular sequences. The proposed trinocular approach (or, say $n+1$-ocular analysis for an n-camera stereo-vision system) appears to be a fairly

indicative tool to highlight issues or good performance of running stereo-analysis techniques.

We examined the masked and the full approach. We illustrated by example that both approaches may lead to their own evaluation results. Using these two approaches it was possible to point out particular weakness or strength of a matching algorithms in dependency of used configuration. Miscalculations in homogeneous areas may not become 'visible' due to ongoing high NCC-indices in the full analysis; however, using the masked approach, a more appropriate evaluation result is possible in general (the reported index corresponded typically 'well' with a visual analysis of the calculated depth map or the calculated virtual view).

For designing an adaptive computer vision approach for vision-based driver assistance, it appears as particularly important to identify frames (or time intervals) where the behaviour of stereo-analysis algorithms "suddenly changes", such that a new optimization can take place for selecting and configuring a suitable matcher.

References

1. N. Atzpadin, P. Kauff, and O. Schreer. Stereo analysis by hybrid recursive matching for real-time immersive video conferencing. *IEEE Trans. Circuits Systems Video Technology*, **14**:321–334, 2004.
2. H. Badino, U. Franke, and R. Mester. Free space computation using stochastic occupancy grids and dynamic programming. In Proc. *Dynamic Vision, ICCV workshop*, pages 1–12, 2007.
3. S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski. A database and evaluation methodology for optical flow. In Proc. *ICCV*, pages 1–8, 2007.
4. J. Banks and P. Corke. Quantitative evaluation of matching methods and validity measures for stereo vision. *Int. J. Robotics Research*, **20**:512–532, 2001.
5. J. Blanco, F. Moreno, and J. Gonzalez. A collection of outdoor robotic datasets with centimeter-accuracy ground truth. *Autonomous Robots*, **27**:327–351, 2009.
6. R. Bolles, H. Baker, and M. Hannah. The JISCT stereo evaluation. *ARPA Image Understanding Workshop*, pages 263–274, 1993.
7. Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Analysis Machine Intelligence*, **23**:1222–1239, 2001.
8. CMU/VASC. Stereo image data base, `http://vasc.ri.cmu.edu/idb/html/stereo/`, retrieved 2011 [online].
9. Computer Vision Group, University of Bonn. Stereo images with ground truth disparity and occlusion, `http://www.uni-bonn.de/~uzs75l/MRTStereo/stereo_data/index.html`, retrieved 2011 [online].
10. DAGM 2011, adverse vision condition challenge, `http://www.dagm2011.org/adverse-vision-conditions-challenge.html`, retrieved 2011 [online].
11. A. Eid and A. Farag. A unified framework for performance evaluation of 3-D reconstruction techniques. In Proc. *CVPRW*, volume 3, pages 33–41, 2004.
12. G. Egnal, M. Mintz, and R. Wildes. A stereo confidence metric using single view imagery with comparison to five alternative approaches. *Image Vision Computing*, **22**:943–957, 2004.

13. *.enpeda..* (Environment Perception and Driver Assistance) project, The University of Auckland, EISATS (*.enpeda..* Sequence Analysis Test Site), Set 2, `http://www.mi.auckland.ac.nz/EISATS`, retrieved 2011 [online].
14. O. Faugeras, P. Fua, B. Hotz, R. Ma, L. Robert, M. Thonnat, and Z. Zhang. Quantitative and qualitative comparison of some area and feature-based stereo-analysis algorithms. In Proc. *Workshop Robust Computer Vision*, pages 1–26, 1992.
15. P. Felzenszwalb and D. Huttenlocher. Efficient belief propagation for early vision. *Int. J. Computer Vision*, **70**:41–54, 2006.
16. L. Florack, B. Romeny, M. Viergever, and J. Koenderink. The Gaussian scale-space paradigm and the multiscale local jet. *Int. J. Computer Vision*, **18**:61–75, 1996.
17. W. Förstner. 10 pros and cons against performance characterization of vision algorithms. In Proc. *ECCV*, pages 13–29, 1996.
18. U. Franke, C. Rabe, H. Badino, and S. Gehrig. 3D-vision: Fusion of stereo and motion for robust environment perception. In Proc. *DAGM*, pages 216–223, 2005.
19. B. Georgescu and P. Meer. Point matching under large image deformations and illumination changes. *IEEE Trans. Pattern Analysis Machine Intelligence*, **26**:674–688, 2004.
20. R. Gherardi. Confidence-based cost modulation for stereo matching. In Proc. *ICPR*, 978-1-4244-2175-6, 2008.
21. E. Gülch. Results of test on image matching of ISPRS WG III/4. ISPRS J. Photogrammetry Remote Sensing, **46**:1–18, 1991.
22. R. Haeusler and R. Klette. Benchmarking stereo data (not the matching algorithms). In Proc. *DAGM*, pages 383–392, 2010.
23. S. Hermann, S. Morales, and R. Klette. Illumination invariant cost functions in semi-global matching. In Proc. *CVVT*, volume 2, pages 245-254, 2010.
24. S. Hermann and T. Vaudrey. The gradient - a powerful and robust cost function for stereo matching. In Proc. *IVCNZ*, 978-1-4244-9631-0, IEEE, 2010.
25. S. Hermann, S. Morales, and R. Klette. Half-resolution semi-global stereo matching. In Proc.*IEEE Symp. IV*, pages 201–206, 2011.
26. S. Hermann, and R. Klette. Evaluation of a New Coarse-to-Fine Strategy for Fast Semi-Global Stereo Matching. In Proc.*PSIVT*, pages 395–406, 2011.
27. H. Hirschmüller. Accurate and efficient stereo processing by semi-global matching and mutual information. In Proc. *CVPR*, volume 2, pages 807–814, 2005.
28. R. Jiang, R. Klette, and S. Wang. Statistical modeling of long-range drift in visual odometry. In Proc. *CVVT, ACCV workshop*, LNCS 6469, pages 214-224, 2011.
29. JISCT Stereo Images, `http://vasc.ri.cmu.edu/idb/html/jisct/index.html`, retrieved 2011 [online].
30. J. Klappstein, T. Vaudrey, C. Rabe, A. Wedel, and R. Klette. Moving object segmentation using optical flow and depth information. In Proc. *PSIVT*, pages 611–623, 2009.
31. A. Klaus, M. Sormann, and K. Karner. Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. In Proc. *ICPR*, 10.1109/ICPR.2006.1033, 2006.
32. R. Klette, N. Krüger, T. Vaudrey, K. Pauwels, M. Hulle, S. Morales, F. Kandil, R. Haeusler, N. Pugeault, C. Rabe, and M. Leppe. Performance of correspondence algorithms in vision-based driver assistance using an online image sequence database. *IEEE Trans. Vehicular Technology*, **60**:2012–2026, 2011.
33. R. Klette, K. Schlüns, and A. Koschan. *Computer Vision: Three-Dimensional Data from Images*. Springer, Singapore, 1998.

34. R. Klette, T. Vaudrey, J. Wiest, R. Haeusler, R. Jiang, and S. Morales. Current challenges in vision-based driver assistance. In *Progress in Combinatorial Image Analysis* (P. Wiederhold and R. Barneva, eds.), pages 3-22, Research Publ. Services, Singapore, 2010.
35. R. Klette and P. Zamperoni. *Handbook of Image Processing Operators*. John Wiley & Sons, Chichester, 1996.
36. J. Kogler, H. Hemetsberger, B. Alefs, W. Kubinger, and W. Travis. Embedded stereo vision system for intelligent autonomous vehicles. In Proc. *IEEE Symp. IV*, pages 64–69, 2006.
37. V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts? *IEEE Trans. Pattern Analysis Machine Intelligence*, **26**:147 –159, 2004.
38. D. Kondermann, S. Meister, and P. Lauer. An outdoor stereo camera system for the generation of real-world benchmark datasets with ground truth. Universität Heidelberg HCI, Technical Rep., 2011.
39. Y. G. Leclerc, Q.-T. Luong, and P. Fua. Measuring the self-consistency of stereo algorithms. In Proc. *ECCV*, pages 282–298, 2000.
40. Z. Liu and R. Klette. Approximate ground truth for stereo and motion analysis on real-world sequences. In Proc. *PSIVT*, pages 874–885, 2009.
41. R. Mohan, G. Medioni, and R. Nevatia. Stereo error detection, correction and evaluation. *IEEE Trans. Pattern Analysis Machine Intelligence*, **11**:113–120, 1989.
42. S. Morales, T. Vaudrey, and R. Klette. Robustness evaluation of stereo-analysis algorithms on long stereo sequences. In Proc. *IEEE Symp. IV*, pages 347 – 352, 2009.
43. S. Morales and R. Klette. A third eye for performance evaluation in stereo sequence analysis. In Proc. *CAIP*, pages 1078–1086, 2009.
44. S. Morales and R. Klette. Ground truth evaluation of stereo-analysis algorithms for real world applications. In *Proc. CVVT, ACCV workshop*, volume 2, pages 152–162, 2010.
45. P. Mordohai. The self-aware matching measure for stereo. In Proc. *ICCV*, pages 1841 –1848, 2009.
46. J. Mulligan, V. Isler, and K. Daniilidis. Performance evaluation of stereo for tele-presence. In Proc. *ICCV)*, volume 2, pages 558–565, 2001.
47. D. Murray and J. Little. Using real-time stereo vision for mobile robot navigation. *Autonomous Robots*, **8**:161–171, 2000.
48. T. Ndhlovu and F. Nicolls. An alternative confidence measure for local-matching stereo algorithms. In Proc *ROBOMECH*, 9780620447218, 2009.
49. Y. Satoh and K. Sakaue. An omnidirectional stereo vision-based smart wheelchair. *EURASIP J. Image Video Processing*, 11 pages, 10.1155/2007/87646, 2007.
50. D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Computer Vision*, **47**:7–42, 2001.
51. D. Scharstein and R. Szeliski. High-accuracy stereo depth maps using structured light. In Proc. *CVPR*, pages 195–202, 2003.
52. K. Schauwecker, S. Morales, S. Hermann, and R. Klette. A comparative study of stereo-matching algorithms for road-modelling in the presence of windscreen wipers. In Proc. *IEEE Symp. IV*, pages 7–12, 2011.
53. M. Seelinger and J.-D. Yoder. Automatic pallet engagement by a vision guided forklift. In Proc. *ICRA*, pages 4068 – 4073, 2005.
54. Safe Camera System SafetyEYE. `https://shop.pilz.com/eshop/cat/en/DE/00014000337042/SafetyEYE-Safe-camera-system`, retrieved 2011 [online].
55. P. Steingrube, S. Gehrig, and U. Franke. Performance evaluation of stereo-analysis algorithms for automotive applications. In Proc. *ICVS*, pages 285–294, 2009.

56. R. Szeliski. Prediction error as a quality metric for motion and stereo. In Proc. *ICCV*, pages 781–788, 1999.
57. Tech & Industry Analysis from Asia. Toyota's Lexus "LS460" employs stereo cameras for pedestrian detection and collision avoidance steering assist, `http://techon.nikkeibp.co.jp/english/NEWS_EN/20060301/113832/`, retrieved 2011 [online].
58. N. Thacker, A. Clark, J. Barron, J. Beveridge, P. Courtney, W. Crum, V. Ramesh, and C. Clark. Performance characterization in computer vision: A guide to best practices. *Computer Vision Image Understanding*, **109**:305–334, 2008.
59. W. van der Mark and M. Gavrila. Real-time dense stereo for intelligent vehicles. *IEEE Trans. Intelligent Transportation Systems*, **7**:38–50, 2006.
60. T. Vaudrey, C. Rabe, R. Klette, and J. Milburn. Differences between stereo and motion behaviour on synthetic and real-world stereo sequences. In Proc. *IVCNZ*, 10.1109/IVCNZ.2008.4762133, IEEE, 2008.
61. T. Vaudrey, S. Morales, A. Wedel, and R. Klette. Generalized residual images effect on illumination artifact removal for correspondence algorithms. *Pattern Recognition*, **44**:2034–2046, 2011.
62. R. Zabih and J. Woodfill. Non-parametric local transforms for computing visual correspondence. In Proc. *ECCV*, volume 2, pages 151–158, 1994.