Evaluation of a new Coarse-to-Fine Strategy for Fast Semi-Global Stereo Matching

Simon Hermann* and Reinhard Klette

The *.enpeda..* project, Department of Computer Science
The University of Auckland, New Zealand

Abstract. The paper considers semi-global stereo matching in the context of vision-based driver assistance systems. The need for real-time performance in this field requires a design change of the originally proposed method to run on current hardware. This paper proposes such a new design; the novel strategy first generates a disparity map from half-resolution input images. The result is then used as prior to restrict the disparity search space for full-resolution computation. This approach is compared to an SGM strategy as employed currently in a state-of-the-art real-time FPGA solution. Furthermore, trinocular stereo evaluation is performed on ten real-world traffic sequences with a total of 4,000 trinocular frames. An extension to the original evaluation methodology is proposed to resolve ambiguities and to incorporate disparity density in a statistically meaningful way. Evaluation results indicate that the novel SGM method is up to 40% faster when compared to the previous strategy. It returns denser disparity maps, and is also more accurate on evaluated traffic scenes.

Keywords: Semi-global matching, driver assistance systems, coarse-to-fine stereo.

1 Introduction

Stereo correspondence analysis by *semi-global stereo matching* (SGM), as proposed by Heiko Hirschmüller [7], is a popular choice for real-time applications that require dense disparity maps at high frame rates. For example, vision-based *driver assistance systems* (DAS) favour the SGM strategy; see Rabe et al. [11]. A major constraint for real-time SGM implementation is the available memory throughput in current hardware. Because SGM integrates along multiple *1-dimensional* (1D) energy paths, a large memory block needs to be updated in off-chip memory.

Current literature on real-time SGM proposes to alter the design to the original method for ensuring high frame rates for image resolutions equivalent to the VGA norm (i.e. 640×480). For example, Hirschmüller [7] recommends to

* The first author thanks the German Academic Exchange Service (DAAD) for financial support.

integrate at least along eight directions to obtain satisfactory results. But Nedevschi et al. [5] propose to integrate only along horizontal and vertical directions, leaving out diagonal energy paths. They justify their approach with the argument that objects recorded from a moving vehicle are usually aligned along the main axis, such that diagonal directions do not contribute as much to the final solution. But by omitting 50% of the accumulation procedure, the requirements on data processing are eased and real-time performance is achieved.

A research group at Daimler A.G. uses another design concept for their FPGA implementation that was proposed by Gehrig et al. [4]. They keep the recommended eight accumulation paths, but calculate a disparity image on a down-scaled image pair first. The result is then scaled-up to full resolution and serves as a disparity prior. In a consecutive step they calculate a disparity map for a specified region-of-interest with SGM on full resolution images, but using only half of the disparity search space. They generate the final result by replacing disparities in the prior image with disparities from the full resolution map, if the prior suggests that a disparity lies inside the reduced search space. Otherwise the prior disparity is taken as the final result. This is based on the argument that sufficient disparity accuracy for close objects can be obtained when computing half-resolution disparity images only. But, as the re-projection error increases quadratically when disparities get smaller and boundaries of objects further away may become vague due to downscaling, it is required to calculate disparities at full resolution to minimize distance uncertainties for those objects.

The SGM design as proposed in this paper follows the Daimler approach and calculates a disparity prior on half-resolution images. However, in contrast we use the prior to actively determine the search space for the full-resolution SGM, instead of having an indication how to merge independently calculated disparity maps. Our approach therefore follows the standard coarse-to-fine concept, where results from lower-resolution images are used to initialize the same algorithm operating on the next higher resolution level. Such coarse-to-fine approaches are nowadays standard in variational motion estimation algorithms to achieve faster convergence; see, for example, the work by Brox et al. [1] or Zach et al. [18].

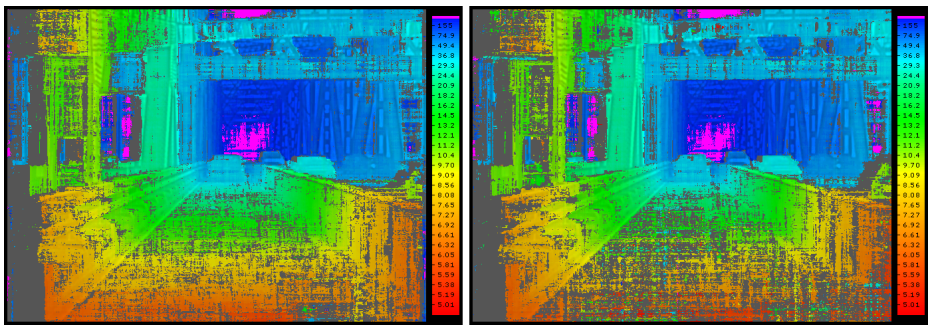

Fig. 1. Disparity results from the new SGM design (left) and the standard SGM design (right). The new design is 60% faster and is much denser especially inside the challenging road area.

To the best of our knowledge, no coarse-to-fine concept as described in the previous paragraph has been proposed so far in combination with SGM, and therefore has not been evaluated either. For the evaluation we propose an extension to an existing methodology [9] that can be used for stereo performance evaluation in the absence of ground truth and employ it on a reasonably large database of real-world traffic sequences. Of course, coarse-to-fine strategies are already employed to improve the performance of stereo matching algorithms in general. For example, a recent publication by Sizintsev and Wildes [14] employs a coarse-to-fine strategy to a block-matching algorithm. Also, in the original SGM design by Hirschmüller [7], a coarse-to-fine strategy is used, but only to support the *mutual information* (MI) cost function. The author recommends to calculate the disparities with SGM at each pyramid level from scratch. So, the prior information is just used to improve the quality of the MI cost function and not to improve the run-time performance of SGM. This defines the place where this paper is positioned, namely somewhere between the original SGM [7] and the SGM design proposed by Gehrig et al. [4]. We use design considerations from the latter work to select a method to be compared with our novel strategy, because of the shared goal to improve the run-time performance of SGM while maintaining stereo accuracy on real-world traffic scenes.

The rest of this paper is organized as follows. In Section 2, relevant details of the SGM algorithm are recalled and parameter settings of the used implementation are given. We present the design consideration as proposed by Gehrig et al., followed by our proposed coarse-to-fine approach. We also provide a discussion about run-time performance. The trinocular evaluation concept as proposed by Morales and Klette [9] is outlined in Section 3; we propose alterations and further extensions to the original method. In Section 4 we present ten real-world sequences, each of 400 trinocular frames, and outline the methodology of our experiments using trinocular evaluation. The results of this study are discussed in detail in Section 5. The paper concludes with a summary in Section 6.

2 Semi-Global Matching

We first recall the SGM algorithm and explain our alterations to the original configuration as reported in [7]. We then compare two SGM design consideration of this reference implementation. The first, called $SGM_{\mathcal{G}}$, is our implementation following the design concept as proposed by Gehrig et al. [4], and this serves as the method of comparison. The second implements our coarse-to-fine approach. We discuss the run-time and disparity analysis performance of both methods.

Cost Accumulation and Cost Function. We introduce the notation for defining the cost accumulation procedure. For a cost accumulation path $L_{\mathbf{a}}$ with direction $\mathbf{a}$, processed between image border and pixel p, we consider the segment $p_0, p_1, \ldots, p_n$ of that path, with p_0 on the image border, and $p_n = p$. The cost at pixel position p for a disparity $d \in \{0, \ldots, D\} \subset \mathbb{N}$ on the path $L_{\mathbf{a}}$ is recursively defined as follows, for $i = 1, 2, \ldots, n$:

$$L_{\mathbf{a}}(p_i, d) = C(p_i, d) + \mathcal{M}_i - \min_{\Delta} L_{\mathbf{a}}(p_{i-1}, \Delta) \tag{1}$$

with

$$\mathcal{M}_i = \min \begin{cases} L_{\mathbf{a}}(p_{i-1}, d) \\ L_{\mathbf{a}}(p_{i-1}, d-1) + c_1 \\ L_{\mathbf{a}}(p_{i-1}, d+1) + c_1 \\ \min_{\Delta} L_{\mathbf{a}}(p_{i-1}, \Delta) + c_2(p_i) \end{cases} \tag{2}$$

where $C(p, d)$ is the similarity cost of pixel p for disparity d, and c_1 and c_2 are the penalties of the smoothness term. The second penalty c_2 is individually adjusted at each pixel p_i to $c_2(p_i)$. The magnitude of the forward difference in direction $\mathbf{a}$ scales the penalty for each p_i with

$$c_2(p_i) = \frac{c_2}{|I(p_{i-1}) - I(p_i)|} \tag{3}$$

where $I(\cdot)$ refers to the intensity at a pixel. For disparities $d = 0$ and $d = D$, the terms $L_{\mathbf{a}}(p_{i-1}, d-1) + c_1$ and $L_{\mathbf{a}}(p_{i-1}, d+1) + c_1$ are removed from $\mathcal{M}_i$, respectively.

The standard SGM algorithm uses eight paths for accumulation (up, down, left, right, and the four in-between angles). To enforce uniqueness, two disparity maps are calculated to perform a left-right consistency check. A disparity passes this test if corresponding disparities do not deviate by more than one disparity level. To identify an occlusion or mismatch, a unique *invalid label* is assigned to pixels whose disparities failed this test. Disparities are calculated with sub-pixel accuracy using the equiangular interpolation method proposed by Shimizu and Okutomi [13]. The penalties are set to $c_1 = 30$ and $c_2 = 150$ for an intensity domain of $[0, 255]$. The input images are smoothed with a small 3×3 mean kernel. As similarity cost, we employ the census cost function which is based on the census transform. Several studies [8, 6] found that this function is very 'descriptive' and robust, even under strong illumination variations, which is crucial for real-world applications.

The census transform [16] assigns to each pixel in the left and right image a signature vector, which is stored as a bit string (i.e. as an integer). This transformation is performed once prior to cost calculation, and signatures are stored in an integer matrix of the dimension of the image. The signature sequence is generated as follows:

$$\text{census}_{\text{sig}} = \Big[\Psi(I_{i,j} \geq I_{i+x,j+y})\Big]_{(x,y)\in\mathcal{N}} \tag{4}$$

where $\Psi(\cdot)$ returns 1 if true, and 0 otherwise. $\mathcal{N}$ denotes a neighbourhood (e.g. 8-neighbourhood) centred at the origin.

The census cost is the Hamming distance of two signature vectors and can be calculated very efficiently [15]. In fact, the cost of calculating the Hamming distance is proportional to the actual Hamming distance and not to the length of the signature string. This is useful in GPU implementations: calculating the cost from scratch is here cheaper than accessing the global memory [3].

Design Considerations. First we introduce some terminology. A standard SGM implementation was described in the previous subsection. We now describe

the design consideration reported by Gehrig et al. [4], denoted by $SGM_{\mathcal{G}}$. Our new approach is denoted by $SGM_{\mathcal{F}}$, where subscript $\mathcal{F}$ stands for "fast".

Both programs, $SGM_{\mathcal{G}}$ and $SGM_{\mathcal{F}}$, calculate a dense disparity map applying standard SGM on half-resolution input images. The images were scaled down using a 5×5 Gauss kernel with $\sigma = 1$. The half-resolution disparity maps are scaled up; in-between pixels are linearly interpolated if both neighbours have a valid disparity assigned to them. When identifying (by the left-right consistency check) a case of occlusion or mismatch, we assign an invalid label to the corresponding 3×3 neighbourhood. This calculated half-resolution disparity map $\mathcal{P}$ serves in both methods as prior for subsequent calculations.

In case of $SGM_{\mathcal{G}}$, a second disparity map F is calculated on full-resolution input images. However, the maximum disparity D is reduced to D/2 to reduce the memory to be processed. The final disparity map R is created as follows

$$R_{i,j} = \begin{cases} \mathcal{P}_{i,j} & \text{if } \mathcal{P}_{i,j} > D/2 - 1 \\ F_{i,j} & \text{otherwise} \end{cases} \tag{5}$$

In case of $SGM_{\mathcal{F}}$, the prior $\mathcal{P}$ is used to define the search space for every individual pixel. For a valid disparity δ in $\mathcal{P}$, we process Equation (1) not for $d \in \{0, \ldots, D\} \subset \mathbb{N}$ but only for $d \in \{\delta - 4, \delta - 3 \ldots, \delta + 3, \delta + 4\} \subset \mathbb{N}$.

In other words we restrict the disparity search space to nine pixels around the prior. In case of disparities close to 0 or D, we do not reduce the search space but shift it accordingly. In case of invalid pixels we simply assign the default search space which would be $d \in \{0, \ldots, D\} \subset \mathbb{N}$, to allow for all possible disparities.

Run-Time Performance. We analyse the approximate run-time performance on images with resolution W×H. We assume that the maximum possible disparity is D. This means that a memory block of W×H×D has to be processed, which resides in off-chip memory. Because one individual integration step consists of a constant number of operations [see Equation (1)], the run-time performance can be related to the size of the memory that needs to be processed. The advantage of this model is its independence from any hardware consideration or implementation.

The memory block used in standard SGM serves as reference to define a coefficient ϱ_X that indicates the ratio of memory needed in SGM_X. Without alterations, we have $\varrho_{\mathcal{S}} = 1$ in standard SGM.

In case of $SGM_{\mathcal{G}}$, we have to process a memory block of size W/2×H/2×D/2 for the half resolution image, and W×H×D/2 for the full resolution image. Adding those two quantities results in $\frac{5}{8}$×W×H×D, which gives a coefficient $\varrho_{\mathcal{G}} = \frac{5}{8}$. We can now measure the performance gain of $SGM_{\mathcal{G}}$ compared to standard SGM, taking into account that

$$1 - \frac{\varrho_{\mathcal{G}}}{\varrho_{\mathcal{S}}} = \frac{3}{8} = 37.5\% \tag{6}$$

In case of $SGM_{\mathcal{F}}$, the individual run-time depends on the density of the half-resolution disparity map, because the whole search space is considered at occlusions in the full-resolution run. We denote the density of this map by φ. The total memory to be processed equals

$$[W/2 \times H/2 \times D/2] + [W/2 \times H/2 \times (9\varphi + (1-\varphi)D)] \stackrel{!}{=} \varrho_{\mathcal{F}} \times W \times H \times D \quad (7)$$

A few algebraic operations lead to

$$\varrho_{\mathcal{F}} = \frac{9}{8} - \varphi \frac{D-9}{D} \quad (8)$$

The gain compared to $SGM_{\mathcal{G}}$ equals

$$1 - \frac{\varrho_{\mathcal{F}}}{\varrho_{\mathcal{G}}} = 1 - \frac{8}{5}\left[\frac{9}{8} - \varphi \frac{D-9}{D}\right] \quad (9)$$

We see that in case of the new design, the run-time performance can actually be worse compared to the standard SGM in cases where the prior disparity map is very sparse. However, in practice this is almost never the case; if it occurs then full-resolution SGM is well justified (i.e. the stereo data is 'challenging'). Consider on the other hand a perfectly dense prior map (i.e. $\varphi = 1$). To obtain the same run time as with $SGM_{\mathcal{G}}$, the minimum disparity range has to be at least $D = 18$. As $\varphi = 1$ is also unlikely, the performance advantage only occurs for larger values of D. For example, a common value such as $D = 128$ defines a possible run-time gain of up to 68%. We measure performance advantages in our experiments by applying Equation (9). Results below show an expected performance gain of about 40%.

3 Trinocular Stereo Evaluation

A predicted-error technique was first employed by Morales and Klette [9] for evaluating stereo analysis on long real-world stereo sequences. It requires at least stereo triples of the same scene, recorded at the same time instance by three calibrated cameras. Two of the three images (i.e reference and match image) are used to calculate a disparity map by the stereo matching algorithm of choice. Each pixel of the reference image is then projected into the position in which it would be located in the third (i.e. control) image C. This *virtual* image V is then compared to the control image C by calculating the *normalized cross-correlation* (NCC) index as follows:

$$NCC(V,C) = \frac{1}{|\Omega|} \sum_{(i,j)\in\Omega} \frac{[V(i,j) - \mu_V][C(i,j) - \mu_C]}{\sigma_V \sigma_C} \quad (10)$$

where μ_V and μ_C denote the means, and σ_V and σ_C the standard deviations of the control and virtual images, respectively. The domain Ω contains only non-occluded pixels (i.e. pixels which are successfully mapped from the reference image to the virtual image domain).

Generating the Virtual Image. In the original work by Morales and Klette [9] it is proposed to use a forward mapping to generate the virtual image. In other words, intensities of the reference image are mapped to positions in the control image and assigned to the closest pixel position. The problem here is that during the mapping process more than one intensity value may be mapped

to the same pixel location. Discarding any of these mappings would cause a bias in the evaluation, as the final index is affected by removing potentially wrong or correct disparities. To avoid this bias we do not calculate a virtual image but rather calculate a control intensity by means of bilinear lookup from the calculated position in the control image that is compared with the intensity of the reference image.

To make this process as easy as possible and to avoid any bias from an otherwise required de-rectification step, we recorded with a horizontally aligned trinocular camera system and rectified the images with respect to the left-most camera. This way we obtain three rectified images where corresponding epipolar lines in all three images are running along the same image row. Thus, a pixel position in the control image has the same y-coordinate as the corresponding pixel in the reference image.

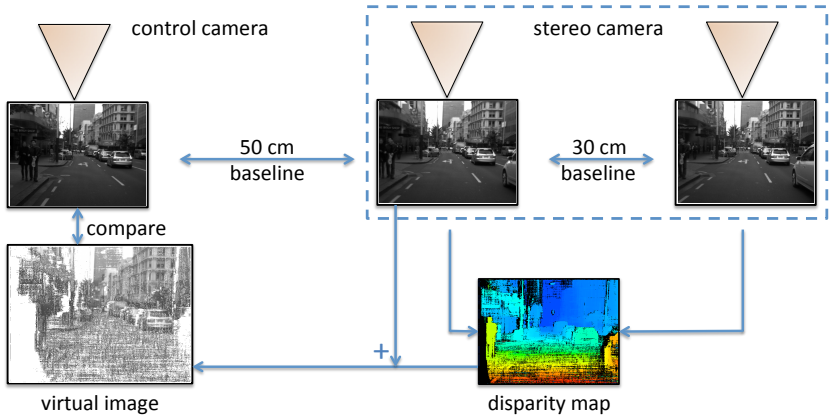

Fig. 2. Setup for the trinocular stereo experiment in this paper, showing one example frame from the experimental database.

The x-coordinate is then calculated as the current location in the reference image plus an offset, which is the product of the current disparity and the ratio of the baselines from the reference image to the control and to the match image. Figure 2 shows the setup in our experiments. The stereo camera has a 30 cm baseline and a disparity map is calculated with the centre image as reference. Then the virtual image is generated and compared with the control image. The scale factor to multiply the disparities with is here $\frac{50}{30}$. Remember that in practice we warp the control image to the image plane of the reference camera as discussed before, but we will stick with the previous terminology as it makes it easier when proposing the following alteration to the original index.

Comparing two Stereo Algorithms. The basic idea of trinocular stereo evaluation is, of course, to have a quality measure to compare the performance of different stereo algorithms in the absence of ground truth. Following the original approach, the difference of the NCC index at each frame for each stereo algorithm is evaluated. In case of only two stereo algorithms, we introduce a

measure ΔNCC that calculates the signed difference of two indices. This makes it easy to compare very similar results as the sign already gives an indication which algorithm performs better.

However, there is a bias in this evaluation. The density of a disparity map is not reflected. Therefore, a sparse stereo algorithm that calculates disparities only at pixels that respond to a robust feature detector would very likely perform much better in this index than a dense algorithm which also assigns disparities in case of weak confidence. The question is how to incorporate the density in the index in a meaningful way. For that, we first introduce some further notation.

We think of images as being random variables X and Y that take intensity values as events. The NCC value can be interpreted as the correlation coefficient $\rho_{X,Y} = Cov(X,Y)/(\sigma_X \sigma_Y)$ with

$$Cov(X,Y) = E[(X - EX)(Y - EY)] \tag{11}$$

So the index reflects a mean of some distribution, and it is possible to calculate the standard deviation of it, referred to by $Cov_\sigma(X,Y)$.

We consider two disparity images D_1 and D_2 that generate two virtual images V_1 and V_2, respectively, both to be compared with a control image C. For the evaluation we consider all pixels of the domain $\Omega_1 \cup \Omega_2$. The total number of this domain is $n = |\Omega_1 \cup \Omega_2|$. We determine for each disparity image the number of invalid pixels as $k_1 = n - |\Omega_1 \setminus (\Omega_1 \cap \Omega_2)|$ and $k_2 = n - |\Omega_2 \setminus (\Omega_1 \cap \Omega_2)|$. We propose for $l = \{1, 2\}$ the following index for the comparison of two stereo algorithms:

$$NCC_\sigma = \frac{1}{n}\left(\left[\sum_{(i,j)\in\Omega_1} \frac{\mathcal{K}}{\sigma_{V_l}\sigma_{C_l}}\right] + \left[\frac{k_1 + k_2}{2}(NCC - Cov_\sigma)\right]\right) \tag{12}$$

where $\mathcal{K} = [V_l(i,j) - \mu_V][C_l(i,j) - \mu_C]$. We omit the arguments (V_i, C) in NCC_σ, NCC, and Cov_σ for better readability.

The index works as follows. Consider $\Omega_1 = \Omega_2$, which results in $k_1 = k_2 = 0$. In this case this index will be identical to the original NCC index as proposed in Equation (10). Now consider the symmetric case that $k_1 > k_2$ and $Cov_\sigma(V_1, C) = Cov_\sigma(V_2, C) = \upsilon$. Again, the index will be identical, because we only add terms that correspond to the pre-calculated mean. However, since we can assume that $\upsilon > 0$, we add terms such that the final index decreases. If the first term is identical for both images, the index that corresponds to the denser disparity map increases. If, on the other hand, $\upsilon_1 > \upsilon_2$, $k_1 = k_2$ and the first term is again identical in both cases, then the index that corresponds to the smaller standard deviation wins. This is reasonable, as we can assume that a smaller standard deviation refers to a 'more consistent' disparity result.

To summarize, with Equation (12) we proposed an alteration to the original evaluation index. It slightly adjusts the original index such that a higher disparity density has a positive impact on the index. We propose to use the standard deviation of the covariance for the index adjustment. This is useful because it relates to the underlying data and therefore gives also an additional quality

measure (see evaluation below). But, the main motivation is that it can annihilate the benefit of a higher disparity density in case that ''additional' disparity values, which do not positively contribute to the index, increase the standard deviation and therefore have a negative affect on the final result. Thus, we regulate the NCC adjustment by two parameters, which can have a compensating or amplifying effect.

4 Evaluation Methodology and Datasets

We evaluate on ten trinocular sequences that show urban and rural environments. Each sequence consists of 400 frames. Figure 3 shows some frame samples. We refer to them by numbers only as we do not discuss them in the context of the scene they are showing, but the sample frames may help to 'read' Table 1.

Fig. 3. From left to right: Example frames of sequence, 3, 5, 9, 10, 6

We evaluated $SGM_{\mathcal{F}}$ and $SGM_{\mathcal{G}}$ on each frame of all sequences using the trinocular evaluation as proposed in Section 3. We calculated the signed difference of several values except the performance where values relating to $SGM_{\mathcal{F}}$ constitute the first summand. This list describes the results provided in Tab. 1:

- ΔNCC: difference of the original index.
- ΔNCC_σ: difference of the adjusted index.

Seq. #	Δ NCC		Δ NCC$_\sigma$		$\Delta\ \sigma$		Δ density		perf. gain	
1	0.73	0.33	2.13	1.29	-16.0	8.30	3.70	1.82	49.2	5.20
2	0.24	0.14	0.79	0.27	3.79	3.80	6.69	1.79	50.5	2.58
3	0.20	0.38	1.00	0.40	4.75	2.11	6.37	0.65	35.8	2.58
4	0.48	0.42	1.62	0.80	6.06	4.19	7.95	0.83	31.4	6.72
5	1.14	0.46	3.19	1.20	-6.42	8.93	4.66	0.62	44.1	4.37
6	0.25	0.40	-0.04	0.39	8.76	4.70	3.19	1.19	42.1	2.94
7	0.32	0.13	1.40	0.73	-3.63	1.97	8.25	1.73	58.9	1.64
8	0.24	0.23	1.54	1.17	-2.17	5.49	6.99	3.25	40.7	8.90
9	0.10	0.22	0.40	0.49	0.37	3.28	4.38	2.51	48.4	7.19
10	0.85	0.21	5.79	1.39	-12.7	8.53	14.8	2.01	45.6	1.60
Mean	0.46	0.27	1.78	0.81	-1.71	5.13	6.70	1.64	44.7	4.37
StdDev	0.34	0.15	1.67	0.42	8.15	2.62	3.34	0.85	7.81	2.54
Median	0.28	0.28	1.47	0.77	-0.9	4.45	6.53	1.76	44.9	3.67

Table 1. Table of evaluation results.

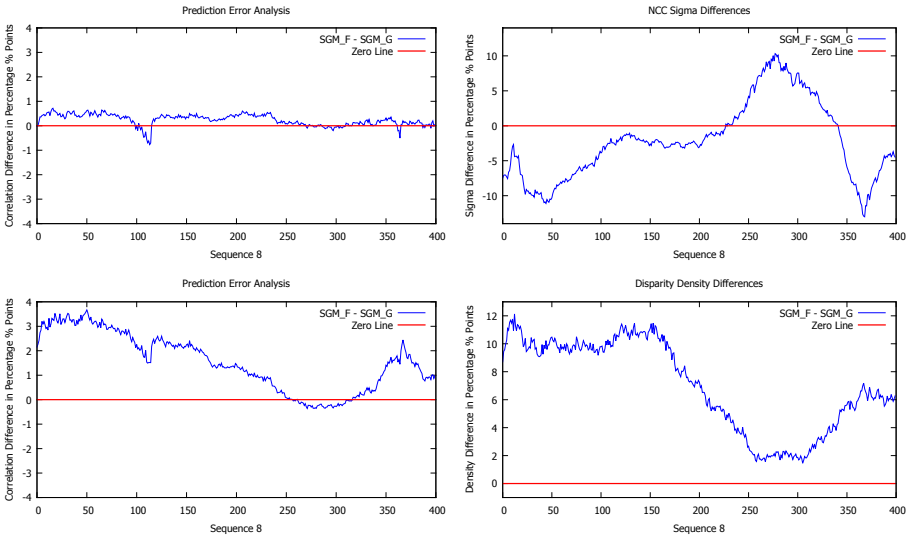

Fig. 4. Results of Sequence 8. Top: ΔNCC and $\Delta\sigma$. Bottom: ΔNCC_σ and Δdensity.

- $\Delta\sigma$: difference of calculated Cov_σ
- Δdensity: difference of the disparity density over the whole image.
- perf. gain: the run-time gain of $SGM_\mathcal{F}$ compared to $SGM_\mathcal{G}$.

The left entry for each item is the mean over the whole image sequence; the right entry is the standard deviation. At the bottom of the table, mean standard deviation and median are given for each item. A positive value favours $SGM_\mathcal{F}$ except for $\Delta\sigma$ where a negative Δ_σ defines 'better'.

Highlighted entries show relatively better performances for $SGM_\mathcal{G}$ (red / sequence 6), and for $SGM_\mathcal{F}$ (green / sequence 10). For a more detailed illustration of one sequence, see frame-by-frame results for Sequence 8 in Fig. 4; values for this sequence are close to medians and thus 'kind of representative'

Disparities of these images increase to up to 84, but we decided to run the algorithms on $D = 128$, for two reasons: First, this disparity limit is the current standard for real-time DAS stereo systems; second, the fact that this way most of the disparity map is taken from the full resolution disparity image in case of $SGM_\mathcal{G}$ is considered beneficial according to Gehrig et al. (page 136,[4]) who state that "Ideally, SGM would be computed everywhere at full resolution".

5 Results

Looking at performance indices ΔNCC and ΔNCC_σ at Tab. 1 a clear tendency in favour for $SGM_\mathcal{F}$ is obvious. All index differences are positive with one exception in Sequence 6. However, since these values refer to percentage point differences, the performance quality of both methods is very similar. But this result comes with a mean run-time improvement of 40% over all sequences for $SGM_\mathcal{F}$. Along with that the new design return 5% to 6% denser disparity maps than $SGM_\mathcal{G}$.

To summarize, our compressed results over 4000 real-world traffic stereo frames suggest that we get slightly denser disparity maps and a positive tendency in stereo performance with a run time improvement of 40% over the method of comparison, which already has a run-time advantage of 37.5% to the standard SGM design. As we already mentioned, we defined the disparity range $D = 128$ but find actual disparities only up to 84. Therefore, the major part of the disparity map generated by $SGM_{\mathcal{G}}$ consists of the full resolution SGM. Thus, qualitative conclusion most likely hold against the standard SGM design. Different configurations and designs will be evaluated in the future, the scope of this paper only allows to introduce the new design and compare it with one design that follows a similar approach and is state-of-the-art.

We can also use the table to check the new evaluation index for consistency. See Sequence 6, where $SGM_{\mathcal{G}}$ performs best w.r.t. the new index. In this sequence we also have a low density advantage for $SGM_{\mathcal{F}}$ which is well below the mean and we have a very high Cov_{σ}. These three values are consistent with our motivation for this index. Also, Sequence 10 where $SGM_{\mathcal{F}}$ performs best has a very low Cov_{σ} and a high disparity density compared to $SGM_{\mathcal{G}}$. This also supports our argument.

For further analysis and to give an example, we picked sequence 8. for frame-by-frame analysis. We choose this sequence as it is close to the median performance (compare with final row of Tab. 1). The graphs for the first four evaluation differences of Tab. 1 can be seen in Fig. 1.

Consider the part from Frame 1-150. The old and the new index follow the same pattern, but the new index pushes $SGM_{\mathcal{F}}$ on a higher index level. Looking at $\Delta\sigma$ and $\Delta density$ we get the explanation. We have a much higher density and lower Cov_{σ} than $SGM_{\mathcal{G}}$. Here both factors work in combination. Between Frames 150 and 250 the original index stays constant and even increases a little. The new index however, slightly decreases. Looking again to the right side of the figure we see that also the density decreases and the Cov_{σ} increases. Again, this effect is visible in the new index. Between Frame 250 and 350 we have a positive impact for the method of comparison. The $\Delta\sigma$ is here in favour for $SGM_{\mathcal{G}}$. There is a slightly higher disparity density for $SGM_{\mathcal{F}}$ that has a small compensation affect. However, this again shows, that the new index works as intended and that results are conform with the expectations. We could not find an example in our results that has a contradicting tendency.

6 Conclusions

We proposed a new design for SGM that employs a coarse-to-fine strategy to reduce computational complexity. We compared this new method to a design that follows a similar approach but with a very different implementation. The common goal of both designs is to reduce the run-time of the algorithm while keeping the quality of results of the original algorithm. We evaluated both designs on 4,000 real-world traffic sequences. For the evaluation we extended an existing trinocular evaluation approach. Our experiments support that the proposed design results in a slightly higher density, has an overall tendency to more

accurate results and also has an average run-time advantage of 40% over the other method. Furthermore, we evaluated a novel evaluation index and found that results are conform with out motivation for defining this index. This new index is of benefit for stereo evaluation when ground truth is missing.

References

1. T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *ECCV*, LNCS 3024, pp. 25–36, 2004.
2. *.enpeda..* image sequence analysis test site. `www.mi.auckland.ac.nz/EISATS`
3. I. Ernst and H. Hirschmüller. Mutual information based semi-global stereo matching on the GPU. In *Int. Symp. Advances Visual Computing*, pages 228–239, 2008.
4. S. K. Gehrig, F. Eberli, and T. Meyer. A real-time low-power stereo vision engine using semi-global matching. In *Computer Vision Systems*, LNCS 5815, pages 134–143, 2009.
5. I. Haller, C. Pantillie, F. Oniga, and S. Nedevschi. Real-time semi-global dense stereo solution with improved sub-pixel accuracy. In *Intelligent Vehicles Symp.*, pages 369–376, 2010.
6. S. Hermann, S. Morales, T. Vaudrey, and R. Klette. Illumination invariant cost functions in semi-global matching. In *Computer Vision Vehicle Technology: Earth to Mars*, ACCV workshop, LNCS, 2011.
7. H. Hirschmüller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *Computer Vision Pattern Recognition*, volume 2, pages 807–814, 2005.
8. H. Hirschmüller and D. Scharstein. Evaluation of stereo matching costs on images with radiometric differences. *IEEE Trans. Pattern Analysis Machine Intelligence*, **31**:1582–1599, 2009.
9. S. Morales and R. Klette. A third eye for performance evaluation in stereo sequence analysis. In *CAIP*, pages 1078–1086, 2009.
10. Y. Ohta and T. Kanade. Stereo by two-level dynamic programming. In Proc. *Int. Joint Conf. Artificial Intelligence*, pages 1120–1126, 1985.
11. C. Rabe, T. Müller, A. Wedel and U. Franke. Dense, Robust, and Accurate Motion Field Estimation from Stereo Image Sequences in Real-Time. In *ECCV*, LNCS 6314, pages 582–595, 2010.
12. T. Saito and J. Toriwaki. New algorithms for n-dimensional Euclidean distance transformation. *Pattern Recognition*, **27**:1551–1565, 1994.
13. M. Shimizu and M. Okutomi. An analysis of subpixel estimation error on area-based image matching. In *Digital Signal Processing*, vol. 2, pages 1239–1242, 2002.
14. M. Sizintsev and R. P. Wildes. Coarse-to-fine stereo vision with accurate 3D boundaries. *Image and Vision Computing*, volume 28:3, pages 352-366, 2010.
15. P. Wegener. A technique for counting ones in a binary computer. *Comm. ACM*, **3**:322, 1960.
16. R. Zabih and J. Woodfill. Non-parametric local transform for computing visual correspondence. In *ECCV*, volume 2, pages 151–158, 1994.
17. T. Vaudrey, C. Rabe, R. Klette, and J. Milburn. Differences between stereo and motion behaviour on synthetic and real-world stereo sequences. In *Image Vision Computing New Zealand*, pages 1–6, 2008.
18. C. Zach, T. Pock, and H. Bischof. A duality based approach for realtime tv-l1 optical flow. In *DAGM*, LNCS 4713, pp. 214–223, 2007.