Half Resolution Semi-Global Stereo Matching

Simon Hermann, Sandino Morales and Reinhard Klette
.enpeda.. group, Department of Computer Science, University of Auckland, New Zealand

***Abstract*— Semi-global matching is a popular choice for applications where dense and robust stereo estimation is required and real-time performance is crucial. It therefore plays an important role in vision-based driver assistance systems. The strength of the algorithm comes from the integration of multiple 1D energy paths which are minimized along eight different directions across the image domain. The contribution of this paper is twofold. First, a thorough evaluation of stereo matching quality is performed when the number of accumulation paths is reduced. Second, an alteration of semi-global matching is proposed that operates on only half of the image domain without loosing any disparity resolution. The evaluation is performed on four real-world driving sequences of 400 frames each, as well as on image pairs where ground truth is available. Results indicate that a reduction of accumulation paths is a very good option to improve the run-time performance without loosing significant quality. Furthermore, operating semi-global matching only on half the image yields almost identical results to the corresponding full path integration. This approach yields the potential to speed up the algorithm by 50% and could also be exploited for other alterations of the algorithm.**

I. Introduction and Related Literature

Stereo estimation by semi-global matching (SGM) [1] is popular for industrial applications where dense and robust real-time stereo matching is required for high frame rates, for example of 30 Hz in current vision-based *driver assistance systems* (DAS). Recently a Daimler AG research group presented their design of an FPGA implementation [2] that reaches this frame rate running on eight accumulation paths. Other authors announced a close to real-time GPU based implementation [3] for images of size 640×480. The original paper by Heiko Hirschmüller [1] recommended the use of 16 (and at least 8) paths for data accumulation to achieve high quality results. The stereo community in general follows this recommendation and uses eight paths.

In [4] a first attempt was made to test the change in stereo quality when costs are accumulated only along four paths. Unfortunately, the evaluation was performed on one data set only and the result is therefore not representative. Another group [5] announced to work with a GPU implementation of SGM that processes at least 24 fps. They reach this frame rate by omitting the diagonal integration paths and therefore omitting 50% of the accumulation procedure. They showed by example and discussion that, according to their experience, the quality decrease on their stereo sequences is only marginal compared to the significant run-time improvement. However, they did not quantify the performance loss for ensuring this run-time gain. To the best knowledge of the authors, no study so far provides a thorough evaluation on traffic sequences that quantifies the performance loss that results when integrating along four instead of eight paths. This study aims to close this gap and tries to quantify the tradeoff in performance that comes with a faster execution of SGM. Furthermore, it is proposed to perform SGM integration on only half of the image resolution to yield faster execution by changing the algorithm design of the accumulation step.

The outline of this paper is as follows. In Section II, relevant details of the SGM algorithm are recalled and parameter settings of the used algorithm are given. Section III presents the different integration strategies which are evaluated. This includes the reduction of accumulation directions from eight to four paths (along the image axes), as well as a possible further reduction down to two paths. The latter strategy is designed such that it does not suffer from the streaking effect known for example from standard dynamic programming [6]. The section concludes by introducing how to accumulate the cost on only half of the image domain without loosing disparity resolution. We simply refer to this as *Half Resolution SGM*. The evaluation methodology is outlined in Section IV. We present four real-world sequences, each of 400 trinocular stereo frames. A quality measure, based on a trinocular stereo evaluation methodology [7], is used to quantify the relative performance of different accumulation strategies. Additionally we run the same algorithm on six different stereo pairs for which ground truth is available. The results of this study are presented and discussed in Section V. They suggest that path reduction is a very good option to gain significant run-time speed-ups. All findings of this study are summarized in Section VI.

II. Semi-Global Matching

This section gives a formal outline of the integration step of the SGM algorithm [1], as we mainly focus on this part of the method.

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) on the path $L_{\mathbf{a}}$ is recursively defined as follows, for $i = 1, 2, \ldots, n$:

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

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

with

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

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 to $c_2(p_i)$. The magnitude of the forward difference in direction $\mathbf{a}$ scales the penalty for each p_i with

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

where $I(\cdot)$ refers to the intensity at a pixel. The reference configuration of the algorithm uses eight paths (up, down, left, right, and the in-between angles) for accumulation. 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. The penalties are set to $c_1 = 30$ and $c_2 = 150$, for an intensity domain of $[0, 255]$. As similarity cost we employ the census cost which is defined in the next section. Several studies [8], [9] found this function to be very descriptive and robust, even under strong illumination variations, which is crucial for real-world applications.

A. Census Cost Function

The census transform [10] assigns to each pixel in the left and right image a signature vector, which is stored as bit string in an integer. This transformation is performed once prior cost calculation and signatures are stored in an integer matrix of the dimension of the image. The signature sequence is generated as follows

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

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

The census cost is the Hamming distance of two signature vectors and can be calculated very efficiently [12]. 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 may become useful in GPU implementations when calculating the cost from scratch could be cheaper than accessing global memory [3]. We employ a 9×3 window as we are working on a 32 bit machine and favour a stronger data contribution along the epipolar line.

III. Integration Strategies

This section recalls the benefits of multiple paths minimization, which is the main reason for the great success of SGM. It also introduces different accumulation strategies which are evaluated according to the methodology as outlined in Section IV.

We use an example frame of a driving sequence that illustrates the following discussion. The left image of Figure 1 shows the result of the reference SGM algorithm when energy minimization is performed only along the epipolar line (here from left to right). This closely corresponds to the standard dynamic programming approach [6]. The right image shows when a second path (right to left) is minimized and results of the cost paths are accumulated. In both images the so called *streaking effect* is apparent, which is a result when cost paths are individually minimized without considering neighbouring image rows. Furthermore, objects seem to be slightly shifted along the accumulation direction when SGM is run with only one path. Introducing a second path that runs in opposite direction resolves this spatial bias. However, it can only slightly resolve the streaking effect. SGM suppresses the streaking effect by introducing non-horizontal cost accumulation paths which contribute to the final costs. Figure 2 shows on the left the improvement when two vertical integration paths are added. The right image of Figure 2 shows the result for 8-path accumulation (adding four diagonal directions). From visual inspection of this example we may conclude that the performance gain from two to four paths is significant, since the streaking effect is resolved by introducing vertical accumulation directions.

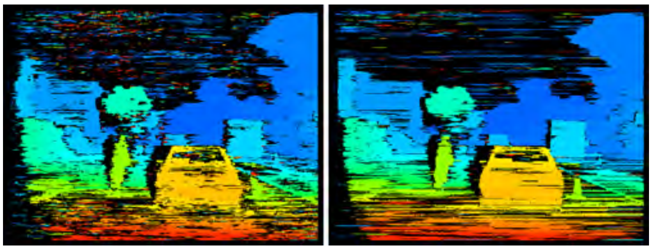

Fig. 1. Result of one horizontal accumulation path (left) and two horizontal paths (right).

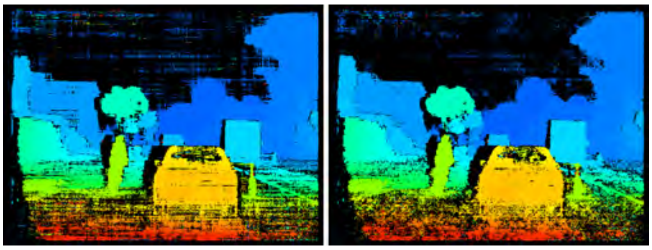

Fig. 2. Result of 4-path integration (left) and 8-path integration (right).

The benefit from four to eight paths, however, appears to be marginal only. This observation coincides with the study [5]. The next section introduces a 2-path accumulation strategy whose design pays special attention to resolve the streaking effect.

A. 2-Path Integration Strategies

The last section highlighted the streaking effect that results from horizontal 2-path integration. Adding non-horizontal paths solves this problem. The solution for removing the streaking comes from the fact that costs are accumulated also along vertical directions, propagating similarity information from neighbouring image rows across the image. So the question arises whether two paths could be sufficient to gain similar results as four paths, by choosing one vertical and one horizontal direction only and therefore also minimizing the energy semi-globally. We can think of two options to apply semi-global 2-path integration strategies. First, we may choose for left and right disparity maps identical paths (e.g. down and right). However, as in case of one direction, this may result in a spatial bias along the chosen directions, such that disparities are slightly propagated across depth discontinuities and objects could be slightly displaced.

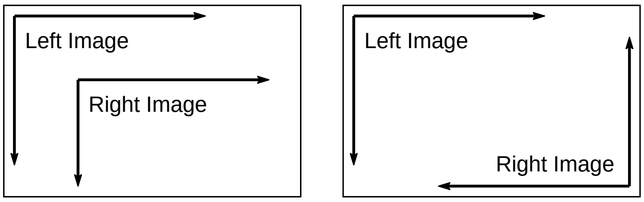

Fig. 3. 2-path integration strategies. Left: Identical setup. Right: Opposite setup.

Another approach to resolve this potential problem is to use opposite path directions for left and right images. The intuition is that potential blur is discarded during the left-right consistency check. Figure 3 sketches both 2-path strategies. Results with those strategies are shown in Figure 4.

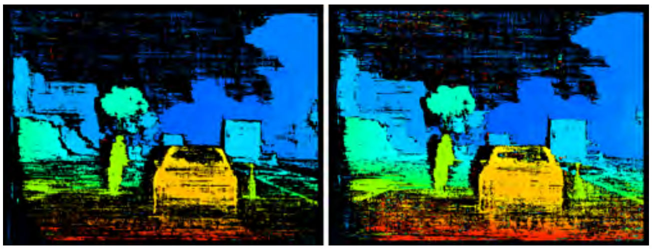

Fig. 4. Result of 2-path integration strategies. Left: Disjoint directions. Right: Identical directions.

By comparing the results we notice that the opposite 2-path strategy lacks a certain amount of denseness. However, we see that this appears only to be true on low textured areas such as the road and along object boundaries. This effect may become of benefit for the application of object segmentation in stereo images.

B. Half Resolution SGM

Another idea to save computational resources is to apply the integration step, for example, only on even pixel indices, i.e. we perform the smoothness operation only on every second pixel of the image. The question is what happens with odd indices. There are two possible solutions. First we simply copy the cost at p_{2i} to p_{2i-1}. Thus all pixels at odd indices get a cost contribution from their left and right neighbouring pixel (as costs are passed back once for each direction). Figure 5 illustrates this strategy. Formally we change the definition of the accumulation path as follows:

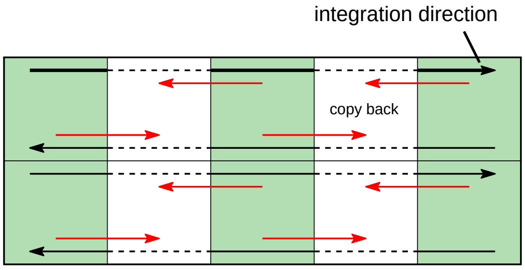

Fig. 5. Example of accumulation step showing two image rows. Calculated costs are passed back to neighbouring pixels.

$$L_{\mathbf{a}}(p_{2i}, d) = C(p_{2i}, d) + \mathcal{M}_{2i} - \min_{\Delta} L_{\mathbf{a}}(p_{i-1}, \Delta) \tag{5}$$

with

$$\mathcal{M}_{2i} = \min \left\{ \begin{array}{c} L_{\mathbf{a}}(p_{2i-2}, d) \\ L_{\mathbf{a}}(p_{2i-2}, d-1) + c_1 \\ L_{\mathbf{a}}(p_{2i-2}, d+1) + c_1 \\ \min_{\Delta} L_{\mathbf{a}}(p_{2i-2}, \Delta) + c_2(p_{2i}) \end{array} \right\} \tag{6}$$

and

$$L_{\mathbf{a}}(p_{2i-1}, d) = L_{\mathbf{a}}(p_{2i}, d) \tag{7}$$

In this study, this approach is applied only on 4-path SGM accumulation using the census cost function. There is no particular reason why half-resolution SGM was left out for 8-paths or even 2-paths, other than the way this approach evolved and was developed. However, we found that the copy operation has a major contribution to the actual runtime, such that the performance gain is significantly less than 50% as one would expect. So far no memory access optimization was done to improve the run-time benefit. However, another option is to omit the copy operation. The disparity map will become more sparse. However, since this sparseness is evenly distributed over the whole image it will ultimately depend on the application whether this stereo map is sufficient for the anticipated purpose. Figure 6 shows the difference between regular 4-path half resolution SGM with (left) and without (right) copy operation. In this evaluation the version with the copy operation is used.

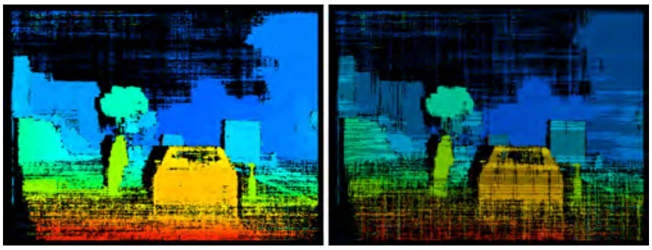

Fig. 6. Left: Half-resolution SGM. Right: Also omitting the copy operation.

IV. Evaluation Methodology and Datasets

The different accumulation strategies as presented in Section III are evaluated on four real-world traffic sequences. Each of these sequences consist of 400 frames and shows typical urban environments. A trinocular stereo evaluation-based approach [7] is employed to quantify the quality of the generated stereo maps. To further support the obtained results of this evaluation we also tested the strategies on stereo data with available ground truth. Results from both methodologies are non-contradicting and support each other.

A. Real-World Data Sets

The data sets used in this evaluation have the names `Queen Street`, `People`, `Harbour Bridge`, `Harbour Barriers` and are available online [13]. They include many challenging scenes with several objects at different distances. Figure 7 shows the left frame of a stereo pair from the `Queen Street` sequence (top) and from the `Harbour Bridge` sequence (bottom). On the right are corresponding disparity maps, generated by the half-resolution SGM approach.

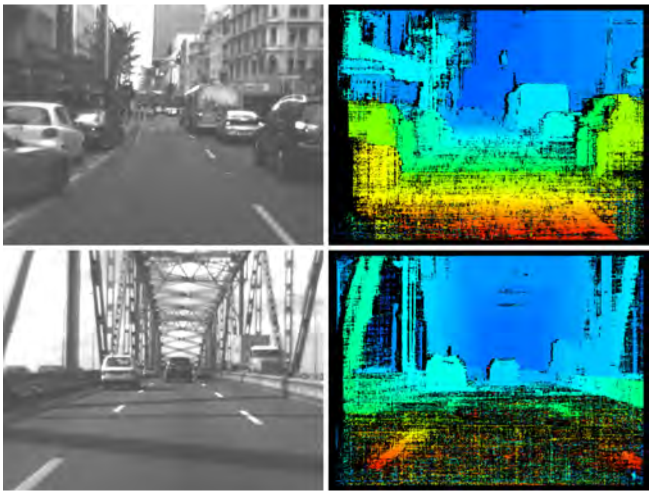

Fig. 7. Top: Frame 75 from the `Queen Street` sequence with stereo map from half-resolution SGM. Bottom: Frame 110 from the `Harbour Bridge` sequence also with generated disparity map.

1) Trinocular Stereo Evaluation: The prediction error technique of [7] for stereo sequences requires at least three different images of the same scene (from different perspectives at the same time instance). The objective is to generate a *virtual* image V from the output of a stereo matching algorithm, and to compare this with an image recorded by an additional *control* camera, that was not used to generate the disparity map. We generate the virtual image by mapping (warping) each pixel of the reference image into the position in which it would be located in the *control image* N (i.e., the image recorded with the control camera). Then, N and V are compared by calculating the *normalized cross-correlation* (NCC) index between them as follows:

$$NCC(N,V) = \frac{1}{|\Omega|} \sum_{(i,j)\in\Omega} \frac{[N(i,j)-\mu_N][V(i,j)-\mu_V]}{\sigma_N \sigma_V} \tag{8}$$

where μ_N and μ_V denote the means, and σ_N and σ_V the standard deviations of the control and virtual images, respectively. The domain Ω is only for non-occluded pixels (i.e., pixels visible in all three images).

However, there is one significant bias in this measure. Inside textureless regions (e.g. the road area), incorrectly calculated disparities will not affect severely enough the final evaluation index if mapped into a wrong position inside the same textureless area. Therefore a mask is generated to reduce the domain Ω by leaving out textureless regions. The mask is produced in two steps. First, a *binarized gradient image* ∇I is calculated. This image is used to generate a Euclidean distance transformation [11] image I_d. We finally get the mask I_m as

$$I_m(i,j) = \begin{cases} 0 & \text{if } I_d(i,j) > T \\ 1 & \text{otherwise} \end{cases} \tag{9}$$

where T is a predefined threshold. The intuition is as follows. Miscalculated disparities at and within a certain distance around pixels with significant gradient (texture), should affect the index more than miscalculated disparities in textureless regions. Textureless regions as defined by the generated mask are discarded from the evaluation.

This approach may not solve the problem entirely but should result in a fairer comparison. However, trinocular stereo evaluation is in general a way to generate quantitative, and meaningful results that correlate well with subjective visual evaluations.

2) Methodology: We focus particularly on

1) The performance difference between 4- and 8-path integration. This is probably the most interesting evaluation because of the significant performance gain that comes with 4-path integration.
2) Performance difference between regular 4-path integration and half-resolution SGM over four paths.
3) To complete the evaluation both presented 2-path strategies are compared with each other and with the regular 4-path integration strategy which serves as reference performance.

B. Synthetic or Engineered Test Data

Synthetic or engineered stereo images provide a way to obtain ground truth, but come with their specifics [14]. Figure 8 shows an example from the data set [15] used in this paper.

1) Ground Truth Evaluation: We calculate the *good pixel percentage* (GPP), defined as follows,

$$GPP = 100\% \times \frac{1}{|\Omega|} \sum_{(i,j)\in\Omega} \begin{cases} 1, & \text{if } |d_{opt} - G_i| \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

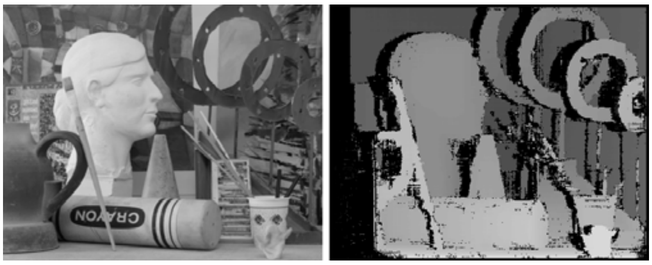

Fig. 8. Left: *Art* image from the ground truth data set. Right: Disparity map using half-resolution SGM.

where G_i encodes the *true disparity* at pixel p_i, and Ω is the set of all pixels with $G_i \neq 0$; value 0 is used to identify occlusions.

2) Methodology: The different accumulation strategies are evaluated on the following six stereo images from this dataset: *Art, Books, Dolls, Laundry, Moebius,* and *Reindeer*. The mean GPP over all six data sets is calculated for each accumulation strategy.

V. Results and Discussion

We mainly focus on the comparison of performance results obtained from the trinocular stereo evaluation approach. Figure 9 shows the results of all four tested sequences, each row corresponding to one evaluation setup as discussed in the previous section. The interesting part of the curves is their general trend in relation to each other, which is why this small scale of the graphs is sufficient for interpretation.

From the charts in Figure 9 (top row) we see that index-curves of 4- and 8-path SGM follow exactly the same pattern, which indicates that both algorithm setups behave almost identical to different complexities and challenges in the stereo data. A frame with a significant index decrease in the 8-path setup results always in a similar decrease in the 4-path setup. There are no exceptions or contradicting tendencies. The only difference is a small decrease of the index level when reducing 8- down to 4-paths. These differences are everywhere less than 5 % points, and most of the time only within 1 or 2 % points. For example, for the `Queen Street` sequence we observe a 2-3 % decrease on an index level of 75 % and higher. This yields a performance decrease of 4 % maximum, but a run-time improvement of about 50% during the expensive accumulation step of the algorithm. Of course, to ultimately quantify the performance loss, an extensive study on real-world sequences with approximate ground truth (e.g. generated by data from a LIDAR system), needs to be performed. However, our results already show that differences in performance, due to different scene complexities, affect both 8- and 4-path in exactly the same way and almost to the same extend.

The performance difference between 4-path SGM and 4-path half-resolution SGM is only marginal. The results are shown in the middle row of Figure 9. The index curves indicate identical, and in one case even superior performance for half-resolution SGM.

The almost identical performance can be explained by used cost function and copy operation. The census cost function considers information of neighbouring pixels which includes the pixels omitted during the accumulation step. Also, since path costs (rather than just data costs) are copied from both neighbouring pixels along the current path, the cost at an omitted pixel also incorporates the SGM smoothness constraint. The downside could be a loss in sub-pixel accuracy at those pixels. Sub-pixel accuracy has not been evaluated in this study.

Another possible option for this approach is to run SGM with two different cost functions that yield different characteristics or advantages. The first run, for example, could employ a cost function on even pixel indices, another run uses a different cost function on odd indices. The benefit of combining two data descriptors in this way instead of merging cost in the data term only, is that the smoothness constraint is already incorporated during the merge step. However, we consider this as an interesting option for future research.

The final evaluation considers the difference in performance between the two proposed 2-path integration strategies, and compares the performance with the SGM 4-path integration as reference. Figure 9 shows the results (bottom row). There is a significant difference of performance when comparing 2-path integration in opposite directions with 2 path integration that yield the same vertical and horizontal directions for left and right input frame. However, performance results are close when comparing the latter 2-path with the regular 4-path integration approach. The 4-path accumulation step however seems overall to slightly outperform the 2-path approach. Although the run-time performance gain of two path accumulation is another 50% compared to the 4-path accumulation, it only yields another 25 % increase when compared to the run time of the 8-path accumulation setup. Nonetheless, the similar performance to 4 path integration is interesting. The weak result of the 2-path strategy using opposite directions is likely due to the lower density of the stereo map.

The results from the ground truth evaluation in Figure 10 confirm the discussed results. There is in fact no difference between 8-path and 4-path SGM. Other than in the trinocular

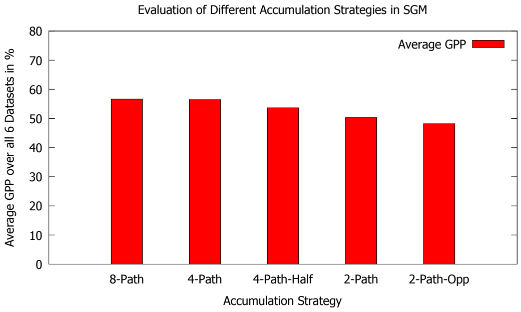

Fig. 10. Left: Art image from the engineered data set. Right: Resulting stereo map using half resolution SGM.

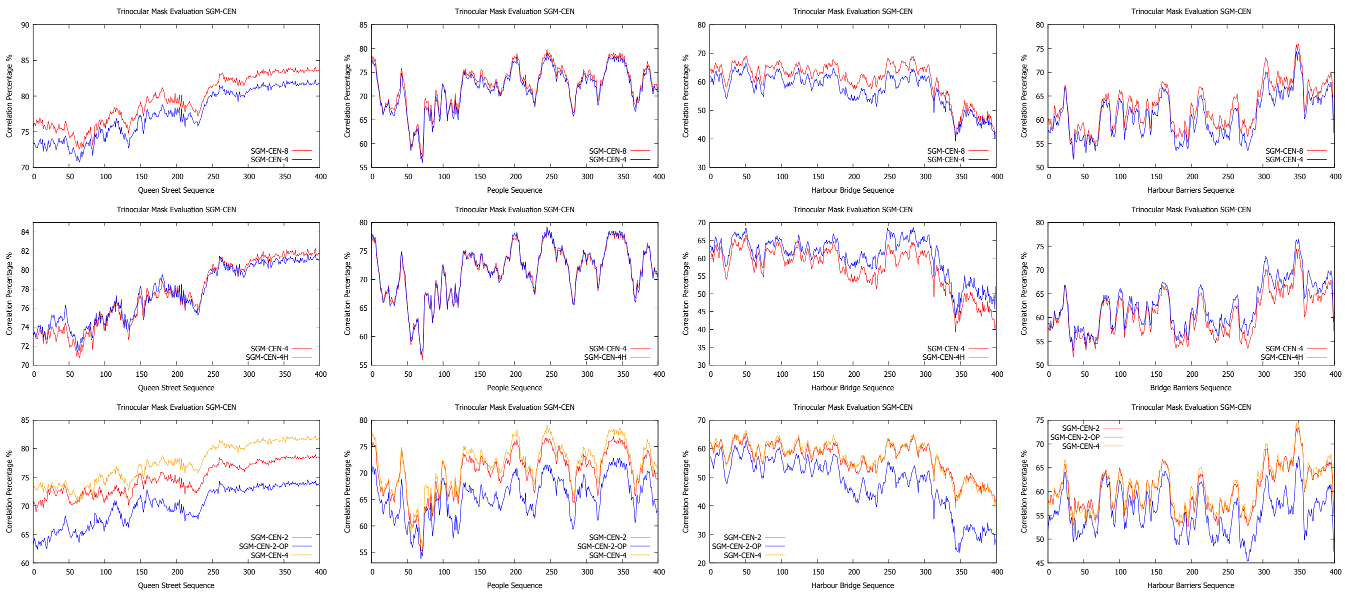

Fig. 9. Correlation percentage curves. Top row: 4- versus 8-path SGM accumulation. Middle row: 4-path standard versus 4-path half-resolution SGM. Bottom row: both 2-path integration strategies.

evaluation we see a slight decrease for half resolution SGM. The performance for the 2-path strategies are again very similar.

Of course, running SGM on real-world traffic sequences with available ground truth ultimately shows how many paths yield which level of performance. However, extensive traffic-scene data sets with available ground truth is unfortunately not available, yet. We used trinocular stereo evaluation for our analysis which gives an indication that reducing the accumulation paths will result in a stereo map with sufficient quality for subsequent applications. Ultimately this is what matters and can give an implicit quality measure:

If an application can operate on a 4-path SGM stereo-map as well as on an 8-path SGM stereo-map, why should one accept a run-time almost twice as long? The argument that processing power increases constantly is not excluding considerations of decreases in run-time in algorithm design; note that usually requirements of applications (e.g., recording speed or image resolution) also advance.

VI. Conclusions

We compared stereo performance of the SGM algorithm when using different accumulation strategies. Results indicate that, depending of the application, an insignificant reduction of quality can be tolerated in return for a significantly faster execution. The study also introduced 2-path integration strategies in SGM that do not suffer from the streaking effect.

Furthermore, running SGM only on half of the resolution of the image (without loosing disparity resolution) was proposed. Performance was almost identical to the regular 4-path SGM. This approach offers new ways for cost integration in a more general sense.

References

[1] H. Hirschmüller. Accurate and efficient stereo processing by semi-global matching and mutual information. *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition(CVPR)*, 2:807–814, June 2005.

[2] S. K. Gehrig, F. Eberli, and T. Meyer. A real-time low-power stereo vision engine using semi-global matching. *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, 5815:134–143, 2009.

[3] I. Ernst and H. Hirschmüller. Mutual information based semi-global stereo matching on the gpu. In *ISVC '08: Proceedings of the 4th International Symposium on Advances in Visual Computing*, pages 228–239, 2008.

[4] S. Hermann, R. Klette and E. Destefanis. Inclusion of a Second-Order Prior into Semi-Global Matching. 3rd Pacific Rim Symposium on Advances in Image and Video Technology. Lecture Notes in Computer Science, vol. 5414, pages 633–644, January 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 Symposium (IV), 2010 IEEE*, pages 369–376, 2010.

[6] Y. Ohta, T. Kanade. Stereo by two-level dynamic programming. *Proc. Int. Joint Conf. Artificial Intelligence* **2** (1985) 1120–1126

[7] S. Morales, T. Vaudrey, and R. Klette. A third eye for performance evaluation in stereo sequence analysis. *Proc. CAIP (2009)* pages 1078–1086

[8] H. Hirschmüller and D. Scharstein. Evaluation of stereo matching costs on images with radiometric differences. *IEEE Trans. Pattern Analysis Machine Intelligence* **31** (2009) 1582–1599

[9] S. Hermann, S. Morales, T. Vaudrey, and R. Klette. Illumination Invariant Cost Functions in Semi-Global Matching. In *Computer Vision in Vehicle Technology: From Earth to Mars (CVVT:E2M)*, 2010.

[10] R. Zabih and J. Woodfill. Non-parametric local transform for computing visual correspondence. *Proc. European Conf. on Computer Vision (ECCV)*, 2:151–158, 1994.

[11] T. Saito and J. Toriwaki. New algorithms for n-dimensional Euclidean distance transformation. *Pattern Recognition*, **27** (1994) 1551–1565

[12] P. Wegener. A Technique for Counting Ones in a Binary Computer. In *Communications of the ACM*, 3:5 page 322, 1960.

[13] *.enpeda..* image sequences analysis test site. `www.mi.auckland.ac.nz/EISATS`

[14] R. Haeusler and R. Klette. Benchmarking stereo data (not the matching algorithms). *Proc. DAGM*, LNCS 6376 (2010) 383–392

[15] Middlebury College, stereo vision page. `vision.middlebury.edu/stereo/`