Inclusion of a Second-Order Prior into Semi-Global Matching

Simon Hermann[1], Reinhard Klette[1], and Eduardo Destefanis[2]

[1] The *.enpeda..* Project, The University of Auckland, New Zealand
[2] Universidad Tecnológica Nacional, Facultad Regional Córdoba, Argentina

Abstract. Today's stereo vision algorithms and computing technology allow real-time 3D data analysis, for example for driver assistance systems. A recently developed Semi-Global Matching (SGM) approach by H. Hirschmüller became a popular choice due to performance and robustness. This paper evaluates different parameter settings for SGM, and its main contribution consists in suggesting to include a second order prior into the smoothness term of the energy function. It also proposes and tests a new cost function for SGM. Furthermore, some preprocessing (edge images) proved to be of great value for improving SGM stereo results on real-world sequences, as previously already shown by S. Guan and R. Klette for belief propagation. There is also a performance gain for engineered stereo data (e.g.) as currently used on the Middlebury stereo website. However, the fact that results are not as impressive as on the *.enpeda..* sequences indicates that optimizing for engineered data does not neccessarily improve real world stereo data analysis.

1 Introduction

Stereo algorithms are currently evaluated either on selected images with calculated ground truth, or on real-world stereo sequences, such as typical for driver assistance systems (DAS). Interestingly, evaluation results differ; for example, algorithms performing well on engineered image examples may fail on real-world sequences [7].

This paper evaluates variants of the SGM algorithm of [5] both on stereo images of the Middlebury stereo website[3] as well as on real-world image sequences of the *.enpeda..* test image website.[4] It discusses various parameter settings and possible preprocessing steps.

1.1 Semi-Global Matching

The SGM algorithm approximates the minimum of a 2D energy function by minimizing multiple 1D energies, employing a dynamic programming scheme. The energy function consists of a data term and two smoothness terms. The first

[3] *vision.middlebury.edu/stereo/*
[4] *www.mi.auckland.ac.nz*, and follow the data link

smoothness term penalizes small disparity changes of neighboring pixels with a rather low penalty c_1 to allow slanted surfaces. The second term penalizes larger disparity changes with a higher penalty c_2. This second penalty is independent of the actual disparity change in order to preserve depth discontinuities. The previously mentioned 1D energies are defined as minimum cost paths $L_{\mathbf{a}}$ that start at each border pixel of the image and are traversed in direction $\mathbf{a}$.

A direction is basically a digitized line, and all digital lines of identical slopes are considered to be equivalent. Usually eight directions are sufficient in SGM to obtain high-quality results. For a digital line in direction $\mathbf{a}$, processed between image border and pixel p, we only consider the segment $p_0 p_1 \dots p_n$ of that digital line, 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 $i = 1, 2, \dots, n$):

$$\begin{aligned} L_{\mathbf{a}}(p_i, d) = \; & C(p_i, d) + \min\big[L_{\mathbf{a}}(p_{i-1}, d), \\ & L_{\mathbf{a}}(p_{i-1}, d-1) + c_1, L_{\mathbf{a}}(p_{i-1}, d+1) + c_1, \\ & \min_{\Delta} L_{\mathbf{a}}(p_{i-1}, \Delta) + c_2\big] - \min_{\Delta} L_{\mathbf{a}}(p_{i-1}, \Delta) \end{aligned}$$

where $C(p, d)$ corresponds to the data term and is the similarity cost of pixel p for disparity d. The costs of paths $L_{\mathbf{a}}$, for all (say, eight) directions $\mathbf{a}$, are accumulated at a pixel p, for all disparities d with $0 \le d \le d_{max}$, and the disparity d_{opt} with the lowest cost is finally selected.

To achieve subpixel accuracy it is proposed to fit a parabolic curve through costs of disparities $d_{opt} - 1$, d_{opt}, and $d_{opt} + 1$, and to take the position of the minimum. Outliers may be filtered by applying a small median filter. For a given stereo pair of images, one image serves as base, and the other one is matched against the base image.

To enforce the uniqueness of a disparity map (for a given stereo pair), roles of base and match images are swapped, which allows to calculate a second disparity image. In a final consistency check, a pixel is labeled valid if the difference of corresponding disparities (in both disparity maps) does not exceed 1; otherwise the pixel is labeled invalid.

[6] identifies invalid disparities either as occlusions or mismatches. For subsequent validation of those, a discontinuity preserving interpolation method is proposed in which valid disparities are propagated into adjacent invalid disparities. This propagation uses, similar to the SGM step, a number of (say, eight) directions, and generates possible values, one for each direction. The original paper suggests to treat mismatches and occlusions differently, by choosing the second lowest value for occlusions (since this value would rather come from the background), and to use the median value as a fair representative for a mismatch. For further details of the algorithm and instructions for implementation, see [5, 6].

1.2 Experimental Setup

We classify potential parameters of an SGM algorithm into primary and secondary parameters.

Primary and Secondary Parameters. Penalties c_1 and c_2 are primary parameters of the cost accumulation step of the algorithm. Hirschmüller suggested to adjust c_2 to the magnitude of the local intensity gradient. As a simple approximation, c_2 is divided by the intensity difference of the current and the previous pixel. If, after such an adjustment, $c_2 \leq c_1$, we set $c_2 = c_1 + 1$.

Any other parameter is considered in this paper to be secondary. The objective now is to derive normative statements about secondary parameters. For that we define a reference configuration of secondary parameters, and evaluate image pairs based on ground truth, for all the possible combinations of c_1 and c_2, with $c_1 = 0, ..., 50$ and c_1 incremented in steps of 5, and $c_2 = 0, ..., 250$ and c_2 incremented in steps of 25.

We then change only one secondary parameter, evaluate for all combinations of c_1 and c_2, and compare the results with the reference configuration. For our reference configuration we implemented the algorithm as described in the previous section but without subpixel accuracy. Also for simplicity reasons we treated occlusions and mismatches equally by simply choosing the lowest valid value of propagated disparities.

Costs are computed using Birchfield and Tomasi's similarity measure [2]. A 3×3 median filter is used for eliminating outliers, and the described consistency check ensures the uniqueness of the solution. No smoothing of the input images is done prior to this processing, and parameter c_2 is adjusted by intensity differences.

For our experiments we decided for the Tsukuba sequence from the Middlebury stereo website, taking image *scene1.row3.col2.ppm* to be the left and *scene1.row3.col3.ppm* to be the right input image. The disparity range was chosen to be limited by $d_{max} = 18$.

We evaluate the error at all pixels, and consider a disparity to be false if it differs from the ground truth. Results (i.e., percentage of bad pixels) are shown in Table 1. The 'Mean 1/4' error value is calculated by taking the mean of the best 25% of the error results, and the 'Mean 1/2' by taking the mean of the best 50%. Numbers in brackets (after median, minimum and maximum values) specify the corresponding (c_1, c_2) configuration. We now describe changes of parameters and present obtained results.

Table 1. Errors in % for the reference configuration of secondary parameters.

	Mean 1/4	Mean 1/2	Mean	Median	Min	Max
Reference parameter	**13.1**	**13.5**	**19.3**	**14.3 / (40,125)**	**12.8 / (20,125)**	**90.1 / (0,0)**

Use of Smoothing or Median Filters. We applied either a 3×3 or a 5×5 smoothing filter on the input images prior to processing them with the SGM algorithm:

$$\frac{1}{16} \cdot \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix} \qquad \text{or} \qquad \frac{1}{100} \cdot \begin{bmatrix} 1 & 2 & 4 & 2 & 1 \\ 2 & 4 & 8 & 4 & 2 \\ 4 & 8 & 16 & 8 & 4 \\ 2 & 4 & 8 & 4 & 2 \\ 1 & 2 & 4 & 2 & 1 \end{bmatrix}$$

Minimum error values are printed in bold in Table 2. This experiment indicates that using a small 3×3 smoothing kernel generally improves the results of SGM, independent of the setting of c_1 and c_2. A larger kernel seems to have a negative influence on results.

Table 2. Results for different smoothing filters.

	Mean 1/4	Mean 1/2	Mean	Median	Min	Max
Reference parameter	13.1	13.5	19.3	14.3 / (40,125)	12.8 / (20,125)	90.1 / (0,0)
Smooth 3x3	**11.4**	**11.9**	**17.7**	**12.9 / (30,125)**	**10.4 / (10,0)**	**86.0 / (0,0)**
Smooth 5x5	13.5	14.2	19.8	15.2 / (35,150)	12.4 / (10,0)	86.3 / (0,0)

Now, a 3×3 median filter is used as part of the reference configuration. We extend the window size of the median filter to 5×5 and 7×7 while leaving the rest of the reference configuration unchanged; see Table 3.

Table 3. Results for different median filters.

	Mean 1/4	Mean 1/2	Mean	Median	Min	Max
Reference parameter	13.1	**13.5**	19.3	14.3 / (40,125)	12.8 / (20,125)	90.1 / (0,0)
Median 5x5	**12.9**	**13.5**	18.8	**14.3 / (40,125)**	**12.6 / (20,125)**	88.6 / (0,0)
Median 7x7	13.1	13.7	**18.7**	14.5 / (40,100)	12.7 / (20,125)	**88.4 / (0,0)**

Best results are typically obtained when using the 5×5 median. In cases of the overall mean and the maximum value, smaller error values are obtained for the 7×7 median. In general it seems that a 5×5 median performs better than a 3×3 median, for any configuration (c_1, c_2). However, the improvement seems to be minor.

Use of Different Numbers of Paths. Hirschmüller suggested in his paper [5] that "the number of paths must be at least 8 and should be 16 for providing a good coverage"; results in Table 4 confirm his statement.

Table 4. Results for different median filters.

	Mean 1/4	Mean 1/2	Mean	Median	Min	Max
Reference parameter	13.1	**13.5**	19.3	**14.3 / (40,125)**	12.8 / (20,125)	**90.1 / (0,0)**
Path 4	14.3	14.5	21.3	14.9 / (50,75)	14.0 / (30,25)	**90.1 / (0,0)**
Path 16	**13.0**	**13.5**	**19.0**	14.5 / (40,100)	**12.7 / (20,175)**	**90.1 / (0,0)**

Eight paths lead to better results than four paths. Improvements are about 1% by comparison. Also, choosing 16 paths results in lower errors. However,

improvements in this experiment are around 0.1 %. In practical applications like DAS, where real time performance is crucial, such a marginal quality gain would not justify any increase in computational time.

2 Use of Second Order Prior and New Cost Function

We suggest a possible improvement of SGM results by adding an additional penalty during the cost accumulation process, based on a second order prior. The idea is that a configuration of disparities should be favored for which the second order derivative at p_i is small. This should equalize the high penalty c_2 which is added regardless of the discontinuity.

2.1 New Smoothness Term

Consider three consecutive pixel positions along a path $L_{\mathbf{a}}$, say p_{i-1}, p_i, and p_{i+1}, with disparities d_{i-1}, d_i and d_{i+1}, respectively. This defines a triangle in 3D space, with disparities being the third coordinate. The angle α at (p_i, d_i) can easily be computed using the formula

$$\alpha = \arccos\left(\frac{a^2 + b^2 - c^2}{2ab}\right)$$

(see Figure 1) with

$$\begin{aligned} a &= ||(p_{i-1}, d_{i-1}), (p_i, d_i)||_2 \\ b &= ||(p_i, d_i), (p_{i+1}, d_{i+1})||_2 \\ c &= ||(p_{i-1}, d_{i-1}), (p_{i+1}, d_{i+1})||_2 \end{aligned}$$

$||\cdot||_2$ is the Euclidean distance. The goal is to favor smooth transitions (i.e., we need a function that increases the penalty when the angle gets smaller, and

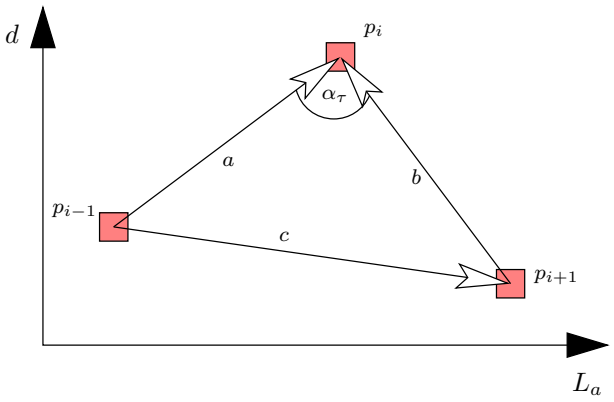

Fig. 1. Estimation of second order prior.

decreases when the angle gets larger). Since the maximum possible angle is π, we choose

$$c_3 = (\frac{\pi}{\alpha} - 1.0) \cdot \tau$$

as a function of α and of an external scalar τ. Positions p_{i-1} and p_{i+1} are determined by pixel position p_i and direction $\mathbf{a}$. Thus, c_3 is basically a function of disparities d_{i-1}, d_i, and d_{i+1} (and of τ). We now need to compute c_3 at every p_i, for every d during the accumulation. Thus, when computing the penalty we already know the disparity at p_i. We have to select an a-priori disparity d_{mx} with the most likely minimum cost at p_{i+1} (i.e., most likely to be selected as d_{opt}). We select

$$d_{mx} = \min_{\Delta} C(p_{i+1}, \Delta)$$

to be a 'good guess'. Now we may write c_3 as a function of the disparity only, chosen for the previous position (p_{i-1}) (i.e., $c_3(d_{prev})$). Define

$$d_{mp} = \min_{\Delta} L_{\mathbf{a}}(p_{i-1}, \Delta)$$

which is the disparity with the current minimum cost on the path at the previous position. Also define the cost at pixel p for disparity d on the path $L_{\mathbf{a}}$ as follows:

$$\begin{aligned} L_{\mathbf{a}}(p,d) = \; & C(p,d) + \; \min\big[L_{\mathbf{a}}(p_{i-1},d), L_{\mathbf{a}}(p_{i-1},d-1) + c_1 + c_3(d-1), \\ & L_{\mathbf{a}}(p_{i-1},d+1) + c_1 + c_3(d+1), \min_{\Delta} L_{\mathbf{a}}(p_{i-1},\Delta) + c_2 + c_3(d_{mp})\big] \\ & - \min_{\Delta} L_{\mathbf{a}}(p_{i-1},\Delta) \end{aligned}$$

For results of this approximation, see Table 5. With the exception of the overall mean, the errors tend to be slightly reduced when using a second order prior. The constant τ was set to be $\frac{3}{2}$. However, this is just an initial experience with including a second order prior. More experiments and modified approaches (say, with other parameter settings for τ or function c_3) should be performed in future; this may just define a new direction of research.

Table 5. Results for 2nd Order Prior

	Mean 1/4	Mean 1/2	Mean	Median	Min	Max
Reference parameter	13.1	13.5	**19.3**	14.3 / (40,125)	12.8 / (20,125)	90.1 / (0,0)
2nd Order Prior	**12.8**	**13.2**	19.5	**14.0 / (40,25)**	**12.2 / (20,200)**	**76.6 / (0,0)**

2.2 New Cost Function based on Signal Deviation

The reference configuration of the SGM algorithm uses the BT cost function [2]. This function computes the cost at pixel p_i as follows: Let I_{p_i} be the intensity

value of pixel p_i in the base image and I_{q_i} the intensity for the corresponding pixel in the match image, for disparity d. Intensities in both images are interpolated using intensities of previous or subsequent pixels along the epipolar line. For example, let $I_{p_{i-1/2}} = \frac{1}{2} \cdot I_{p_i} + \frac{1}{2} \cdot I_{p_{i-1}}$ be an interpolated value at p_i, just using the previous pixel. The absolute difference of $\min(I_{p_{i-1/2}}, I_{p_i}, I_{p_{i+1/2}})$ and $\min(I_{q_{i-1/2}}, I_{q_i}, I_{q_{1+1/2}})$ is then used for the final matching cost.

This new scheme for cost calculations considers a 1D window around pixels p_i and q_i. Usually, this window should have a size of $\omega = 5$ or $\omega = 7$. We take the mean of the sum of absolute intensity differences,

$$\frac{1}{\omega} \cdot \sum_{j=i-\frac{\omega}{2}}^{i+\frac{\omega}{2}} \delta_j$$

with three options for δ_j. This value is one of the following:

(1) $\delta_j = \left| I_{p_j} - I_{q_j} + (I_{q_i} - I_{p_i}) \right|$
(2) $\delta_j = \left| I_{p_j} - I_{q_j} \right|$
(3) $\delta_j = \left| I_{p_j} - I_{q_j} \right| - |I_{p_i} - I_{q_i}|$

The first two options can be interpreted as a mean deviation from the intensity signal of the match image compared to the signal of the base image. Thus, this similarity is not only (as in BT) based on intensity differences at pixel locations, but also on the 'structure' of the signal. See Table 6.

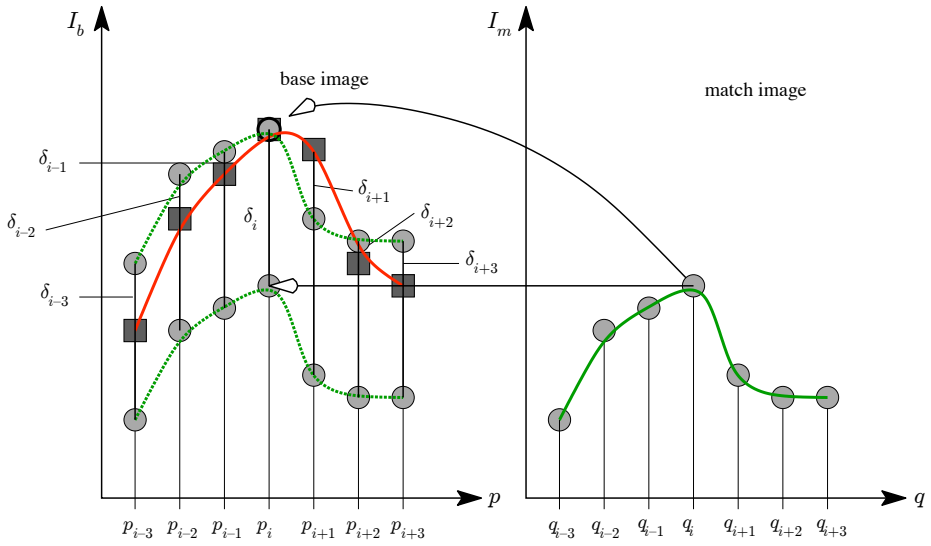

Fig. 2. The cost function and the structure of the signal: the intensity value of q_i is shifted such that we have $I_{p_i} - I_{q_i} = 0$. The cost at p_i is the mean of all absolute differences within the selected neighborhood.

Table 6. Results for different window sizes for new cost function.

	Mean 1/4	Mean 1/2	Mean	Median	Min	Max
Reference parameter	13.1	13.5	19.3	14.3 / (40,125)	12.8 / (20,125)	90.1 / (0,0)
Cost opt.1 w=5	13.1	13.3	16.5	13.8 / (50,200)	12.9 / (35,25)	52.1 / (0,0)
Cost opt.1 w=7	14.1	14.3	18.0	15.0 / (25,175)	14.0 / (45,50)	46.4 / (0,0)
Cost opt.2 w=5	12.4	12.6	16.3	13.0 / (25,225)	12.1 / (35,50)	44.8 / (0,0)
Cost opt.2 w=7	12.7	12.9	16.2	13.3 / (25,100)	12.6 / (30,25)	**41.8 / (0,0)**
Cost opt.3 w=5	**11.3**	**11.7**	**14.6**	**12.4 / (35,125)**	**10.9 / (35,50)**	52.6 / (0,0)
Cost opt.3 w=7	**11.3**	**11.7**	**14.6**	12.7 / (10,50)	11.0 / (20,75)	47.2 / (0,0)

For option (1), by shifting the intensities by offset $(I_{q_i} - I_{p_i})$, the difference of intensities at $j = i$ becomes zero. See, for example Figure 2. The intensity signal around q_i is shifted, and differences are taken at new positions. This option emphasizes almost completely the structure of the signal, and not so much intensity differences. This might be of value if changes in lighting occur between both images of a stereo pair. However, results are similar to the reference configuration if input images do not show such changes in lighting.

Option (2) leaves intensity values unshifted, and simply computes the mean of the sum of differences. This option emphasizes intensity differences as well as the structure of the signal. See the lower signal of the match image in Figure 2. Results are about 1% better than for the reference configuration.

Option (3) improves results by about 2% compared to the reference configuration, which is certainly very good! The difference to option (i) is that we subtract the absolute value of the offset. A geometric interpretation of (iii) is still missing.

2.3 Best Configuration

Finally we choose a best configuration by picking from every analyzed secondary parameter the, to our opinion, best option (i.e., we choose eight paths for the accumulation, also considering the computational cost, use a 5×5 median filter for outliers, a 3×3 smoothing kernel, and option (2) for the cost function because we have a geometrical motivation and improvement).

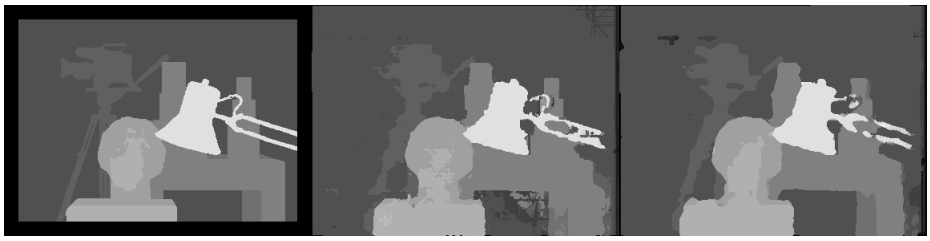

Fig. 3. *Left: ground truth of Tsukuba. Middle: result of SGM using the reference configuration. Right: result of SGM using the best configuration.*

Table 7. Results for the best configuration

	Mean 1/4	Mean 1/2	Mean	Median	Min	Max
Reference parameter	13.1	13.5	19.3	14.3 / (40,125)	12.8 / (20,125)	90.1 / (0,0)
Best parameter	**9.9**	**10.2**	**12.6**	**10.8 / (40,50)**	**9.1 / (15,0)**	**32.5 / (0,0)**

The second order prior is included into the cost accumulation step. Results outperform, as expected, any result obtained for modifying just a single parameter; see Table 7.

Figure 3 shows the Tsukuba ground truth on the left. The image in the middle shows the obtained result when using the reference parametrization, and the image on the right the resulting disparity map for our identified 'optimum configuration'. Obviously, there are some major improvements.

3 Application to *.enpeda..* Sequences

We also applied the discussed versions of the SGM algorithm to the sequences of Set 1 of the *.enpeda..* test image website. Our experiments confirmed that Sobel preprocessing for those sequences is beneficial, as already shown for belief propagation [4]; see Table 8 for edge results on Tsukuba image sequence.

Table 8. Results for edge preprocessing on Tsukuba images

	Mean 1/4	Mean 1/2	Mean	Median	Min	Max
Reference parameter	13.1	13.5	19.3	14.3 / (40,125)	12.8 / (20,125)	**90.1 / (0,0)**
Sobel Preprocessing	**12.5**	**12.8**	**18.5**	**13.3 / (35,50)**	**12.2 / (20,75)**	94.7 / (0,0)

Figure 4 illustrates results for frame 106 of the *construction site* sequence (not using the original depth of 12 bits but scaled to 8 bits).

The image in the upper row, left, shows the right input image of the stereo pair, and in upper row, right, its Sobel edge image. The depth maps in this figure have value $200 - d_{opt} \cdot 5$ if d_{opt} is calculated at that pixel, with $d_{max} = 40$.

The images in the middle row shows the result of applying SGM to the original image data using the reference configuration with $c_1 = 20$ and $c_2 = 125$, which was the suggested primary parameter setting (left: original input, right: Sobel images as input).

Resulting depth maps appear to be, obviously, more accurate in general with Sobel preprocessing. (Studies for approximated ground truth are a subject for future work.)

The images in the bottom row are results for our 'optimum configuration' as described above (also with $c_1 = 20$ and $c_2 = 125$), again either on the original data (left) or on Sobel image pairs (right).

In our experiments, we processed the sequences of Set 1 (Daimler sequences) of the *.enpeda..* test image website, using throughout our 'optimum configuration' on the Sobel input data. Figure 5 illustrates examples; each row

Fig. 4. *Top: image of original input sequence (left) and its Sobel image (right). Middle: results of SGM (reference configuration) on original image pair (left) and on Sobel image pair (right). Bottom: results of SGM (using our optimized configuration f) on original image pair (left) and on Sobel image pair (right).*

has an original image on the left and our optimized SGM result on the right. From top to bottom, the rows are showing the *intern on bike*, *save turn*, *dancing light*, and *squirrel* sequences, in this order.

The *squirrel sequence* was taken at night which possibly contributes to the difficulty here. The daylight sequences seem to perform reasonably well. Improvements from the reference to the optimized configuration are obvious especially on Sobel preprocessed images.

Fig. 5. *Left: example of a right input image of the processed sequence. Right; depth maps, after Sobel preprocessing, and using SGM with the optimized configuration.*

4 Conclusions

This paper proposes a new cost function and tested it with the SGM algorithm. It also contributes by presenting a first attempt to include an additional penalty

to the accumulation step, based on a second order prior. Results indicate that there is a potential for performance gain and justifies more experiments for this subject in future.

We also tested SGM on Sobel images of the Tsukuba image sequence on the Middlebury stereo page. Results indicate that edge preoprocessing can improve the quality of the algorithm (see Table 8). Especially the outcome of our experiments on real-world sequences suggest that processing SGM on edge images can also result in a big performance gain.

Obviously, the discussed options of variations in primary and secondary SGM parameters allow for many more optimization experiments, also with respect to possible preprocessing. However, [7] indicates that the Sobel operator compares well against other edge operators (Canny, Kovesi-Owens) in case when using belief propagation for disparity calculation. However, performance gains are much better on real world sequences than on engineered data. Therefore it would be interesting to quantify how much real world stereo analysis really benefit from optimizating for engineered data.

5 Acknowledgement

The authors would like to thank Thomas Pock for the idea to include a second order prior into the cost accumulation step of the algorithm.

References

1. H. Badino. A robust approach for ego-motion estimation using a mobile stereo platform. In Proc. *Int. Workshop Complex Motion*, LNCS 3417, pages 198–208, Springer, Berlin, 2004.
2. S. Birchfield and C. Tomasi. Depth discontinuities by pixel-to-pixel stereo. *Int. J. Computer Vision*, **35**:269–293, 1999.
3. S. Gehrig and U. Franke. Improving stereo sub-pixel accuracy for long range stereo. Daimler A.G., Internal Report, Sindelfingen, 2007.
4. S. Guan and R. Klette. Belief-propagation on edge images for stereo analysis of image sequences. In Proc. *Robot Vision*, LNCS 4931, pages 291–302, Springer, Berlin, 2008.
5. H. Hirschmüller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *IEEE Conf. Computer Vision Pattern Recognition*, volume 2, pages 807–814, 2005.
6. H. Hirschmüller. Stereo vision in structured environments by consistent semi-global matching. In *IEEE Conf. Computer Vision Pattern Recognition*, volume 2, pages 2386–2393, 2006.
7. R. Klette. Evaluation of stereo and motion techniques on real-world video sequences. Dagstuhl seminar *Statistical and Geometrical Approaches to Visual Motion Analysis*, http://kathrin.dagstuhl.de/08291/Materials2/, 2008