

EVALUATION OF PENALTY FUNCTIONS FOR SEMI-GLOBAL MATCHING COST AGGREGATION

Christian Banz, Peter Pirsch, and Holger Blume

Institute of Microelectronic Systems
Leibniz Universität Hannover, Hannover, Germany
{banz,pirsch,blume}@ims.uni-hannover.de

KEY WORDS: Stereoscopic, Quality, Matching, Vision, Reconstruction, Camera, Disparity Estimation, Semi-Global Matching

ABSTRACT:

The stereo matching method semi-global matching (SGM) relies on consistency constraints during the cost aggregation which are enforced by so-called penalty terms. This paper proposes new and evaluates four penalty functions for SGM. Due to mutual dependencies, two types of matching cost calculation, census and rank transform, are considered. Performance is measured using original and degenerated images exhibiting radiometric changes and noise from the Middlebury benchmark. The two best performing penalty functions are inversely proportional and negatively linear to the intensity gradient and perform equally with 6.05 % and 5.91 % average error, respectively. The experiments also show that adaptive penalty terms are mandatory when dealing with difficult imaging conditions. Consequently, for highest algorithmic performance in real-world systems, selection of a suitable penalty function and thorough parametrization with respect to the expected image quality is essential.

1 INTRODUCTION

Calculating depth information by stereo matching (disparity estimation) is a common image processing task in many remote sensing applications. Typical applications of range cameras based on stereo imaging include advanced driver assistance systems, robotics, and keyhole surgery assistance systems. Crucial aspects for real-world suitability is accuracy and density of the depth map, which are especially difficult to achieve at in untextured areas. These requirements are further impacted by noise and difficult lighting conditions. Naturally, all of these effects occur in real-world scenarios.

The semi-global matching algorithm (SGM) (Hirschmüller, 2008) is among the top-performing algorithms in the ongoing Middlebury benchmark (Scharstein and Szeliski, 2012). The benchmark originated from the studies in (Scharstein and Szeliski, 2002) comparing state-of-the-art stereo methods using a controlled set of test images with complex scene structure and varying texture. It has also been shown that SGM is able to effectively deal with the aforementioned issues (Hirschmüller and Scharstein, 2009). Several combinations of matching cost functions and stereo methods were evaluated using original and degraded test images (e. g. noise, exposure differences).

Furthermore, it has recently been shown that SGM can be implemented in real-time on a variety of platforms. For example, an FPGA implementation (Banz et al., 2011b) and a GPU implementation (Banz et al., 2011a) both reach over 60 fps for VGA images with 128 pixel disparity range. The high algorithmic performance and real-time capability make SGM very attractive for a wide range of applications including low power embedded vision systems and desktop system with off-the-shelf hardware.

Of major relevance to the performance are the smoothness constraints that are imposed by SGM during the cost aggregation step. These constraints are adapted to the image content by means of so-called penalty functions which penalize abrupt changes in the depth information when, according to image content, a change of objects is unlikely. Therefore, the choice of penalty functions has a significant influence on the algorithmic performance and

robustness. Despite the many surveys on SGM, the influence of the penalty functions has not yet been investigated.

In this paper, new penalty functions for the cost aggregation step of SGM are proposed and evaluated. Due to the mutual dependency of matching cost function and penalty function, two matching cost functions for initial correspondence hypothesis are considered. These are based on the rank transform and the census transform (Zabih and Woodfill, 1994), both of which are often used in systems for disparity estimation due to their good performance and efficient implementation possibilities. Each penalty function is parametrized for both matching cost functions using the established data sets with ground truth disparities from (Scharstein and Szeliski, 2002) with and without additional controlled radiometric changes of intensity similar to (Hirschmüller and Scharstein, 2009) as well as noise. Evaluation is performed in terms of, firstly, accuracy and density of the disparity map and, secondly, the insensitivity to the degraded input images.

Section 2 reviews algorithmic background on semi-global matching and disparity estimation. Section 3 details the methodology, experiments and results for the different test sets. Conclusions are drawn in Section 4.

2 STEREO MATCHING

It is important to distinguish between the initial a similarity measure (*matching costs*) between two pixels in the base and match image (or left and right image, respectively) and the aggregation method that uses these costs. In this work, rank transform and census transform (Zabih and Woodfill, 1994) are considered as matching costs functions and semi-global matching (Hirschmüller, 2008) is used for cost aggregation. Final disparity selection is performed by a winner-take-all (WTA) approach.

2.1 Rank Transform

Matching costs $C(\mathbf{p}, d)$ based on the rank transform (RT) of the base and match image R_b and R_m are calculated as

$$C(\mathbf{p}, d) = |R_b(p_x, y) - R_m(p_x - d, y)| \quad (1)$$

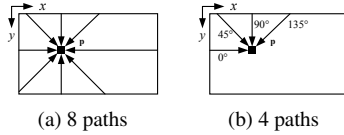


Figure 1: Investigated path orientations for eight and four paths.

where $\mathbf{p} = [p_x, p_y]^T$ is the pixel location in the left image and R is the area-based non-parametric rank-transform. It is defined as the number of pixels \mathbf{p}' in a square $M \times M$ (here: $M = 9$) neighborhood $A(\mathbf{p})$ of the center pixel \mathbf{p} with a luminous intensity I less than $I(\mathbf{p})$

$$R(\mathbf{p}) = \|\{\mathbf{p}' \in A(\mathbf{p}) \mid I(\mathbf{p}') < I(\mathbf{p})\}\|. \quad (2)$$

2.2 Census Transform

The census transform (CT) maps the square $M \times M$ (here: $M = 5$) neighborhood $A(\mathbf{p})$ of the center pixel \mathbf{p} to a bit string where pixels with an intensity I less than $I(\mathbf{p})$ are 1, else 0. Thus,

$$R(\mathbf{p}) = \bigotimes_{\mathbf{p}' \in A(\mathbf{p})} \xi(I(\mathbf{p}'), I(\mathbf{p})) \quad (3)$$

with \bigotimes the concatenation and $\xi(a, b) = 1 \quad \forall \quad a < b, \quad 0 \quad \text{else}$. The matching cost between two pixels is the Hamming distance of the corresponding census transformed pixels, i. e.

$$C(\mathbf{p}, d) = \sum_{i \in 0 \dots M^2 - 1} |R_{b,i}(p_x, y) \neq R_{m,i}(p_x - d, y)| \quad (4)$$

where i indexes the bit position. The census transform is also non-parametric and area based.

Both methods are able to deal with radiometric differences since they depend on the ordering of the pixel's intensities rather than the absolute values. In contrast to the rank transform, spatial information is retained by the census transform.

2.3 Semi-Global Matching

In many cases, pixel-wise calculated matching costs (i.e. locally calculated) yield non-unique or wrong correspondences due to low texture and ambiguity. Therefore, semi-global matching introduces global consistency constraints by aggregating matching costs along several independent, one-dimensional paths across the image. Thereby, SGM aims to approximate a global energy minimization problem which is NP-hard. The paths are formulated recursively by the definition of the path costs $L_r(\mathbf{p}, d)$ along a path \mathbf{r} .

$$L_r(\mathbf{p}, d) = C(\mathbf{p}, d) + \min [L_r(\mathbf{p} - \mathbf{r}, d), L_r(\mathbf{p} - \mathbf{r}, d - 1) + P_1, L_r(\mathbf{p} - \mathbf{r}, d + 1) + P_1, \min_i L_r(\mathbf{p} - \mathbf{r}, i) + P_2] - \min_l L_r(\mathbf{p} - \mathbf{r}, l) \quad (5)$$

The first term describes the initial matching costs. The second term adds the minimal path costs of the previous pixel $\mathbf{p} - \mathbf{r}$ including a penalty P_1 for disparity changes and P_2 for disparity discontinuities, respectively. Discrimination of small changes $|\Delta d| = 1$ pixel (px) and discontinuities $|\Delta d| > 1$ px allows for slanted and curved surfaces on the one hand and preserves disparity discontinuities on the other hand. The last term prevents constantly increasing path costs. For a detailed discussion refer to (Hirschmüller, 2008). P_1 is an empirically determined constant.

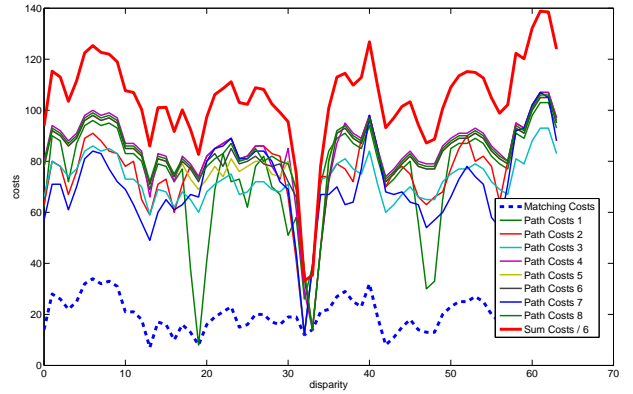


Figure 2: Matching costs, aggregated path costs, and sum costs for the pixel $\mathbf{p} = [183; 278]$ of the Teddy image calculated with census transform and SGM.

P_2 can also be empirically determined or can be adapted to the image content. The selection of these penalty functions is focus of this contribution and will be discussed in 3 in detail.

Quasi-global optimization across the entire image is achieved by calculating path costs from multiple directions to a pixel, as shown in Fig. 1. The aggregated costs S are the sum of the path costs

$$S(\mathbf{p}, d) = \sum_{\mathbf{r}} L_r(\mathbf{p}, d). \quad (6)$$

The disparity map $D_b(p_x, p_y)$ from the perspective of the base camera is calculated by selecting the disparity with the minimal aggregated costs

$$\min_d S(\mathbf{p}, d) \quad (7)$$

for each pixel. For calculating $D_m(q_x, q_y)$, the minimal aggregated costs along the corresponding epipolar lines are selected:

$$\min_d S(q_x + d, q_y, d). \quad (8)$$

The effect of the path costs aggregation and the disparity selection is illustrated in Fig. 2. The initial matching costs (dotted line) exhibit a high level of ambiguity. Seven of the eight aggregated paths costs already show distinct minima. The summed path costs (thick red line) clearly identify the minimum at a disparity level of 32 resolving all ambiguities. However, the cost difference for the positions 32 and 33 is minimal indicating that the correct position is located a sub-pel precision. Half-pel accuracy is obtained by quadratic curve fitting through the neighboring sum costs around the minimum. An evaluation of sub-pel interpolation methods can be found in (Haller and Nedeveschi, 2012).

Both, uniqueness check and left/right check are performed to ensure that only valid disparities with high confidence level are produced. The uniqueness check sets disparities invalid if the minimum $\min_d S(\mathbf{p}, d)$ is not unique. The left/right-check sets disparities invalid if the disparity $D_b(\mathbf{p})$ and its corresponding disparity of D_m differ by more than 1 px. No post-processing steps, e. g. hole-filling or interpolation, are performed. An overview of the processing steps is given in Fig. 3. From (5) and Fig. 2 the crucial role of a "good" selection of P_1 and P_2 for the overall performance is obvious. It is suggested in (Hirschmüller, 2008) to adapt $P_2(I)$ to the image content in order to penalize abrupt disparity changes when, according to image content, a depth discontinuity is unlikely. Therefore, it is clearly necessary to assess different functions for $P_2(I)$ and their influence on the algorithmic performance, which is the scope of this paper.

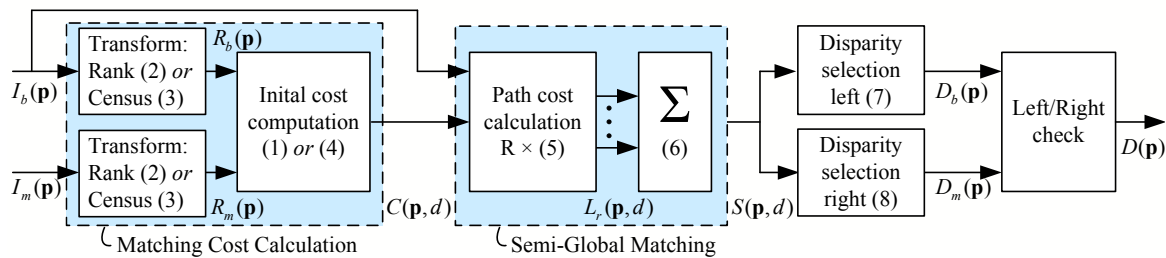


Figure 3: Processing steps for disparity estimation using rank transform/census transform and semi-global matching.

3 EVALUATION AND RESULTS

The four evaluated penalty functions are:

(a) empirically determined constant value, i. e.

$$P_{2,c} = const. \quad (9)$$

(b) negatively proportional to the absolute luminous intensity gradient of the currently processed pixels along the path, i. e.

$$P_{2,l} = -\alpha \cdot |I(\mathbf{p}) - I(\mathbf{p} - \mathbf{r})| + \gamma \quad (10)$$

(c) inversely proportional to the absolute luminous intensity gradient of the currently processed pixels along the path. This follows the original proposal from SGM.

$$P_{2,i} = \frac{\alpha}{|I(\mathbf{p}) - I(\mathbf{p} - \mathbf{r})| + \beta} + \gamma \quad (11)$$

(d) negatively proportional to the variance of the luminous intensity in a local window, i. e.

$$P_{2,v} = -\alpha \cdot \text{Var}(A(\mathbf{p})) + \gamma \quad (12)$$

In all cases it has to be ensured that $P_2 \geq P_1$. Therefore, a lower bound is introduced $P_{2,\min}$ to which the values are clipped. An upper bound is not required because penalty higher than $C_{max} + P_1$ cause that value never to be taken in the outer min-term in Eq. (5). It follows that (b) does not require a parameter β for shift in x direction. This is implicitly done by adjusting γ . Cases (b) and (c) are based on the hypothesis that depth changes are often visible as luminance changes. Case (d) is based on the hypothesis that matching costs in highly structured areas are highly discriminative and luminance changes not only occur due to object changes.

3.1 Methodology and Middlebury Images

For the first set of experiments the established Middlebury stereo data set (Cones, Teddy, Venus and Tsukuba) is used (Scharstein and Szeliski, 2002). These were taken under controlled laboratory conditions. Intensity differences and noise are expected to be minimal. The disparity ranges are 64 px for Cones and Teddy, 32 px for Venus, and 16 px for Tsukuba. Each penalty function is parametrized for each image with both matching cost functions for 4 and 8 paths. The resulting disparity maps are evaluated by counting the number of erroneous disparities in non-occluded areas. An erroneous disparity differs by more than a defined threshold from ground truth. Two thresholds are considered: $|\Delta| > 1$ px and $|\Delta| > 0.5$ px. Percentages stated in the following are the number of erroneous pixels of all non-occluded pixels (not the entire image). Ignoring occluded areas, i. e. where disparities cannot be computed, allows to focus on the performance of the disparity estimation algorithm rather than any post-processing steps. Otherwise, the results would be biased by the

quality of the hole interpolation algorithm. For the same reasons no post-processing steps are applied to the disparity maps.

Questions the first set of experiments is aimed to answer are: Is there a clear favorite among the penalty functions? How sensitive is the performance towards the parametrization of the penalty function? Is the parametrization robust across different images taken with different setups and cameras? These questions are of relevance for real world system since insensitivity towards non-optimal parametrization and camera imposed differences are mandatory.

Fig. 4 shows the results computed with census and 8 paths for the four test images as the parametrization of each function is changed. The parameter configurations for each penalty function are sorted with increasing error and the best 100 configurations are shown. The parameters of each function (P_1 , $P_{2,\min}$, α , β , and γ) are changed systematically with carefully determined step sizes big enough to ensure sufficiently different configuration sets on the one hand and small enough not to miss local minima on the other hand.

Setting P_2 constant performs well if carefully adjusted to the particular image but quality degrades quickly as these values are changed. The adaptive functions $P_{2,l}$ and $P_{2,i}$ perform significantly better with up to 1 percentage points improvement. Both are comparable in terms of quality and superiority is minimal depending on the particular image. The variance based approach performs significantly worse than the other adaptive approaches and sometimes even worse than the fixed approach. This could be due to the fact that $P_{2,v}$ does not calculate penalties along the currently processed path but from the local window giving the same penalty value for all path directions. For the census-based matching costs $P_{2,l}$ and $P_{2,i}$ are the best functions.

The second row of Fig. 4 shows the data re-grouped according to penalty function, this time over all configurations analyzed. All functions are insensitive to a certain degree of non-optimal parametrization to the image content. However, it is also clear that good parametrization is essential for obtaining the maximum of correct information.

The third row of Fig. 4 assess if optimal configurations coincide from image to image. The configurations are now ordered according to the parameter values and same configurations are on the same x -position. Clearly, performance of a particular configuration coincides across all images. Further, the best configuration for one image is usually found for the other images when allowing a minimal 0.5% percentage point error margin. When going from 8 paths to 4 paths (data not shown) the same observations and conclusions can be made with just slightly increased error counts. For half-pel error thresholds there are no changes in configurations (data not shown).

Results employing the rank transform are shown in Fig. 4 fourth row. Error counts for best performance are always slightly higher

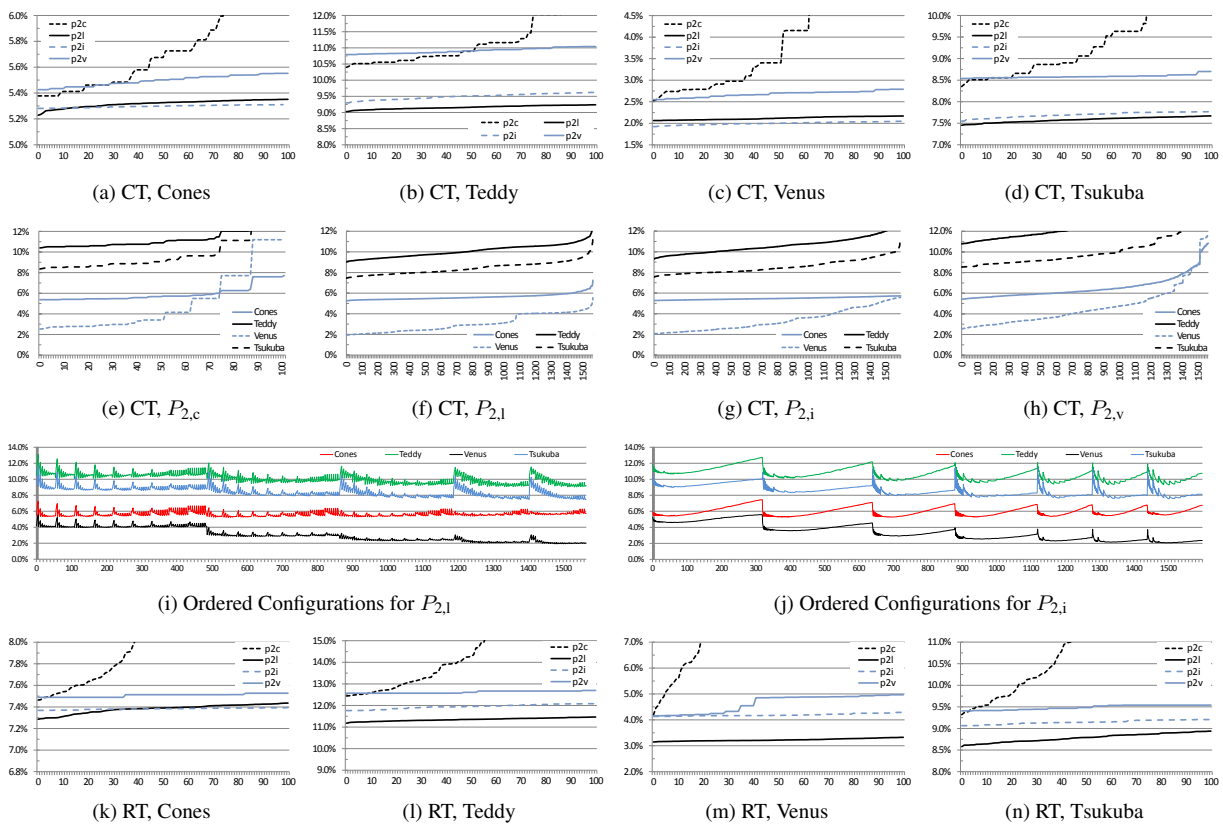


Figure 4: Errors in unoccluded areas for all penalty functions employing census transform examining behavior on the same image (row one), over different images (row two), stability of configurations across images (row three), and employing rank transform (row four).

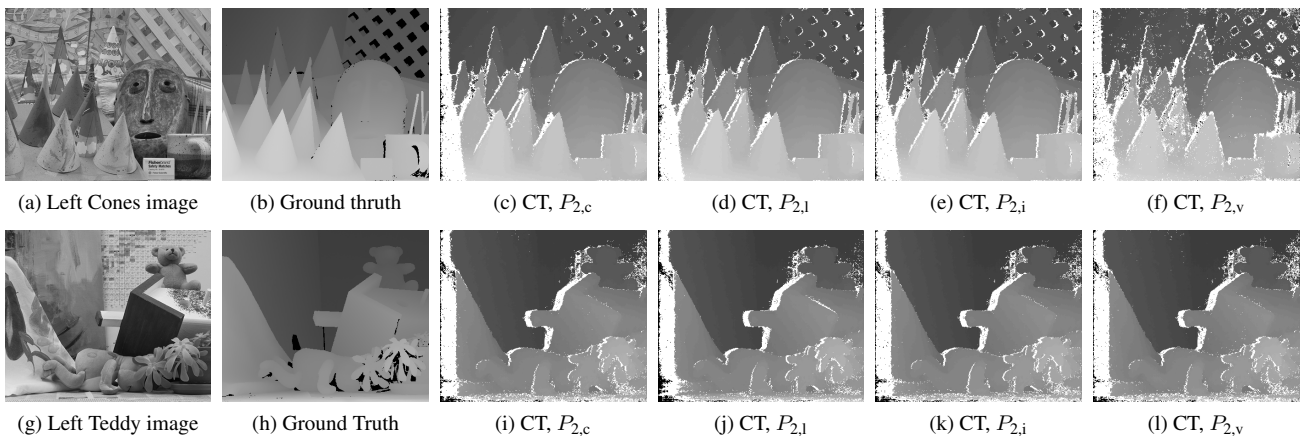


Figure 5: Disparity maps obtained with optimally parametrized penalty functions for Cones (top row) and Teddy (bottom row).

than for the census transform which is due to the missing spatial information of the rank transform. As with the census transform $P_{2,c}$ performs acceptably and $P_{2,v}$ poorly for an adaptive function. However, in opposite to census, $P_{2,l}$ always outperforms $P_{2,i}$; in 3 cases quite significantly. All functions are similarly robust towards parametrization offset as with census (data not shown). Again, good configurations coincide across all images (data not shown).

The error levels obtained with optimally adapted penalty functions are summarized in Table 1. For the two test images Cones and Teddy the resulting disparity maps are shown in Fig. 5. Even visually highly noticeable, $P_{2,v}$ introduces a significant amount of errors. Using $P_{2,i}$ the small structures in Cones are retained, otherwise there is no significant difference between $P_{2,l}$ and $P_{2,i}$. For

comparison, these two functions are plotted with their optimal configuration in Fig. 6 showing an obvious similarity between the two functions over x .

3.2 Simulated Degenerated Images

For the second set of experiments the left input image of the Cones data set is artificially degraded whereas the right remains unchanged. Two types of radiometric differences and two types of noise are considered:

- Additive white Gaussian noise (AWGN) with SNR = 12 dB
- Salt-and-Pepper noise with 14 % of the image degenerated
- Linear brightness (gain) change across the half the image
- Non-linear brightness (gamma) change

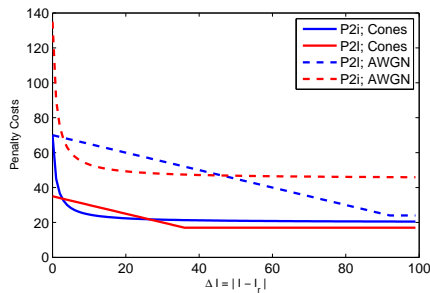


Figure 6: Optimal penalty functions ($P_{2,1}$ and $P_{2,i}$) for the Cones image (solid lines) and in the presence of AWGN (dotted lines).

The degenerated images are shown in Fig. 8 top row. The evaluation methodology remains as with the first experiments. This experiment is aimed at answering the following questions: Are the penalty functions robust towards a variety of different image interferences and which performs best? How robust is the parametrization across these interferences? Is the choice of penalty function and parametrization the same as for the non-degenerated images? The errors curves for the four types of simulated degeneration are shown in Fig. 7. Fixed penalty functions degrade quickly for census and rank, as the constant values are now trained to the type and amount of noise in the image. Adaptive functions are able to cope with the noise and radiometric changes. As with non-degenerated images, $P_{2,1}$ and $P_{2,i}$ perform best and in many cases similar. For rank, $P_{2,1}$ outperforms $P_{2,i}$ except for salt-and-pepper noise. For census, the two functions perform equally except for salt-and-pepper noise where the $P_{2,i}$ significantly outperforms the linear function. The variance based approach is always outperformed by the other adaptive approaches. For qualitative evaluation the resulting disparity maps are shown in Fig. 8.

Comparing configurations across the different types of degeneration shows that good configurations coincide (data not shown). As before, good configurations from one image to the next can often be found within a 0.5 percentage points error margin. However, comparing good configurations to configurations from the non degenerated images shows that now higher dynamic range and higher penalties are chosen. For example, the best parameter set from the original Cones image for $P_{2,1}$ is $\{P_1 = 11, P_{2,\min} = 17, \gamma = 35, \alpha = 0.5\}$ resulting in 5.23 % erroneous disparities. For the AWGN case it is $\{P_1 = 20, P_{2,\min} = 24, \gamma = 70, \alpha = 0.5\}$ and for the salt-and-pepper case $\{P_1 = 14, P_{2,\min} = 24, \gamma = 40, \alpha = 0.5\}$ resulting on the original image in 6.27 % and 5.37 %, respectively. For comparison, the optimal penalty functions for the AWGN case have been included in Fig. 6. Consequently, proper selection and parametrization of penalty functions can make disparity estimation robust to high levels of interferences with only minimal performance decrease in ideal cases. However, it also shows that for high-end applications targeting highest quality disparity maps sophisticated image preprocessing is required.

3.3 Real World Images

For real world image data the lack of ground truth makes it extremely difficult to setup automated parametrization. However, optical inspection using real world data from (Ess et al., 2007) was performed with the parametrizations obtained from the non-degenerated and degenerated images. Special attention has been paid to planar, little textured areas, edges, and small structures (e. g. lamp posts). Generally, better results were obtained when using the configurations from the degenerated images. This is in accordance with the argumentation from above.

f	Cones	Teddy	Venus	Tsukuba
Census Transform				
$P_{2,c}$	5.38 %	10.40 %	2.53 %	8.35 %
$P_{2,l}$	5.23 %	9.03 %	1.92 %	7.45 %
$P_{2,i}$	5.43 %	9.30 %	2.06 %	7.55 %
$P_{2,v}$	5.28 %	10.79 %	2.55 %	8.54 %
Rank Transform				
$P_{2,c}$	7.46 %	12.44 %	4.20 %	9.32 %
$P_{2,l}$	7.29 %	11.17 %	3.15 %	8.58 %
$P_{2,i}$	7.37 %	11.76 %	4.13 %	9.06 %
$P_{2,v}$	7.49 %	12.57 %	4.16 %	9.41 %

Table 1: Errors in non-occluded areas with a threshold of 1 disparity obtained with optimally parametrized penalty functions.

f	Baseline	AWGN	Salt	Shadow	Gamma
Census Transform					
$P_{2,c}$	5.38 %	26.35 %	7.63 %	7.86 %	5.41 %
$P_{2,l}$	5.23 %	18.91 %	8.27 %	7.27 %	5.27 %
$P_{2,i}$	5.43 %	18.94 %	7.40 %	7.26 %	5.30 %
$P_{2,v}$	5.28 %	30.70 %	8.40 %	8.16 %	5.45 %
Rank Transform					
$P_{2,c}$	7.46 %	40.64 %	10.99 %	10.27 %	7.60 %
$P_{2,l}$	7.29 %	32.61 %	11.74 %	7.27 %	7.41 %
$P_{2,i}$	7.49 %	40.61 %	11.11 %	9.94 %	7.47 %
$P_{2,v}$	7.37 %	45.84 %	11.99 %	10.54 %	7.61 %

Table 2: Errors in non-occluded areas with a threshold of 1 disparity obtained with optimally parametrized penalty functions on the cones image under various types of degeneration.

4 CONCLUSIONS AND FUTURE WORK

In conclusion, the choice of penalty function and its parametrization has significant influence on the performance of SGM, especially under difficult imaging conditions (e. g. noise, exposure differences). While for highly structured images taken under near ideal conditions constant penalty functions ($P_{2,c}$) perform well, they tend to become overfitted to the particular imaging conditions and performance is not stable over different conditions. Among the adaptive functions, the linear ($P_{2,1}$) and inversely proportional ($P_{2,i}$) functions significantly outperform the variance based approach. They are also robust to interferences in the images making adaptive penalty terms mandatory for robust disparity estimation. Even then, the quality of the disparity map significantly depends on a suitable penalty function for SGM. Using inversely proportional penalty functions, as originally proposed with SGM, does not result in any performance improvement compared to linear dependencies, which is of interest for computationally limited implementations. Nevertheless, thorough parametrization according to the employed matching cost function is essential. Since parametrization using difficult images results in more robust parameter sets real world systems should be parametrized under these conditions. For all penalty functions, employing census transform instead of rank transform exhibits better disparity maps with less edge blurring because census transform retains spatial information. Future work includes testing the penalty functions for other types of matching cost functions, e. g. mutual information.

ACKNOWLEDGEMENTS

This work has been supported in part by the Hans.-L.-Merkle Stiftung (Stifterverband für die deutsche Wissenschaft).

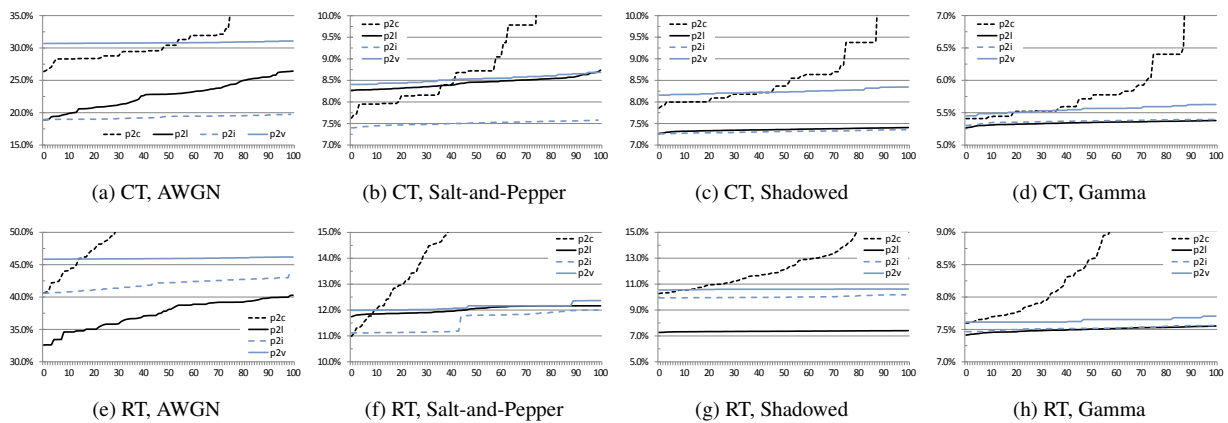


Figure 7: Errors in unoccluded areas for all penalty functions on degenerated input images employing census and rank transform.

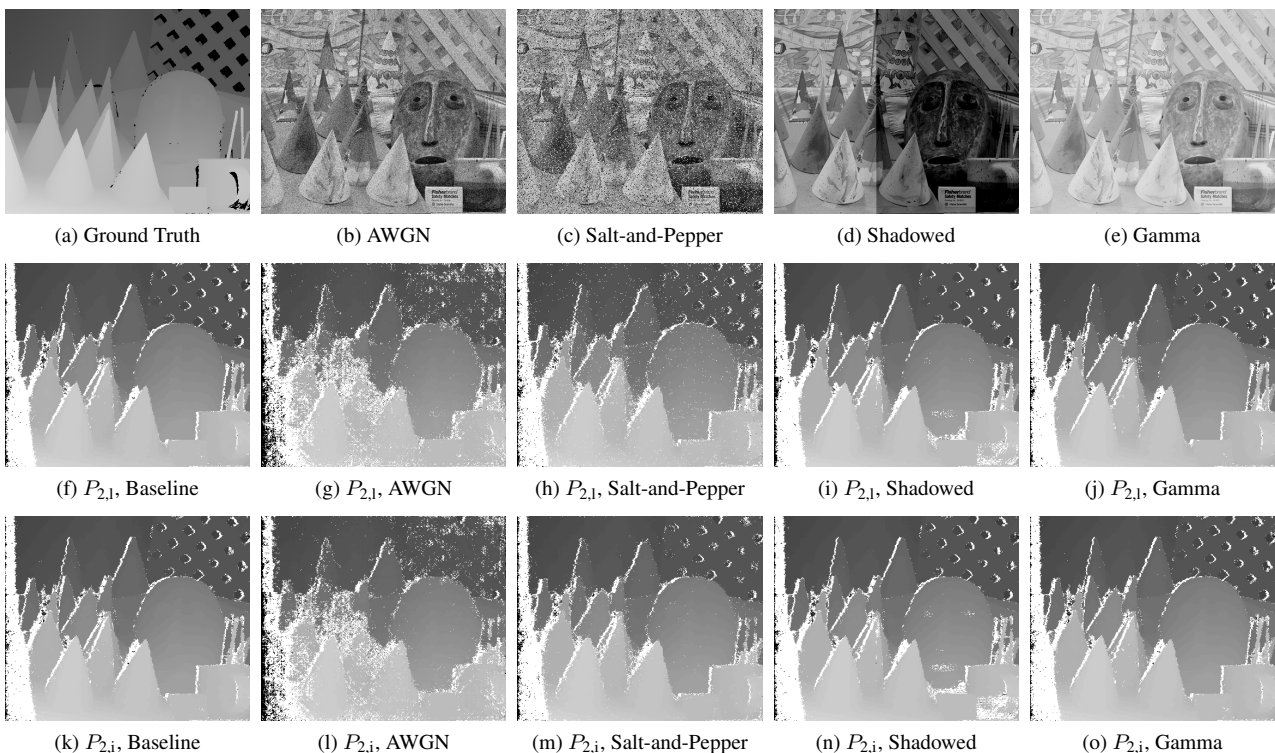


Figure 8: Degenerated input images (row one) and corresponding disparity maps obtained with $P_{2,1}$ (row two) and $P_{2,i}$ (row three) using the census transform.

REFERENCES

- Banz, C., Blume, H. and Pirsch, P., 2011a. Real-time semi-global matching disparity estimation on the gpu. In: Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on, pp. 514–521.
- Banz, C., Hesselbarth, S., Flatt, H., Blume, H. and Pirsch, P., 2011b. Real-time stereo vision system using semi-global matching disparity estimation: Architecture and fpga-implementation. Transactions on High-Performance Embedded Architectures and Compilers, Springer.
- Ess, A., Leibe, B. and Van Gool, L., 2007. Depth and appearance for mobile scene analysis. In: IEEE Intl. Conf. Computer Vision, pp. 1–8.
- Haller, I. and Nedeveschi, S., 2012. Design of interpolation functions for subpixel-accuracy stereo-vision systems. Image Processing, IEEE Transactions on 21(2), pp. 889–898.
- Hirschmüller, H., 2008. Stereo Processing by Semiglobal Matching and Mutual Information. IEEE Trans. Pattern Analysis and Machine Intelligence 30(2), pp. 328–341.
- Hirschmüller, H. and Scharstein, D., 2009. Evaluation of stereo matching costs on images with radiometric differences. IEEE Trans. Pattern Analysis and Machine Intelligence 31(9), pp. 1582–1599.
- Scharstein, D. and Szeliski, R., 2002. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Intl. Journal of Computer Vision 47(1), pp. 7–42.
- Scharstein, D. and Szeliski, R., 2012. The Middlebury Stereo Pages. <http://vision.middlebury.edu/stereo/>.
- Zabih, R. and Woodfill, J., 1994. Non-parametric local transforms for computing visual correspondence. In: European Conference on Computer Vision, pp. 151–158.