Stereo Matching in the Presence of Sub-Pixel Calibration Errors

Heiko Hirschmüller
Institute of Robotics and Mechatronics
German Aerospace Center, Oberpfaffenhofen
heiko.hirschmueller@dlr.de

Stefan Gehrig
Daimler AG
Group Research & Advanced Engineering
stefan.gehrig@daimler.com

Abstract

Stereo matching commonly requires rectified images that are computed from calibrated cameras. Since all underlying parametric camera models are only approximations, calibration and rectification will never be perfect. Additionally, it is very hard to keep the calibration perfectly stable in application scenarios with large temperature changes and vibrations. We show that even small calibration errors of a quarter of a pixel are severely amplified on certain structures. We discuss a robotics and a driver assistance example where sub-pixel calibration errors cause severe problems. We propose a filter solution based on signal theory that removes critical structures and makes stereo algorithms less sensitive to calibration errors. Our approach does not aim to correct decalibration, but rather to avoid amplifications and mismatches. Experiments on ten stereo pairs with ground truth and simulated decalibrations as well as images from robotics and driver assistance scenarios demonstrate the success and limitations of our solution that can be combined with any stereo method.

1. Introduction

Stereo matching commonly requires rectified images that are computed from calibrated cameras. Parametric camera models that are used for calibration are only approximations of physical cameras, especially the lens distortion model. Nevertheless, many methods permit calibration down to sub-pixel accuracy [15, 17], e.g. 0.1 to 0.2 pixel. This is the mean accuracy. Calibration errors can be higher at some parts of the image.

It is typically assumed that these small errors do not cause problems for stereo matching, although it is understood that 3D reconstructions will be slightly biased then. Furthermore, it is assumed that camera calibration is valid as long as the cameras are not physically changed, i.e. lens or cameras re-mounted.

We show theoretically and practically that even cameras that are calibrated to sub-pixel accuracy can cause large matching errors, depending on the scene. We give the example of a service robotics scene with round objects, e.g. glasses, where calibration errors of just 0.25 pixel cause artificial disparity discontinuities of ± 2 pixel.

Another application is stereo vision from a car for driver assistance tasks. The cameras are mounted behind the windscreen, near the rearview mirror, and are affected by large temperature changes and vibration, which causes calibration parameters to drift over time.

It may be argued that this problem could be avoided by mechanical solutions. However, such solutions make the whole system much larger and more expensive. Both is very critical for the main stream use of such a system. Another solution would be online self calibration. However, for our application, self calibration must be performed fully automatic, very robustly and must not lead to temporary unavailability of the system. Dang et al. [4] determined for self calibration an average deviation from the epipolar geometry to be less than 0.2 pixel. However, at outer image parts, deviations of up to 0.6 pixel were measured.

Since calibration errors always exist and cause problems, we are aiming to make binocular stereo matching less susceptible to decalibrations. Our solution is based on signal theory and effectively filters critical structures. It is worth noting that we do not aim to correct decalibration itself. Wrong calibration will still affect reconstruction. We just aim to avoid the amplifying effect that decalibration has on stereo matching.

2. Related Work

We do not know any paper that explicitely considers calibration errors for stereo matching. Typically, cameras are either pre-calibrated using well known methods [17, 15] or self calibrated during operation [4]. Planar rectification transforms the images such that epipolar lines become parallel and coincide with image rows.

There are many stereo images with ground truth [7, 10, 11]. Evaluations on these image sets include comparing different stereo methods [10, 12], different ways of color matching [3] and matching costs on images with radiomet-

978-1-4244-3991-1/09/$25.00 ©2009 IEEE

ric differences [7, 8]. This paper uses ten of these images with ground truth and performs an evaluation that concentrates on the aspect of decalibration.

3. Calibration Error Insensitive Matching

We consider images from a nearly calibrated stereo camera after planar rectification. The epipolar lines are assumed to coincide with image rows.

3.1. Problem Description

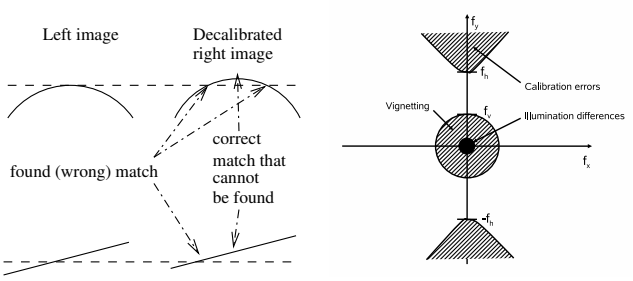
Decalibration means that the camera model does not accurately describe the projection of the real camera. In fact, all (parametric) models can only be approximations of the reality, even with optimally chosen parameters. Thus, a calibration error is always present. We consider the ordinary pinhole camera model with intrinsic parameters like focal length, aspect, skew, principal point and two parameters for radial lens distortion. The extrinsic model describes the 3D transformation between the cameras.

Due to our experience, intrinsic camera calibration is stable under most conditions, provided lenses are internally fixed (e.g. no zoom lenses, no mechanical movement) and they are properly mounted. Extrinsic parameters change more easily due to temperature and vibration. Among extrinsic parameters, rotations are the most critical. Translation is much more stable and can to some extent also be explained by rotations.

Regardless of the kind of decalibration, it leads after rectification to epipolar lines that are not any more aligned to image rows. Fig. 1(a) shows the effect of vertically shifted epipolar lines. Even small decalibrations can lead to high disparity errors or even to completely wrong matches on critical structures. Consider the slanted line in the lower part of the figure: The disparity error e_d can be computed from the epipolar error e_e by $e_d = e_e \cdot \tan(90^\circ - \delta)$ with δ being the angle between horizontal (epipolar) line and the considered line.

Thus, most problematic are nearly horizontal structures, which amplify decalibration errors dramatically. In contrast, vertical structures are matched correctly and the calibration error only propagates into small reconstruction errors. In man made environments, most structures are either horizontal or vertical. Thus, using the stereo camera in a way that epipolar lines are diagonal would reduce the problem. Unfortunately, also lines on the ground that go to infinity (e.g. lane markings for driver assistance) appear diagonal in the images. Therefore, this is not an option for our applications. Alternatively, using a third camera in an L-shape configuration and seeking consistency of both matching directions would solve the problem, but makes the system more expensive and puts a higher burden on processing.

Strong decalibration artifacts are mostly seen in dense



(a) Amplification at vertically shifted epipolar lines

(b) 2D frequency spectrum with unreliable frequencies marked

Figure 1. Description of Problem.

stereo algorithms as sparse stereo implementations usually apply an interest operator to determine stereo only at points with sufficient vertical structure [5]. Dense local stereo methods obtain wrong disparities at horizontal structures. Global stereo methods often even propagate the wrong disparity of horizontal structures into untextured regions. This can lead to phantom objects in the context of an application (see Fig. 10r).

In terms of signal theory, we consider a 2D spectrum of frequencies. The point $(0,0)$ of the frequency spectrum corresponds to a constant offset (i.e. brightness) change. This is commonly avoided for compensating some radiometric differences. Avoiding (f_x, f_y) with $f_x^2 + f_y^2 < f_v^2$, i.e. low frequencies as shown in Fig. 1(b), regardless of direction, removes a dependency on the vignetting effect. Finally, avoiding horizontal frequencies beyond f_h reduces matching problems caused by small decalibrations. In contrast, medium frequencies between f_v and f_h may be valuable for stereo matching.

3.2. Solution Strategy

According to this analysis, one solution is to use a Sobel filter for computing the gradient in x direction (i.e. vertical structure) only. Thus, we perform a convolution with the kernel,

$$S_x = \frac{1}{4}\begin{pmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{pmatrix}. \qquad (1)$$

This eliminates a constant offset due to differentiation as well as all horizontal structures, i.e. vertical frequencies. We will call it XSobel. The effect of the Sobel filter on the spectrum of the image is a high pass filter that suppresses the low frequencies around the f_y axis. Hence all negative effects described above are (almost) erased at the small expense of some information loss around the f_y axis between f_v and f_h.

Our second solution is a modification of background subtraction [1] using a bilateral filter [14], which has been

shown to perform very well for matching images with radiometric differences [8]. A bilateral filter removes all high frequencies, without blurring intensity edges above a certain radiometric threshold. It uses a spatial distance σ_s and a radiometric distance σ_r as parameters. Subtracting the original image by the bilaterally filtered image inverses the outcome and removes all low frequencies (i.e. offset change and vignetting) and is referred to as background subtraction [1]. Our modification consists in applying the bilateral filter only horizontally as,

$$I_f(x,y) = I(x,y) - \frac{\sum_{x' \in N_x} I(x,y) e^s e^r}{\sum_{x' \in N_x} e^s e^r}, \qquad (2)$$
$$s = -\frac{(x'-x)^2}{2\sigma_s^2}, r = -\frac{(I(x',y)-I(x,y))^2}{2\sigma_r^2}.$$

This effectively removes all horizontal structures as well. We will call it HBilSub. Interesting about this filter is, that it is actually much faster to compute than the original bilateral background subtraction. The outcome of both filters are shown in Fig. 2. The effect on the spectrum of the image is similar to that of the Sobel filter. Filtered images can be matched simply with the absolute difference.

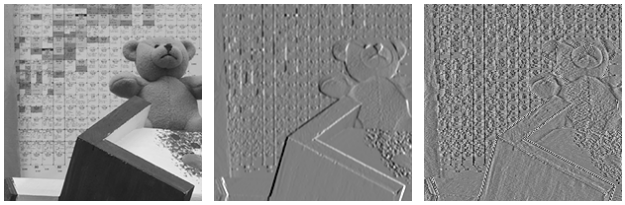


Figure 2. Original image, XSobel and HBilSub filtered with increased contrast for visualization.

4. Results

We use a standard correlation stereo method as well as semi-global matching (SGM) [6] for evaluating the proposed methods. The correlation method uses a 9×9 window, selects the disparity by winner takes all and applies a left/right consistency check. SGM performs pixel-wise matching and approximates the minimization of a global energy by combining pathwise optimizations from all directions through the image. No post filtering is used for both methods. The gaps of consistency checking are closed using pathwise interpolation [6] for both methods.

The XSobel and HBilSub filters, are tested in conjunction with the sum of absolute differences (SAD) for correlation and the pixel-wise, sampling insensitive absolute difference (BT) [2] for SGM. We compare their performance to SAD and BT without filtering. Furthermore, we implemented the BilSub [1], Census [16], using a 9×7 window, and Hierarchical Mutual Information [6] as these costs showed the best performance on images with radiometric differences [8], which plays also a role in our applications.

4.1. Synthetic Decalibration

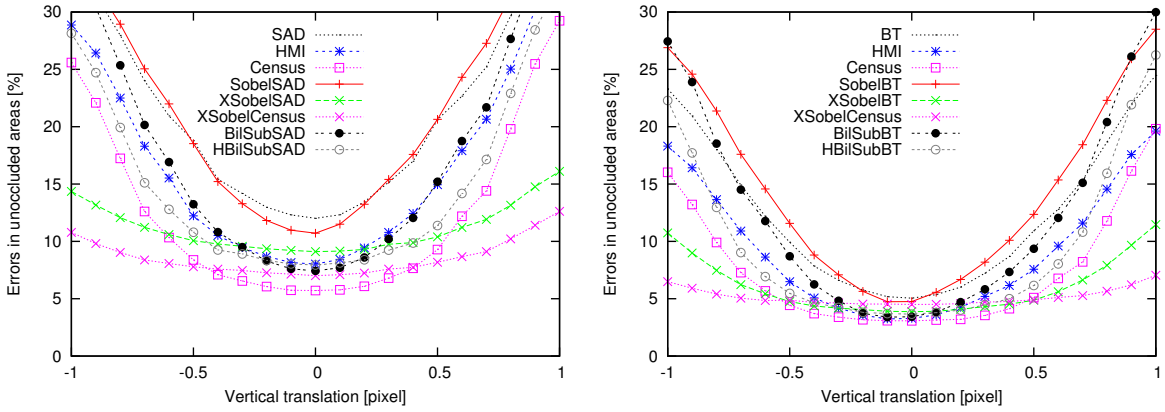
The smoothness parameters of SGM were optimized for each matching cost individually using the unchanged Tsukuba, Venus, Teddy and Cones images (Fig. 3) [10, 11]. After this optimization, the parameters were fixed. We converted all color images into intensity images for matching. The error is measured as the mean percentage of erroneous pixels in unoccluded areas over all images. A pixel is erroneous, if its disparity deviates by more than one pixel from the ground truth.

The first experiment considers a vertical translation of up to ± 1 pixel of the right images. Resampling is done by bilinear interpolation as used in efficient rectification schemes. As shown in Fig. 4, the error increases very quickly with matching costs that are not prepared for decalibration, like SAD/BT, BilSub, HMI and Census. The order of curves is the same as in the study about matching costs for images on radiometric differences [8]. The HBilSub cost performs better than BilSub, as expected. However, XSobel in combination with SAD/BT performs much better for high decalibrations. Surprisingly, the HBilSub result is outperformed by Census at all levels of decalibration and regardless of the stereo method. Also, Census results in lower errors than XSobel, if decalibration is below ± 0.5 pixels. This is due to the fact that the Census implementation uses a much larger window. Large windows increase the stability in general. We found that this is also true for tackling calibration errors. However, larger windows cause blurred depth discontinuities when used with parametric matching costs. This results in higher errors. In contrast, Census as non-parametric cost greatly reduces the blurring problem and benefits from its large window in the presence of calibration errors.

The good performance of XSobel and Census inspired us to combine both. As expected, this yields better results than XSobel with SAD/BT matching, although Census alone performs still a little bit better for small decalibrations. This is an indication that the vertical frequencies between f_v and f_h that the XSobel filter removes are actually useful for the Census matching method.

The same experiment has been performed with the images Art, Books, Dolls, Laundry, Moebius and Reindeer (Fig. 5) [7]. The smoothness parameters of SGM have not been adapted to the new image sets. Since the focal lengths of these images are known, we rotated the right cameras around all three angles. Fig. 6(a) and 6(b) show rotations around the x-axis of the camera coordinate system. This has the same effect as the vertical translation of the last experiment. A rotation of $\pm 0.046^\circ$ results in ± 1 pixel translations. Although the errors are generally higher on the more demanding images, all curves have the same shape as in the first experiment. An interesting observation of Fig. 6(a) and 6(b) is that the lowest error of all costs is not exactly at zero

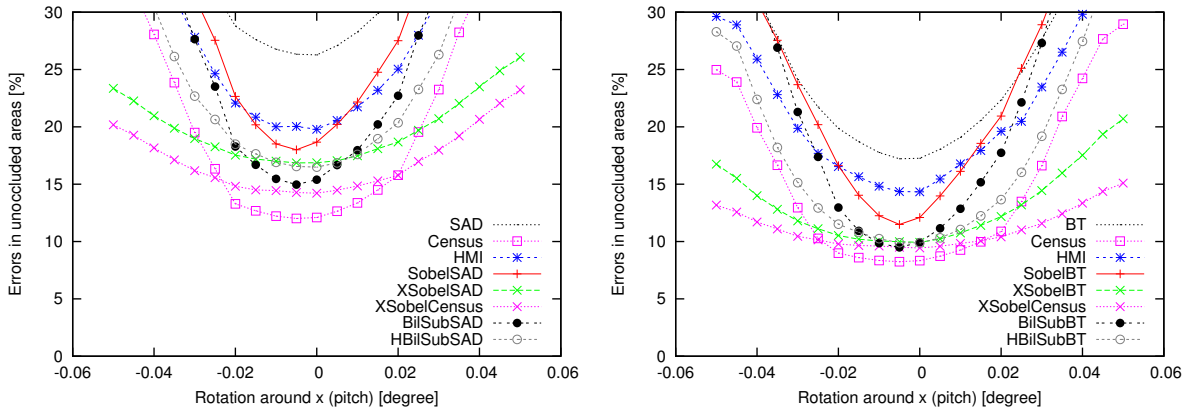
Figure 3. The left images of the Tsukuba, Venus, Teddy, and Cones stereo pairs.



(a) Correlation (b) SGM

Figure 4. Mean errors over the Tsukuba, Venus, Teddy, Cones stereo image set. The right images are vertically shifted.

Figure 5. The Art, Books, Dolls, Laundry, Moebius, and Reindeer stereo pairs.



(a) Correlation, pitch rotation (b) SGM, pitch rotation

Figure 6. Mean errors of the Art, Books, Dolls, Laundry, Moebius, Reindeer stereo image set. The right images are rotated as described.

rotation. This indicates that even these carefully calibrated images are not free of calibration errors.

We performed the same experiments with rotating around the y and z-axis, but do not show the figures due to space limitations. A small rotation around the y-axis keeps the epipolar lines within the image rows. The decalibration error is directly added to the disparities without affecting matching itself. Therefore, all costs degrade in the same way. Census has always the lowest and XSobelCensus always the second lowest error. This rotation is in

practice less critical as it leads only to small reconstruction errors. Rotations around the z-axis affect the outer areas of the image more than the inner areas. In our experiments, the rotation must be about 10 times higher to cause the same amount of error than in the case of the x rotation.

All experiments were repeated using Graph Cuts [13]. The shapes of the curves as well as their relative order are very similar to that of the other stereo methods. Figures are not shown due to space limitations.

In Fig. 7 we show the increment of errors for all images individually using the SGM stereo method. The increment is computed as the error at a rotation of 0.025° (i.e. about half a pixel) around x, divided by the error without any rotation. Although the error increase depends on the individual image, the increase of the best costs, e.g. XSobelCensus, is almost consistently the lowest, which shows that our results are almost independent of the scene content.

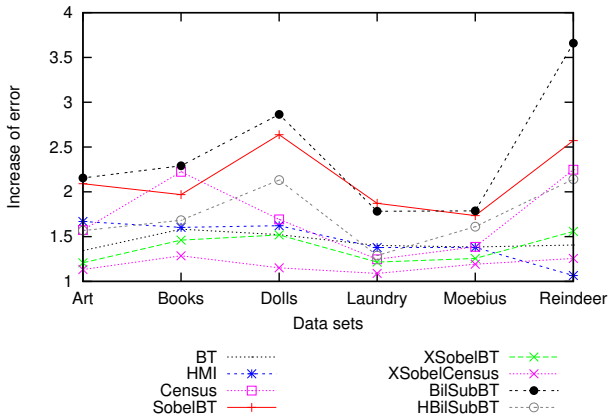


Figure 7. Increase of error of individual stereo images using the SGM stereo method, computed as error at pitch rotation of 0.025° (i.e. about half a pixel), divided by the error without any rotation.

4.2. Stereo Vision for Service Robotics

We studied the effects of our findings for a service robotics scenario in which a humanoid robot is required to detect and classify glasses and bottles using stereo vision [9]. The robot and a typical scenario is shown in Fig. 8. This is a very difficult problem for stereo vision due to the lack of texture and presence of transparency and reflections. Therefore, for glasses, only the upper circular boundary is used for detection and classification.

We tried SGM as well as correlation. SGM produces less blurred object boundaries, but this is not important for the detection and classification of the circular boundary. Therefore, we use the much faster correlation method in this application. Some filters are used additionally to the left/right consistency check for removing outliers on the desk, etc. Furthermore, no interpolation was used. The left part of Fig. 9 shows the disparity images of Fig. 8(b) using different matching costs. White means invalid disparities, i.e. filtered out. The magnified upper part of the left glass is always shown on the right of the full disparity images. Census and XSobelCensus appear to have the best performance, i.e. they cause the smallest gaps on the glass. Both, BilSub and HBilSub perform surprisingly poor.

(a) Service robot (b) Left image of scenario

Figure 8. A service robot and a typical scene.

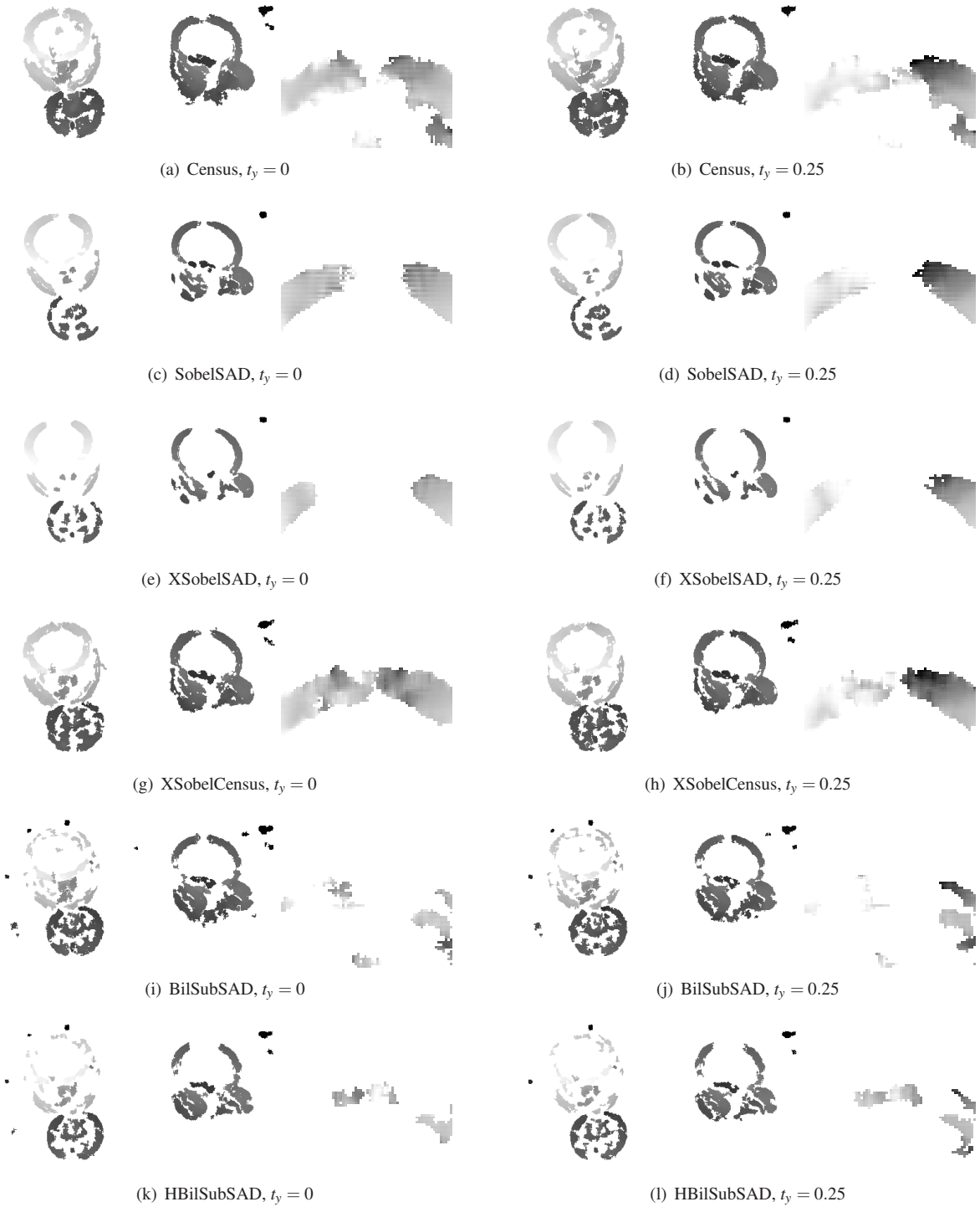
The right part of Fig. 9 shows results of the same methods, but with the right stereo images vertically shifted by about 0.25 pixel. In terms of gaps, Census and XSobelCensus still work best, but all results contain severe amplifications of the decalibration errors. The magnified parts of the disparity images are scaled from their contrast such that the range of white to black corresponds to a disparity range of ± 2 pixel around the true disparity. Thus, decalibration of just a quarter of a pixel causes a disparity error of about ± 2 pixels. Even worse, decalibration causes artificial depth discontinuities that let the circular boundary of the glass appear like a double helix. This kind of error has a very negative influence on the detection and classification of glasses. Using SGM instead of correlation leads to the same phenomenon.

The reason for the failures of our proposed filters is actually simple. Our filters try to suppress horizontal structures in the hope that there is some vertical structure that does not lead to amplifications. However, this extreme application does not have any vertical structure at the upper parts of the glass.

We conclude that the only solution to use binocular stereo in this application is to establish and maintain a very precise calibration, e.g. less than 0.1 pixel calibration error. Another solution would involve using a third camera in an L-shape configuration for trinocular stereo matching.

4.3. Stereo Vision for Driver Assistance

Another application of our proposed filter solution is in a driver assistance scenario. The stereo camera system is mounted behind the wind screen, next to the rear mirror and must robustly operate under very different temperatures and in the presence of vibrations. We found it virtually impossible to keep the calibration stable under these conditions.



(a) Census, $t_y = 0$ (b) Census, $t_y = 0.25$

(c) SobelSAD, $t_y = 0$ (d) SobelSAD, $t_y = 0.25$

(e) XSobelSAD, $t_y = 0$ (f) XSobelSAD, $t_y = 0.25$

(g) XSobelCensus, $t_y = 0$ (h) XSobelCensus, $t_y = 0.25$

(i) BilSubSAD, $t_y = 0$ (j) BilSubSAD, $t_y = 0.25$

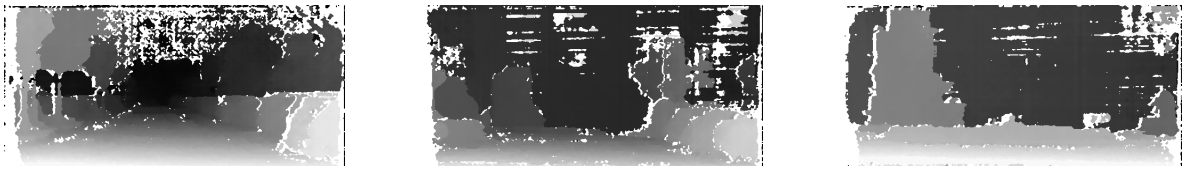
(k) HBilSubSAD, $t_y = 0$ (l) HBilSubSAD, $t_y = 0.25$

Figure 9. Results of the service robot scenario with good calibration (first column) and vertical decalibration by $\frac{1}{4}$th pixel (third column). The second and fourth column show the magnified upper part of the left glass with increased contrast, i.e. ± 2 pixel around the true disparity.

We use the SGM method for processing. The cameras capture intensity images in 12 bits per pixel. The high radiometric depth is used to find enough matches on the otherwise rather textureless street, which is an important issue in this application. Due to the bigger radiometric range, we had to adapt the smoothness parameters.

The top of Fig. 10 shows three typical scenes. The camera is slightly decalibrated. It is important to note that this

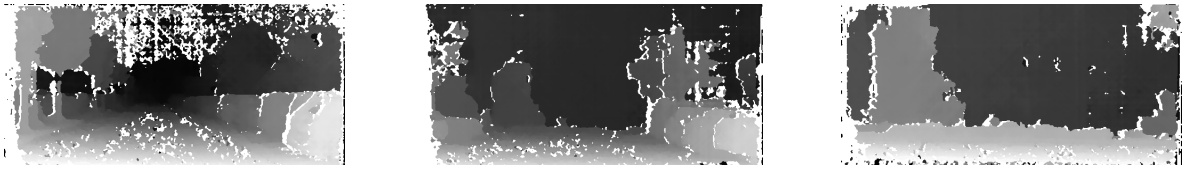
(a) Left image, scene 1 (b) Left image, scene 2 (c) Left image, scene 3



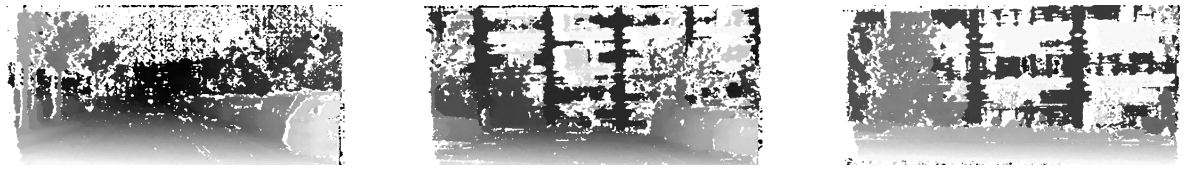
(d) Census (e) Census (f) Census

(g) SobelBT (h) SobelBT (i) SobelBT

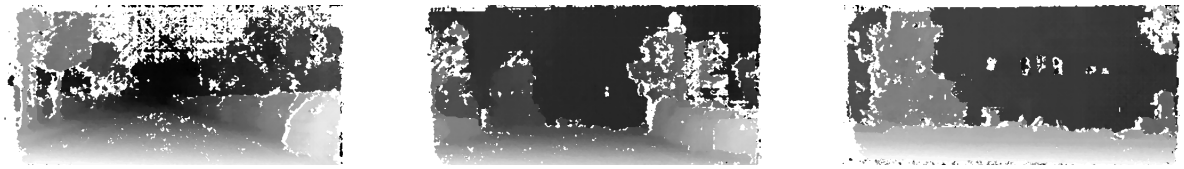
(j) XSobelBT (k) XSobelBT (l) XSobelBT



(m) XSobelCensus (n) XSobelCensus (o) XSobelCensus



(p) BilSubBT (q) BilSubBT (r) BilSubBT



(s) HBilSubBT (t) HBilSubBT (u) HBilSubBT

Figure 10. Three different scenes with slight calibration errors. White means invalid. SGM was used as stereo method.

decalibration is not artificially imposed. The calibration was performed 2 months before the scene was taken and a slight rotational decalibration occurred. Furthermore, due to missing calibration points in the upper third of the image, the intrinsic parameters were slightly wrong estimated which increased the deviation from epipolar geometry to about 0.5 pixel in the upper image part. Fig. 10 shows disparity images of SGM using the left/right consistency check, but without any other post-processing or interpolation. White means invalidated by the consistency check. Since the images contain some radiometric differences as well, we only show results of matching costs that can handle radiometric differences, i.e. no BT.

It can be clearly seen that SobelBT and BilSubBT have severe problems on the nearly horizontal structure, due to inherent decalibration errors. As shown in previous sections, Census works pretty well. XSobelBT, XSobelCensus as well as HBilSubBT all deliver good results and it is not easy to decide from these examples, which one is best. However, XSobelCensus appears to have the lowest number of invalid pixels that are shown in white.

5. Conclusions

We have explained that calibration can never be perfect, because the parametric camera model is only an approximation of a physical camera. Additionally, there are applications, where calibration cannot be kept stable due to large temperature drifts or vibrations. Calibration errors may only be a fraction of one pixel, but we have shown that errors are extremely amplified on nearly horizontal structures. Thus, considering calibration errors is very important for certain applications.

We have proposed to filter out critical image frequencies using either the Sobel filter in x direction only (XSobel) or a horizontal bilateral background subtraction filter (HBilSub). Both filters remove horizontal structures. This is not intended to reduce decalibration errors, but for avoiding the amplifying and destructive effect that it has on matching. The proposed solutions were used with correlation and SGM and compared to common matching costs on images with simulated and real decalibrations.

We found that although HBilSub shows an advantage over BilSub, it is completely outperformed by Census, which has not been designed for compensating decalibrations. In contrast, the XSobel filter resulted in much less errors at decalibrations of more than 0.5 pixels. According to our results, we suggest matching XSobel filtered images with Census as this appeared to be most advantageous.

The limitations of our approach have become visible in one application. Filtering out horizontal structures will only be beneficial, if there is some vertical structure that can be used for matching. Nevertheless, our results in a driver assistance scenario are very encouraging.

References

[1] A. Ansar, A. Castano, and L. Matthies. Enhanced real-time stereo using bilateral filtering. In *3DPVT*, 2004.

[2] S. Birchfield and C. Tomasi. A pixel dissimilarity measure that is insensitive to image sampling. *IEEE TPAMI*, 20(4):401–406, April 1998.

[3] M. Bleyer, S. Chambon, U. Poppe, and M. Gelautz. Evaluation of different methods for using color information in global stereo matching approaches. In *Intern. Arch. of the Photogr., Remote Sensing and Spatial Inform. Sciences*, volume XXXVII, B3a, Bejing, China, 2008.

[4] T. Dang, C. Hoffmann, and C. Stiller. Self-calibration for active automotive stereo vision. In *IEEE Intelligent Vehicles Symposium*, pages 364–369, Tokyo, Japan, June 2006.

[5] U. Franke and A. Joos. Realtime stereo vision for urban traffic scene understanding. In *Intelligent Vehicles*, 2000.

[6] H. Hirschmüller. Stereo processing by semi-global matching and mutual information. *IEEE TPAMI*, 30(2):328–341, February 2008.

[7] H. Hirschmüller and D. Scharstein. Evaluation of cost functions for stereo matching. In *IEEE CVPR*, Minneapolis, Minnesota, USA, 18-23 June 2007.

[8] H. Hirschmüller and D. Scharstein. Evaluation of stereo matching costs on images with radiometric differences. *to appear in IEEE TPAMI*, 2009.

[9] C. Ott, O. Eiberger, W. Friedl, B. Bauml, U. Hillenbrand, C. Borst, A. Albu-Schaffer, B. Brunner, H. Hirschmüller, S. Kielhofer, R. Konietschke, M. Suppa, T. Wimbock, F. Zacharias, and G. Hirzinger. A humanoid two-arm system for dexterous manipulation. In *6th IEEE-RAS International Conference on Humanoid Robots*, pages 276–283, Genoa, Italy, December 2006.

[10] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 47(1/2/3):7–42, April-June 2002.

[11] D. Scharstein and R. Szeliski. High-accuracy stereo depth maps using structured light. In *IEEE CVPR*, volume 1, pages 195–202, Madison, Winsconsin, USA, June 2003.

[12] D. Scharstein and R. Szeliski. Middlebury stereo website. www.middlebury.edu/stereo, 2008.

[13] R. Szeliski, E. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agrawala, M. Tappen, and C. Rother. A comparative study of energy minimization methods for markov random fields with smoothness-based priors. *IEEE TPAMI*, 30(6):1068–1080, June 2008.

[14] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. *IEEE ICCV*, pages 836–846, 1998.

[15] R. Y. Tsai. A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf- tv cameras and lenses. *IEEE Journal of Robotics and Automation*, 3(4):323–344, August 1987.

[16] R. Zabih and J. Woodfill. Non-parametric local transforms for computing visual correspondance. In *Proceedings of the European Conference of Computer Vision*, pages 151–158, Stockholm, Sweden, May 1994.

[17] Z. Zhang. A flexible new technique for camera calibration. *IEEE TPAMI*, 22(11):1330–1334, November 2000.