Parameter Optimization of the SAD-IGMCT for Stereo Vision in RGB and HSV Color Spaces

Kristian Ambrosch
AIT Austrian Institute of Technology
A-1220 Vienna, Austria
Email: kristian.ambrosch@ait.ac.at

Martin Humenberger
AIT Austrian Institute of Technology
A-1220 Vienna, Austria
Email: martin.humenberger@ait.ac.at

***Abstract*—Dependable 3D perception modules are essential for safe operation of robotic platforms. Furthermore, robot navigation and localization as well as object recognition tasks also require processing 2D color camera images. This information could be synchronously delivered by stereo vision sensors with the 3D information automatically mapped onto the 2D camera image. However, embedded real-time stereo vision sensors are often restricted to grayscale images due to limited computational resources. Therefore, we present how a previously designed real-time stereo matching algorithm, the SAD-IGMCT, can be optimized for the RGB and HSV color spaces, further reducing the algorithm's complexity, while still allowing for a high accuracy.**

I. Introduction

For safe operation in uncontrolled environments of robot platforms, dependable 3D perception modules are needed for a reliable description of the surrounding area. Such 3D sensors have to be embedded because processing power is limited on robot platforms and thus should not be additionally stressed with 3D data calculation. Beside that, the 3D perception has to be capable for real-time systems which means that the processing has to be fast and the processing time has to be known and scene independent. Commonly used embedded real-time 3D sensors for mobile robots are laser range finders or laser scanners (LIDAR, light detection and ranging) and time-of-flight (TOF) cameras. Both have the advantage of delivering accurate 3D data, but suffer from low resolution. A promising alternative for robot navigation and mapping is stereo vision. In contrast to time-of-flight and laser, stereo vision delivers 3D information and camera images of the captured environment synchronously. This makes it very well suited for robot applications because the camera images can be additionally used for other tasks such as scene classification. State-of-the-art embedded real-time stereo sensors use grayscale cameras for processing because the image quality, in terms of resolution, noise and sharpness, is better than for Bayer patterned color cameras. Additionally, when using color as matching criterion, the algorithm complexity increases in comparison to using grayscale only. A well known embedded stereo vision sensor is the Small Vision System from Videre Design [9]. It uses Sum of Absolute Differences (SAD) as matching algorithm and a Field Programmable Gate Array (FPGA) as purely embedded processing platform. Another SAD-based stereo sensor is the Mobile Ranger [1] from Mobile Robots Inc. It uses a PCI board, equipped with an FPGA for stereo processing. A different stereo matching algorithm is used by the DeepSeaG2 processor from Tyzx Inc. [10], [11]. It is based on the non-parametric Census (in detail described later on) transform [12] and uses a special stereo processor chip for the matching task.

However, color information is essential for other computer vision tasks on robot platforms such as segmentation for object recognition or scene classification, even if recent works put the achievable gain in accuracy into perspective [3], [6]. In this paper, we show how an embedded and real-time capable stereo matching algorithm, introduced in a previous work [2], can be optimized for the use on RGB and HSV color spaces, while keeping the algorithmic complexity low.

II. Absolute Difference and Gradient-Based Modified Census Transform

The Census transform [12], is a non-parametric algorithm with a high robustness to illumination variations. The Census transform consists of a comparison function ξ, which is used to compare the center pixel's intensity value i_1 with the pixel intensity values i_2 in the neighborhood region, i.e. a block with the dimensions $s_t \times s_t$.

$$\xi(i_1, i_2) = \begin{cases} 1 & | \quad i_1 > i_2 \\ 0 & | \quad i_1 \leq i_2 \end{cases} \qquad (1)$$

Its result is then concatenated ($\bigotimes$) to a bit vector. Thus, the transformation function T_{census} is defined as

$$T_{census}(I, x, y, s_t) = \bigotimes_{[n,m]} \xi[I(x,y), I(x+n, x+m)] \qquad (2)$$

where

$$n, m \;\epsilon\; [-\frac{s_t - 1}{2}, \frac{s_t - 1}{2}]. \qquad (3)$$

For the cost function, the Hamming distance is calculated over the bit vectors. Since we are working on color images, we compute each color channel separately and aggregate the matching costs before the cost selection.

As revealed in [2], using the original Census transform on gradient images does not allow for an increase in accuracy, because it is not able to handle blocks with a saturated center pixel. Therefore, to allow the extension of the Census transform to gradient images, it is necessary to use the Modified Census Transform (MCT) [4], where the center pixel in the transform is replaced by the mean value, which is calculated over the whole block. Since using MCT on the gradient images additionally to the original one triples the algorithm's complexity, we are using a so called sparse computation for the Hamming distance, where only every fourth bit within the bit-vector is used. This way, the overall complexity can be reduced, while the drop in accuracy is at a minimum level.

For the computation of the MCT on the gradient images, we compute the Sobel filtered images for each color channel in x and y direction. Then, we are aggregating the matching costs over an image region with block size $s_a \times s_a$. Thus, the total block size s_b for the stereo matching algorithms is defined as $s_b = s_t + s_a$.

Since the MCT is a non-parametric algorithm we also compute the absolute difference of the blocks' center pixels, to introduce a small parametric measure as well. For a detailed analysis on the absolute difference and gradient-based MCT, called SAD-IGMCT in the following, as well as the sparse computation of the Hamming distance, see [2].

This algorithm consists of three sub-algorithms: the MCT on the intensity images, the MCT on the gradient images, and the absolute difference of the blocks' center pixels. The matching costs for these three algorithms are aggregated after applying a weighting factor to the MCT on the gradient images (W_{grad}) and the absolute differences (W_{ad}). Finally, the matching costs for all color channels are aggregated.

Now, the most accurate matching costs have to be searched for. Their position defines the resulting disparity map's pixel value $d_{map_{x,y}}$. Here, we are using the Winner Takes All (WTA) algorithm, due to its small complexity which suits real-time implementations very well.

For the sub-pixel refinement it is necessary to interpolate the resulting matching costs from the stereo matching, refining the position of the absolute minimum to values that can be in between two pixels. In this work, parabola fitting [8] is used as described in equation 4.

$$\hat{d}_{map} = \frac{a_{x,y,d_{map}-1} - a_{x,y,d_{map}+1}}{2a_{x,y,d_{map}-1} - 4a_{x,y,d_{map}} + 2a_{x,y,d_{map}-1}} \tag{4}$$

Due to the cameras' different viewpoints it can occur that some regions that are visible by one camera, are occluded in the other camera's perspective. For these image areas, it is not possible to find a correlation, and therefore the disparity values for these areas cannot be correct. Thus, it increases the quality of the disparity map, if all matches within occluded areas are removed.

The left/right consistency check [5] takes the disparity map having the left camera image as the primary image and compares it to the one having the right camera image as primary, or vice versa. If the disparity maps' values have too high deviances for the same object point they are disregarded. This way wrong matches are detected and removed. Here, we are using a maximum deviation of just 0.5 pixels. This way, the left/right consistency check not only removes mismatched areas but also most of the incorrect matched surfaces.

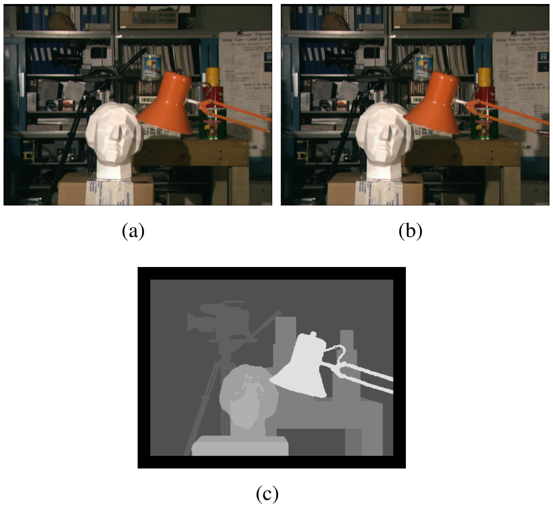
In this work, we are not using further post-processing, such as the interpolation of mismatched or occluded regions, because we are focusing on the optimization of the stereo matching algorithm's accuracy and extensive post-processing might lead to misleading results.

III. Algorithmic Optimizations for RGB and HSV Color Spaces

When using color images for stereo vision, the algorithm's complexity is multiplied with the number of color channels used. However, real-time implementations have limited computational resources. Hence, a stereo matching algorithm suitable for embedded real-time systems must have a low computational complexity. Therefore, we analyzed how the complexity of the stereo matching algorithm can be reduced, while allowing only for a minor drop in accuracy.

To discuss this analysis and the impact on the stereo matching algorithm, we are using the Tsukuba stereo images [7] as depicted in figure 1 from the Middlebury dataset. We chose this dataset, since it presents a typical scene for a domestic robot. Furthermore, in difference to other stereo image pairs in the Middlebury Ranking, it does not have colors with high contrast and can therefore assumed to be rather challenging for color stereo vision.

In this work we are using $s_t = 15$ and $s_a = 5$ for further analysis. These block sizes proved to enable a good accuracy on the Middlebury dataset as well as on real world images, captured with industrial cameras. Thus, we chose this block size even if the Tsukuba images would encourage the use of very large block sizes. Since it would not be useful to discuss this analysis based on a block size suiting one image pair only, we use this generic configuration.



(a) (b)

(c)

Fig. 1. Tsukuba dataset: (a) left image; (b) right image; (c) ground truth.

A. RGB Color Space

For the RGB color space we analyzed the optimum parameters for the SAD-IGMCT. Here, we used a parameter variation for the gradient weight W_{grad} from 0 to 10 with step size 1 and for the absolute differences we varied W_{ad} from 0 to 50 with step size 10. The results are very interesting, as the best accuracy was achieved using $W_{grad} = 10$ and $W_{ad} = 50$. Here, the number of pixels within 0.5 pixel deviation are 64.48%. While our previous works on gray scale images resulted in rather average weighting for the gray scale and gradient images, the results on the RGB images are quite different.

Figure 2 presents the algorithm's accuracy for the SAD-IGMCT where $W_{ad} = 0$, i.e., without the computation of the center pixels' absolute differences. The results show, that the accuracy is highly increasing with the gradient weight and with an accuracy of 64.24% the result for $W_{grad} = 10$ and $W_{ad} = 0$ is 0.24% slightly lower than for $W_{ad} = 50$. Furthermore, it also shows the impact of the gradient images, as the computation of the original RGB channels only, i.e., with $W_{grad} = 0$ and $W_{ad} = 0$, only allows for an accuracy of 55.47%.

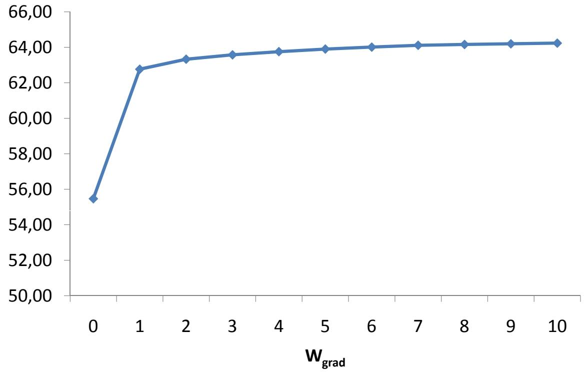


Fig. 2. Accuracy of the SAD-IGMCT on RGB images depending on the gradient weight W_{grad}, with $W_{ad} = 0$.

Hence, we performed the stereo matching on the gradient images of the RGB channels only, varying just the weight for the absolute difference W_{ad} from 0 to 200 having step size 10. The results are depicted in figure 3. Here, the accuracy for $W_{ad} = 0$ is with 64.24% nearly exactly the same as $W_{grad} = 10$ and $W_{ad} = 0$ when the original RGB channels are included in the computation. The best overall accuracy can be achieved having $W_{ad} = 10$ leading to 64.73% correct matches.

However, when considering the limited resources of embedded real-time system, this tiny increase in accuracy does not compensate the computational resources required for the computation of the absolute differences. Thus, it can be assumed that the ideal configuration for using the Census transform on RGB color images is the computation of the MCT on the gradient of the color channels in x and y direction. The disparity image resulting from this configuration is presented in figure 4.

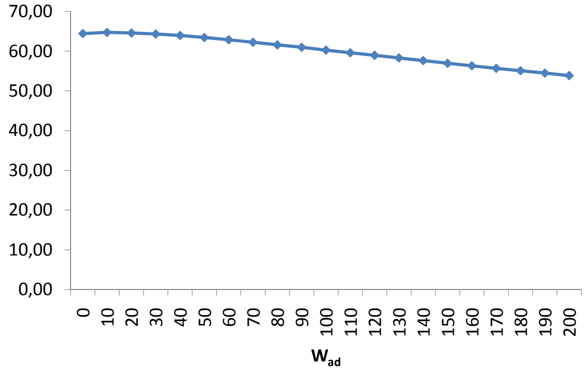


Fig. 3. Accuracy of the SAD-IGMCT the RGB gradient images only, depending on the absolute differences weight W_{ad}.

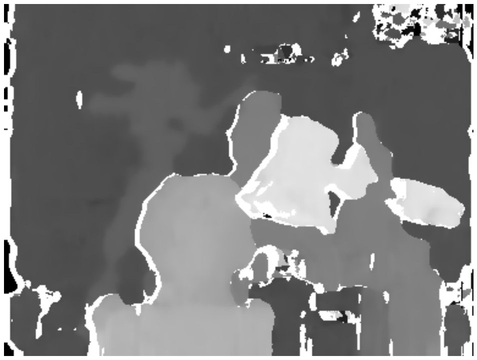


Fig. 4. Disparity calculated for the Tsukuba image set using the gradient of RGB color channels only.

Hue		Saturation		Value	
W_{grad}	W_{ad}	W_{grad}	W_{ad}	W_{grad}	W_{ad}
0	0	3	0	10	50

TABLE I
OPTIMIZED VALUES FOR THE SAD-IGMCT ON THE HSV COLOR SPACE.

B. HSV Color Space

While we were required to treat all color channels of the RGB color space equally to ensure a generic result that is not depending on the dedicated color values appearing in our test image, we analyzed the channels of the HSV color space separately. Here, we varied W_{grad} from 0 to 10 with step size 1 and W_{ad} from 0 to 50 with step size 10 for all three channels of the HSV color space, computing 287496 samples. The best result in this color space was 63.06%, with the parameters presented in table I.

Based on these results, we selected the gradient of the saturation channel as well as the gradient and absolute differences of the value channel as the dominant ones for the matching performance. Thus, we performed the computation on these values only, i.e., setting $W_{grad} = 1$ for the saturation channel and varying W_{grad} from 0 to 10 for the value channel, leading to the results depicted in figure 5. Here, the best result is 63.21% correct matches when using $W_{grad} = 3$ for the value channel. The disparity map resulting from this configuration is presented in figure 4.

The accuracy depending on the weight of the absolute

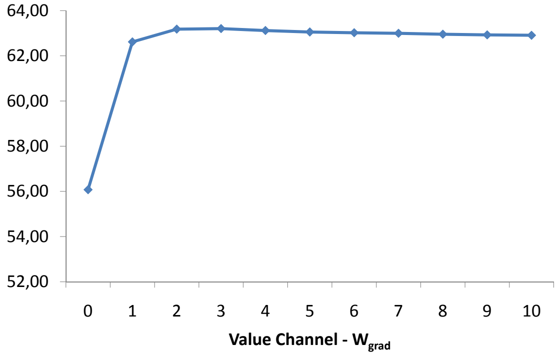


Fig. 5. Accuracy of the SAD-IGMCT on the gradient of the saturation and value channels only, with different W_{grad} for the value channel.

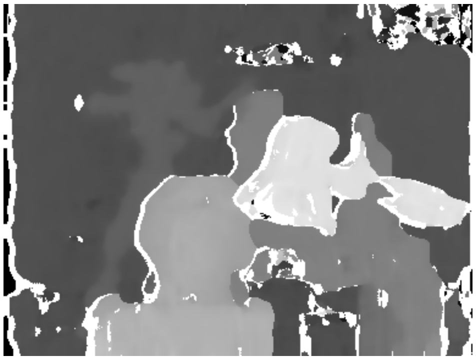


Fig. 6. Disparity calculated for the Tsukuba image set using only the gradient of the saturation having $W_{grad} = 1$ and value channel having $W_{grad} = 1$.

difference W_{ad} on the value channel is presented in figure 7. Here the best result is 63.49% which is once again only slightly better than the results without absolute difference. Hence, reducing the SAD-IGMCT to the MCT on the gradient images of the saturation and value channels of the HSV color space results in nearly exactly the same accuracy, leading to a highly reduced complexity.

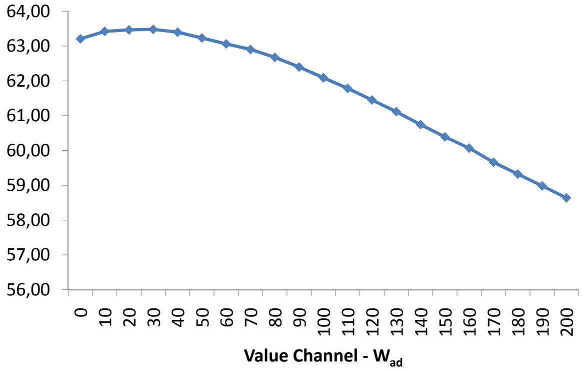


Fig. 7. Accuracy of the SAD-IGMCT on the gradient of the saturation having $W_{grad} = 1$, gradient of the value channel having $W_{grad} = 3$, and different W_{ad} for the value channel.

IV. Conclusions and Future Work

Our analysis on the SAD-IGMCT on the RGB and HSV color space revealed that in difference to grayscale images the algorithm can be reduced to matching the gradient of specific channels only. For the RGB channel, the best results can be expected when using the gradient values of all three color channels, while the HSV allows for a reduction to the saturation and value channels only. Even if the HSV color space requires matching of two channel gradients only, the RGB color space resulted in a slightly better accuracy.

Even if the Tsukuba image is a very good stereo set when focusing on real-time stereo vision for domestic robots, this is only a first step towards a generic analysis on Census-based stereo vision on color images. Thus, we will extend our work on evaluating the optimum parameters for the whole Middlebury dataset, containing multiple stereo image pairs with highly contrasted colors. However, even if the behavior of stereo matching algorithms on images with colors that are rich in contrast is very interesting, a stereo vision system designed for robotics applications will be required to handle both, images with highly and poor contrasted color. Thus, we do not expect the parameters for the final stereo system to be based mainly on image sets without strong color such as the Tsukuba images.

Acknowledgments

The research leading to these results has received funding from the European Community's Sixth Framework Programme (FP6/2003-2006) under grant agreement n° FP6-2006-IST-6-045350 (robots@home).

References

[1] Mobile robots inc. mobileranger, datasheet.

[2] Kristian Ambrosch. *Mapping Stereo Matching Algorithms to Hardware*. PhD thesis, Vienna University of Technology, 2009.

[3] Michael Bleyer and Sylvie Chambon. Does Color Really Help in Dense Stereo Matching? In *Proceedings of the International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT) 2010*, 2010.

[4] Bernhard Froeba and Andreas Ernst. Face detection with the modified census transform. In *Proceedings of the Sixth IEEE Conference on Automatic Face and Gesture Recognition*, 2004.

[5] Pascal Fua. A parallel stereo algorithm that produces dense depth maps and preserves image features. *Machine Vision and Applications*, 6:35–49, 1993.

[6] Heiko Hirschmueller. Evaluation of Stereo Matching Costs on Images with Radiometric Differences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31:1582–1599, 2009.

[7] Daniel Scharstein and Richard Szeliski. A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. *International Journal of Computer Vision*, 47(1–3):7–42, 2002.

[8] Masao Shimizu and Masatoshi Okutomi. Precise Sub-pixel Estimation on Area-Based Matching. In *Proceedings of the eight IEEE International Conference on Computer Vision*, 2003.

[9] STOC Data-Sheet. Videre Design, 865 College Avenue CA 94025 USA, www.videredesign.com, 2008.

[10] Tyzx, Inc. *DeepSeaG2 Product Datasheet*.

[11] John Iseling Woodfill, Gaile Gordon, Dave Jurasek, Terrance Brown, and Ron Buck. The Tyzx DeepSea G2 Vision System, A Taskable, Embedded Stereo Camera. In *Proceedings of the 2006 Conference on Computer Vision and Pattern Recoginition Workshops*, 2006.

[12] Ramin Zabih and John Iseling Woodfill. Non-parametric Local Transforms for Computing Visual Correspondence. In *Proceedings of the 3rd European Conference on Computer Vision*, 1994.