# Stereo Vision in Structured Environments by Consistent Semi-Global Matching

Heiko Hirschmüller
Institute of Robotics and Mechatronics Oberpfaffenhofen
German Aerospace Center (DLR), 82230 Wessling, Germany.
heiko.hirschmueller@dlr.de

## Abstract

*This paper considers the use of stereo vision in structured environments. Sharp discontinuities and large untextured areas must be anticipated, but complex or natural shapes of objects and fine structures should be handled as well. Additionally, radiometric differences of input images often occur in practice. Finally, computation time is an issue for handling large or many images in acceptable time. The Semi-Global Matching method is chosen as it fulfills already many of the requirements. Remaining problems in structured environments are carefully analyzed and two novel extensions suggested. Firstly, intensity consistent disparity selection is proposed for handling untextured areas. Secondly, discontinuity preserving interpolation is suggested for filling holes in the disparity images that are caused by some filters. It is shown that the performance of the new method on test images with ground truth is comparable to the currently best stereo methods, but the complexity and runtime is much lower.*

## 1. Introduction

Structured environments can contain many difficulties for stereo vision. There are often sharp depth discontinuities and large untextured areas. However, there may also be more natural, rounded or fine structured objects like chairs or plants. All of these situations need to be handled for computing accurate disparity images for applications like reconstruction. Additionally, radiometric differences of input images can often not be avoided in practice, especially if individual images are taken at different times and autocalibrated later. Finally, processing time is often an important issue for processing either large or many images in acceptable time.

Section 2 reviews some stereo algorithms that currently produce the best results. The Semi-Global Matching (SGM) method is selected as it addresses already most of the requirements. SGM is reviewed and its behavior analyzed on images in structured environments in Section 3. Two novel refinements are suggested for tackling specific problems in Section 4. Finally, Section 5 shows the performance of the new method on standard test images as well as images of typical indoor scenes.

## 2. Literature review

Since the publication of Scharstein and Szeliski's taxonomy of stereo algorithms [9], many authors have participated in an on-line evaluation [8]. The addition of more complex test images [10] has lead to the new Tsukuba, Venus, Teddy and Cones data set. Especially the last two image pairs are quite complex and realistic.

Almost all of the currently top-ranked algorithms [11, 13, 2, 5, 7, 14] on this data set define a global energy function that is minimized for finding the disparities. This energy function always includes a data term and a smoothness term. The former evaluates the matching of individual pixels, while the latter supports piecewise smooth disparity selections. Some methods use more terms for penalizing occlusions [2, 7] or alternatively treating visibility [11, 13]. Furthermore, some methods [11, 13, 14, 5] enforce the consistency of the disparity of all used stereo images (e.g. left/right consistency or symmetry).

The strategies for finding the minimum of the global energy function differ. The classical approach is Graph Cuts [7], which casts the problem into finding the minimum cut in a graph. Belief Propagation [11] iteratively sends messages between neighboring nodes in the four connected image grid for minimizing the global cost. One of the best ranked variant forces symmetrical matching and uses segmentation as soft constraints. Layered approaches [2, 14] perform image segmentation and use the assumption that disparities vary smoothly (*e.g.* planar) within each segment. An initialization by a simple method like correlation is typically complemented by an iterative refinement of the disparity selection. Furthermore, there are also methods [13] that combine Belief Propagation with segmentation and plane fitting in an iterative loop. Finally, the Semi-Global Matching [5] method sums for each pixel the costs along 1D paths from several directions. In contrast to all other methods,

*IEEE Conference on Computer Vision and Pattern Recognition (CVPR), New York, NY, USA, June 17-22, 2006.*

its pixelwise matching cost is not based on comparing intensities directly, but on Mutual Information [12] between the stereo images. This makes it very robust against radiometric differences and some violations of the assumption of lambertian surfaces, *e.g.* reflections.

The complexity of a global algorithm is usually high and can depend on the complexity of the scene [2]. Consequently, most of these methods [11, 13, 2, 7] have a runtime of at least a minute on typical images. In contrast, the Semi-Global Matching method [5] has a complexity of $O(WHD)$ (i.e. number of pixels times disparity range) and a runtime of just 1-2s under similar conditions.

The Semi-Global Matching method has been selected for the discussed problem, due to its accuracy at depth discontinuities, robustness of matching in the presence of radiometric differences and execution speed.

## 3. Semi-Global Matching (SGM)

A stereo algorithm uses a base image $I_b$ and a match image $I_m$ for calculating a disparity image $D$ that corresponds to the base image. It is assumed that the epipolar geometry between the images is known. An epipolar line $e_{bm}(\mathbf{p},d)$ in the match image is defined by the pixel $\mathbf{p}$ in the base image and the disparity $d$ as line parameter. For rectified images $e_{bm}(\mathbf{p},d) = [\mathbf{p}_x - d, \mathbf{p}_y]^T$. It is noteworthy that certain camera geometries (*e.g.* pushbroom cameras that do not move on a straight path) do not allow an exact rectification of the resulting images [6]. Therefore, a general definition using arbitrarily defined epipolar lines is preferred.

### 3.1. Review

The Semi-Global Matching (SGM) method [5] aims to determine the disparity image $D$, such that the global energy $E(D)$ is a minimum.

$$\begin{aligned} E(D) = \sum_{\mathbf{p}} \Big( C(\mathbf{p}, D_{\mathbf{p}}) &+ \sum_{\mathbf{q} \in N_{\mathbf{p}}} P_1 \, \mathrm{T}[|D_{\mathbf{p}} - D_{\mathbf{q}}| = 1] \\ &+ \sum_{\mathbf{q} \in N_{\mathbf{p}}} P_2 \, \mathrm{T}[|D_{\mathbf{p}} - D_{\mathbf{q}}| > 1] \Big) \end{aligned} \tag{1}$$

The first term of equation (1) calculates the sum of a pixelwise matching cost $C(\mathbf{p}, D_{\mathbf{p}})$ for all pixels $\mathbf{p}$ at their disparities $D_{\mathbf{p}}$. The cost function can either be Birchfield and Tomasi's sampling insensitive intensity difference [1] or Mutual Information [5]. The latter one has the advantage that it takes complex relations between corresponding intensities into account. This has been shown to be very robust against radiometric differences that often occur in practice [5, 6]. The function T[] is defined to return 1 if its argument is true and 0 otherwise. Thus, the second term of the energy function penalizes small disparity differences of neighboring pixels $N_{\mathbf{p}}$ of $\mathbf{p}$ with the cost $P_1$. Similarly, the third term penalizes larger disparity steps (*i.e.* discontinuities) with a higher penalty $P_2$.

The value of $P_2$ does not depend on the size of the disparity step, which preserves discontinuities. However, it has been found advantageous to adapt $P_2$ to the local intensity gradient, because discontinuities are often visible as intensity changes. Thus, the penalty should be reduced where intensities differ, which is expressed as $P_2 = \frac{P_2'}{|I_{b\mathbf{p}} - I_{b\mathbf{q}}|}$.

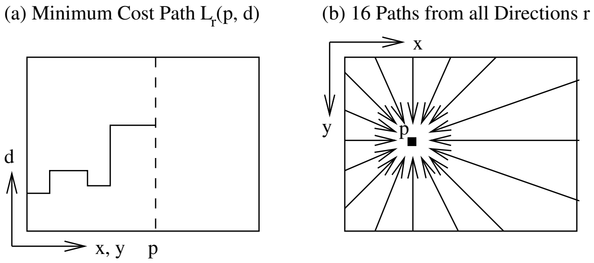

(a) Minimum Cost Path L_r(p, d)
(b) 16 Paths from all Directions r
d
x, y
p
x
y
p

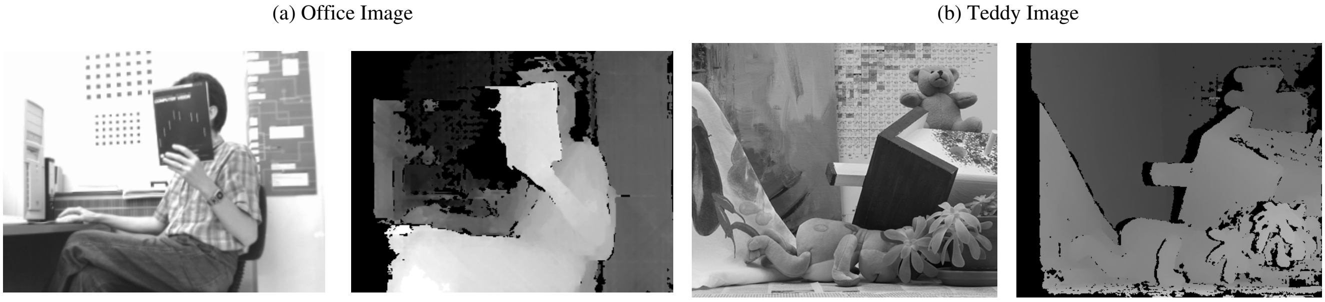
Figure 1. Aggregation of costs.

Finding the global minimum of equation (1) for the whole 2D image is known to be an NP-complete Problem. SGM calculates $E(D)$ efficiently along 1D paths from either 8 or 16 directions towards each pixel as shown in Figure 1. The cost to reach a pixel $\mathbf{p}$ at the disparity $d$ from the direction $\mathbf{r}$ is defined according to (1) recursively as,

$$\begin{aligned} L_{\mathbf{r}}(\mathbf{p}, d) = C(\mathbf{p}, d) + \min(&L_{\mathbf{r}}(\mathbf{p} - \mathbf{r}, d), \\ &L_{\mathbf{r}}(\mathbf{p} - \mathbf{r}, d-1) + P_1, L_{\mathbf{r}}(\mathbf{p} - \mathbf{r}, d+1) + P_1, \\ &\min_i L_{\mathbf{r}}(\mathbf{p} - \mathbf{r}, i) + P_2) - \min_k L_{\mathbf{r}}(\mathbf{p} - \mathbf{r}, k). \end{aligned} \tag{2}$$

The first term is the pixelwise matching cost for the current pixel. The second term adds the minimum of the cost at the previous pixel on the path, including the appropriate penalty. According to equation (1), there is no penalty added to the cost at the same disparity. The penalty $P_1$ is added to costs at the next lower or higher disparity and $P_2$ is added, if a cost at another disparity is the minimum. The last term of function (2) does not have any influence on the subsequent calculation, but it guarantees that $L \leq C_{max} + P_2$. Without this term, $L$ would always increase along each path and its value could exceed the used data type. The calculation of this function can be done in $O(ND)$ steps, where $N$ is the number of pixels along the path and $D$ is the number of disparities.

The costs along the paths from all directions $\mathbf{r}$ are summed $S(\mathbf{p}, d) = \sum_{\mathbf{r}} L_{\mathbf{r}}(\mathbf{p}, d)$. For each pixel $\mathbf{p}$ the disparity $d$ is chosen that corresponds to the minimum cost, *i.e.* $D_{\mathbf{p}} = \mathrm{argmin}_d S(\mathbf{p}, d)$. For sub-pixel estimation, a quadratic curve is fitted through the neighboring costs (*i.e.* at the next higher and lower disparity) and the position of the minimum is calculated. The result is a disparity image $D_b$ that corresponds to the base image $I_b$.



(a) Office Image

(b) Teddy Image

Figure 2. SGM results on images with untextured background areas. Black represents filtered (*i.e.* invalid) disparities.

The disparity image $D_m$ that corresponds to the match image $I_m$ can either be derived from the same cost array $S()$ or calculated from scratch. A consistency check compares the disparities of $D_b$ with $D_m$ and invalidates those that differ, *i.e.*

$$D_{\mathbf{p}} = \begin{cases} D_{b\mathbf{p}} & \text{if } |D_{b\mathbf{p}} - D_{m\mathbf{q}}| \leq 1, \mathbf{q} = e_{bm}(\mathbf{p}, D_{b\mathbf{p}}), \\ D_{inv} & \text{otherwise.} \end{cases} \qquad (3)$$

The complexity of the method is linear to the number of pixels $WH$ and the disparity range $D$, *i.e.* $O(WHD)$, if intermediate costs are reused appropriately. An efficient implementation temporarily stores the costs $S()$ of all pixels and all disparities. The regular structure of calculation and an appropriate choice of data types allows speeding up computation using Single Instruction, Multiple Data (SIMD) commands that are available in modern processing units.

Tests on several stereo images with ground truth showed that the quality of disparity images of SGM is comparable to that of global methods [8]. However, its complexity is equivalent to typical local methods and its execution speed is nearly real time with just 1s on small images [5].

SGM has been applied to the problem of fully automatically reconstructing huge urban areas (*e.g.* whole cities) from high resolution images of an airborne multi-line pushbroom camera [6]. This application benefits not only from the accuracy of SGM at sharp object boundaries (*e.g.* houses), but also from the Mutual Information based matching cost, because the used images have high radiometric differences. More than 17000 $km^2$ of aerial images in resolutions between 15-20cm/pixel have already been processed by SGM.

### 3.2. Problems in structured environments

Despite the success of SGM on aerial images, there are some problems in the presence of large, partly untextured background areas as typically found in structured environments. Figure 2 shows two examples.

Both disparity images contain well handled untextured foreground areas like the journal in Figure 2a or the chimney and walls of the house in Figure 2b. However, foreground object boundaries are blurred into untextured background areas, as seen at the head in Figure 2a and the background on the right of the Teddy in Figure 2b. In contrast, object borders in front of textured background appear correctly. It is interesting to note that these kinds of problems do not occur in the application of aerial imaging. This is probably because at the used level of resolution, there are no really untextured areas in aerial images, at least not behind foreground objects.

SGM handles untextured areas by the smoothness term in equation (1), which penalizes the change of disparity by $P$. Disparity changes are accepted, if the sum of pixelwise matching cost $C(\mathbf{p}, d_i)$ at another disparity $d_i$ and the penalty $P$ is lower than the cost $C(\mathbf{p}, d)$. The reason for the change of pixelwise matching cost of nearby pixels is texture. Thus, untextured areas are interpolated smoothly, by using the support of neighboring, better textured areas.

A problem arises if foreground objects are in front of a partly untextured background. In this case, the required step from foreground to background disparity can be placed anywhere next to or within the untextured area without changing the global energy (1). Thus, the placement of the disparity step mainly depends on noise. The correct placement can be supported by making the penalty $P_2$ adaptive to intensity changes, such that disparity changes within untextured areas are more costly [5]. Thus, placing the disparity change at the border of the untextured area is preferred.

However, the adaptive penalty has already been used for calculating the disparity images of Figure 2. The reason for the still seen misplaced object borders is that SGM propagates the costs along straight paths. In Figure 2b, there are no straight paths leading from the textured background to the place between the arm and leg on the right side of the Teddy. Thus, the algorithm has at this place no information that the disparity should not be continued smoothly between the arm and the leg. Equally difficult is the background around the head in Figure 2a. Additionally, the number of paths that meet in a point and the magnitude with which they suggest a certain disparity influences the correct placement of a disparity step. The adaptive penalty can only work correctly, if the cost is gathered from all directions equally well, which is usually the case with small, compact untextured areas.

Additionally, it is undesirable for some applications like reconstruction to have invalid disparities (*e.g.* black areas in Figure 2), as caused by the left/right consistency check of SGM. These areas need to be interpolated, without smoothing discontinuities of the disparity image.

## 4. Proposed refinements of SGM

The problems at partly untextured background areas and the interpolation of invalid disparities have to be solved for applying SGM successfully on images of structured environments. Furthermore, it is important that the solutions must not increase the complexity of SGM, as efficiency is one of the major advantages of SGM over methods with comparable quality.

### 4.1. Intensity consistent disparity selection

The solution for avoiding the blurring of foreground objects into partly untextured background requires some assumptions:

1. Discontinuities in the disparity image do not occur within untextured areas.
2. On the same physical surface as the untextured area is also some texture visible.
3. The surface of the untextured area can be approximated by a plane.

The first assumption is mostly correct, as depth discontinuities usually cause at least some visual changes in intensities. Otherwise, the discontinuity would be undetectable. The second assumption states that there are at least some points on an untextured surface for which the disparity can be determined. The disparity of an absolutely untextured background surface would be indeterminable. The third assumption is clearly the weakest. Its justification is that identifying untextured areas as areas of nearly constant intensity will result in patches that can be treated as planar. Untextured surfaces with varying distance usually appear with varying intensities. Piecewise constant intensity can be treated as piecewise planar.

The identification of untextured areas is done by a fixed bandwidth Mean Shift Segmentation [3] on the intensity image $I_b$. The radiometric bandwidth $\sigma_r$ is set to $P_1$, which is usually 4. Thus, intensity changes below the smoothness penalty are treated as noise. The spatial bandwidth $\sigma_s$ is set to a rather low value for fast processing (*i.e.* 5). Furthermore, all segments that are smaller than a certain threshold (*i.e.* 100 pixels) are ignored, because small untextured areas are usually handled well by SGM as discussed in Section 3.2. The segmentation result of the Teddy image is shown in Figure 3a.

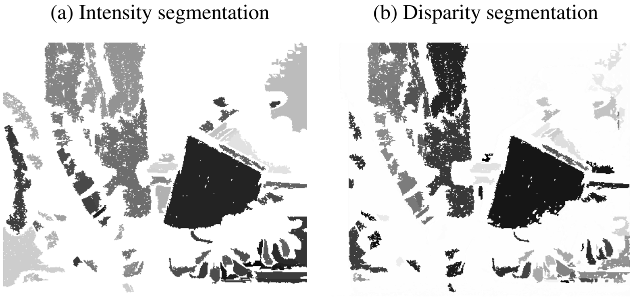

(a) Intensity segmentation (b) Disparity segmentation

Figure 3. Intensity and subsequent disparity segmentation.

A feature of SGM is the propagation of disparities from textured into untextured areas (Section 3.2). This feature, together with the assumption that some neighboring textured areas are on the same surface (*i.e.* assumption 2), lead to the realization that some disparities within each segment $S_i$ should be correct. Furthermore, using penalties $P_2$ that do not depend on the size of the disparity change prefers sudden disparity changes rather than smooth ones. Thus, several hypotheses for the correct disparity of $S_i$ can be identified by segmenting the disparity within each segment $S_i$. This is done by simple segmentation, which connects neighboring equivalent pixels within the 4-connected image grid [4]. Equivalent pixels differ by at most 1 disparity. This fast segmentation results in several segments $S_{ik}$ for each segment $S_i$ as shown in Figure 3b.

Next, the surface hypotheses $F_{ik}$ are created by calculating the best fitting planes through the disparities of $S_{ik}$. The choice for planes is based on assumption 3. Very small segments (*i.e.* $\leq 12$ pixel) are ignored, as it is unlikely that such small patches belong to the correct hypothesis. Then, each hypothesis is evaluated within the area of $S_i$ by,

$$E_{ik}(D') = \sum_{\mathbf{p} \in S_i \setminus occ} (C(\mathbf{p}, D'_\mathbf{p}) + \sum_{\mathbf{q} \in N_\mathbf{p}} P_1 \mathrm{T}[|D'_\mathbf{p} - D'_\mathbf{q}| = 1] + \sum_{\mathbf{q} \in N_\mathbf{p}} P_2 \mathrm{T}[|D'_\mathbf{p} - D'_\mathbf{q}| > 1]) \tag{4}$$

$$D'_\mathbf{p} = \begin{cases} F_{ik}(\mathbf{p}) & \text{if } \mathbf{p} \in S_i \\ D_\mathbf{p} & \text{otherwise.} \end{cases} \tag{5}$$

Thus, all disparities within segment $S_i$ are replaced by the surface hypothesis and the cost $E_{ik}$ calculated for all pixels within $S_i$. The main difference to equation (1) is that pixelwise matching costs are not considered at occlusions. A pixel $\mathbf{p}$ is occluded, if another pixel with higher disparity maps to the same pixel $\mathbf{q}$ in the match image. This detection is performed by first mapping $\mathbf{p}$ into the match image by $\mathbf{q} = e_{bm}(\mathbf{p}, D'_{\mathbf{p}})$. Then, the epipolar line of $\mathbf{q}$ in the base image $e_{mb}(\mathbf{q}, d)$ is followed for $d > D'_{\mathbf{p}}$. Pixel $\mathbf{p}$ is occluded if there is an intersection of the epipolar line with the disparity surface defined by $D'_{\mathbf{p}}$.

For each constant intensity segment $S_i$ the surface hypothesis $F_{ik}$ with the minimum cost $E_{ik}$ is chosen. All disparities within $S_i$ are replaced by values on the chosen surface for making the disparity selection consistent to the intensities of the base image (i.e. fulfilling assumption 1).

$$F_i = F_{ik'} \text{ with } k' = \operatorname*{argmin}_k E_{ik} \tag{6}$$

$$D'_{\mathbf{p}} = \begin{cases} F_i(\mathbf{p}) & \text{if } \mathbf{p} \in S_i \\ D_{\mathbf{p}} & \text{otherwise.} \end{cases} \tag{7}$$

The complexity of fixed bandwidth Mean Shift Segmentation of the intensity image and the simple segmentation of the disparity image is linear in the number of pixels. Calculating the best fitting planes involves visiting all segmented pixels. Testing of all hypotheses requires visiting all pixels of all segments, for all hypotheses (*i.e.* maximum $N$). Additionally, the occlusion test requires going through at most $D$ disparities for each pixel.

Thus, the upper bound of the complexity is $O(WHDN)$. However, segmented pixels are usually just a fraction of the whole image and the maximum number of hypotheses $N$ for a segment is commonly small and often just 1. In the latter case, it is not even necessary to calculate the cost of the hypothesis.

The presented approach is similar to some other methods [2, 13, 14] as it uses image segmentation and plane fitting for refining an initial disparity image. In contrast to other methods, the initial disparity image is due to SGM already quite accurate so that only untextured areas above a certain size are modified. Thus, only critical areas are tackled without the danger of corrupting probably well matched areas. Another difference to other methods is that disparities of the considered areas are selected in one step by considering a number of hypotheses that are inherent in the initial disparity image. There is no time consuming iteration.

## 4.2. Discontinuity preserving interpolation

The left/right consistency check invalidates disparities due to occlusions (*e.g.* $\mathbf{p}_1$ in Figure 4), but also due to other kinds of mismatches (*e.g.* $\mathbf{p}_2$ in Figure 4). For interpolating invalid disparities, both cases need to be treated differently. Occlusions must not be interpolated from the occluder, but only from the occludee to avoid incorrect smoothing of discontinuities. Thus, an extrapolation of the background into occluded regions is necessary. In contrast, holes due to mismatches can be smoothly interpolated from all neighboring pixels.

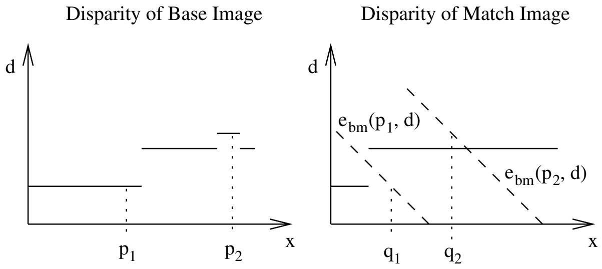


Figure 4. Disparity of the base and match image.

Occlusions and mismatches can be distinguished as part of the left/right consistency check. Figure 4 shows that the epipolar line of the occluded pixel $\mathbf{p}_1$ goes through the discontinuity that causes the occlusion and does not intersect the disparity function $D_m$. In contrast, the epipolar line of the mismatch $\mathbf{p}_2$ intersects with $D_m$. Thus, for each invalidated pixel, an intersection of the corresponding epipolar line with $D_m$ is sought, for marking it as either occluded or mismatched.

Additionally to the consistency check, a segmentation filter may be used that invalidates small disparity segments (*e.g.* 20 pixel), because they are mostly due to errors. The filtered disparities are also marked as mismatches. Figure 5 shows the occlusions and mismatches of the Teddy images that were identified by the consistency check.

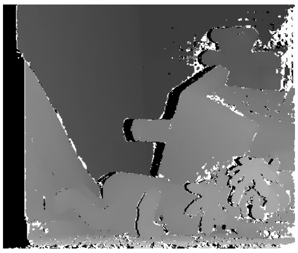
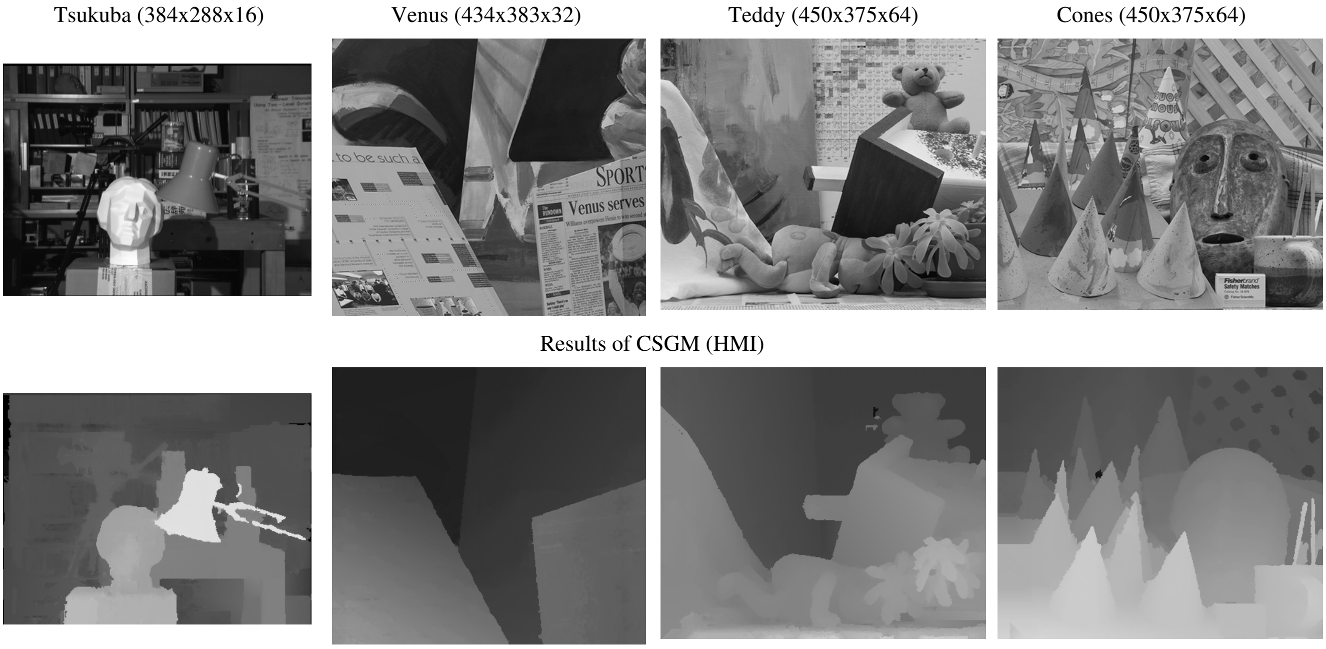


Figure 5. Occluded pixel (black) and mismatches (white).

For interpolation purposes, mismatched pixel areas that are direct neighbors of occluded pixels are treated as occlusions, because these pixels must also be extrapolated from valid background pixels. Interpolation is done by propagating valid disparities through neighboring invalid disparity areas. This is done similarly to SGM along paths from 8 directions. For each invalid pixel, all 8 values $v_{\mathbf{p}i}$ are stored. The final disparity image $D'$ is created by,

$$D'_{\mathbf{p}} = \begin{cases} \text{seclow}_i\, v_{\mathbf{p}i} & \text{if } \mathbf{p} \text{ is occluded,} \\ \text{median}_i\, v_{\mathbf{p}i} & \text{if } \mathbf{p} \text{ is mismatched,} \\ D_{\mathbf{p}} & \text{otherwise.} \end{cases} \tag{8}$$

The first case ensures that occlusions are interpolated from the lower background by selecting the second lowest


Tsukuba (384x288x16)

Venus (434x383x32)

Teddy (450x375x64)

Cones (450x375x64)

Results of CSGM (HMI)

Figure 6. Results of the new method (CSGM) on stereo images with ground truth [9, 10].

| Algorithm | Rank | Tsukuba | | | Venus | | | Teddy | | | Cones | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | nonoc | all | disc | nonoc | all | disc | nonoc | all | disc | nonoc | all | disc |
| AdaptingBP | 1.7 | 1.11 | 1.37 | 5.79 | **0.10** | **0.21** | **1.44** | 4.22 | 7.06 | 11.80 | **2.48** | **7.92** | **7.32** |
| DoubleBP [13] | 2.3 | **0.88** | **1.29** | **4.76** | 0.14 | 0.60 | 2.00 | **3.55** | 8.71 | **9.70** | 2.90 | 9.24 | 7.80 |
| SymBP+occ [11] | 4.9 | 0.97 | 1.75 | 5.09 | 0.16 | 0.33 | 2.19 | 6.47 | 10.70 | 17.00 | 4.79 | 10.70 | 10.90 |
| Segm+visib [2] | 5.0 | 1.30 | 1.57 | 6.92 | 0.79 | 1.06 | 6.76 | 5.00 | **6.54** | 12.30 | 3.72 | 8.62 | 10.20 |
| **CSGM** | 5.9 | 2.61 | 3.29 | 9.89 | 0.25 | 0.57 | 3.24 | 5.14 | 11.80 | 13.00 | 2.77 | 8.35 | 8.20 |
| RegionTreeDP | 6.6 | 1.39 | 1.64 | 6.85 | 0.22 | 0.57 | 1.93 | 7.42 | 11.90 | 16.80 | 6.31 | 11.90 | 11.80 |
| AdaptWeight | 7.1 | 1.38 | 1.85 | 6.90 | 0.71 | 1.19 | 6.13 | 7.88 | 13.30 | 18.60 | 3.97 | 9.79 | 8.26 |
| **SemiGlob** [5] | 8.8 | 3.26 | 3.96 | 12.80 | 1.00 | 1.57 | 11.30 | 6.02 | 12.20 | 16.30 | 3.06 | 9.75 | 8.90 |
| GC+occ [7] | 10.4 | 1.19 | 2.01 | 6.24 | 1.64 | 2.19 | 6.75 | 11.20 | 17.40 | 19.80 | 5.36 | 12.40 | 13.00 |
| Layered [14] | 10.5 | 1.57 | 1.87 | 8.28 | 1.34 | 1.85 | 6.85 | 8.64 | 14.30 | 18.50 | 6.59 | 14.70 | 14.40 |
| Currently 12 more entries ... | | | | | | | | | | | | | |

Table 1. Table from Middlebury Stereo Evaluation [8] using a maximum disparity difference threshold of 1.

value, while the second case emphasizes the use of all information. The median is used instead of the mean for maintaining discontinuities in cases where the mismatched area is at an object border. It can be seen in Figure 5 that there are a few mismatched areas at the right border of objects (*i.e.* un-occluded areas). In contrast to occlusions, there is no preference to either a lower or higher disparity.

The presented interpolation method has the advantage that it is independent of the used stereo matching method. The only requirements are a known epipolar geometry and the calculation of the disparity images for the base and match image for distinguishing between occlusions and mismatches.

The complexity of interpolation is linear to the number of pixels, *i.e.* $O(WH)$, as there is a constant number of operations for each invalid pixel.

## 5. Experimental results

The proposed Consistent Semi-Global Matching (CSGM) method has been tested on stereo images with ground truth as well as on stereo images of structured environments.

### 5.1. Stereo images with ground truth

Figure 6 shows four stereo images with ground truth [9, 10] on which many recent stereo algorithms have been tested. The disparity images have been calculated with constant parameters of CSGM on all images. Hierarchical Mutual Information (HMI) has been chosen as matching cost, which has its main benefit on the Cones images [5]. It can be seen that the critical untextured area on the right of the Teddy as well as all other untextured areas are handled well.

Furthermore, all discontinuities appear sharp and small details of objects are maintained.

The results have been submitted to the Middlebury Stereo Evaluation [8] and compared to other stereo methods. The evaluation compares the disparity images with the corresponding ground truth individually at non-occluded areas, at all pixels and near discontinuities. Pixels, where the disparity differs by more than 1 from the ground truth are counted as errors. Table 1 shows the errors in percent of the corresponding area for the currently top-ranked algorithms.

It can be seen that the proposed refinements of SGM reduce the errors compared to SGM (i.e. referred to as SemiGlob in Table 1). The new method is the second best on the Cone images and performs quite well on Venus and Teddy. It does not perform as well as some other methods on Tsukuba, which results in the 5th place of currently 22 algorithms. However, reducing the maximum disparity differences between calculated disparities and the ground truth to 0.75 or 0.5 rises CSGM to the best performing algorithm on these test images. The reason is probably the lack or bad performance of sub-pixel disparity estimation of other algorithms. This demonstrates the the accuracy of CSGM.

The execution time of SGM increases by the proposed extensions by about 30-50% on the test images of Figure 6. Most of the time is consumed by Mean Shift Segmentation. The total execution time on the Teddy or Cones images with 64 pixels disparity range is just a few seconds on a 2.8GHz PC. This is much lower than the execution time of the most other top-ranked algorithms.

## 5.2. Stereo images of structured scenes

Figure 7 shows a few examples of structured environments, which were taken from a stereo sequence of a walk through an indoor environment. The disparity images without the proposed extensions (i.e. SGM) are shown in the second row. It can be seen that object borders in front of untextured areas appear fuzzy. The third row shows the disparity images with the proposed intensity consistent disparity selection. Object borders in front of untextured areas appear much cleaner and more correct, especially in the office image. Black areas represent unknown disparities, due to occlusions or filtered mismatches. The result of interpolating these areas is shown in the last row. Object borders are maintained during interpolation. Fine structures like at the Chairs or the small objects in the Kitchen appear well maintained.

## 5.3. Limitations of the method

Despite the obvious improvements due to the proposed extensions of SGM, there are cases in which they can fail. The circle in disparity image of the Office scene in Figure 7 marks a place where a part of the poster at a column in front of the wall is wrongly labeled with the same depth as the wall. This happens, because there is no visual change between the background wall and the white part of the poster (violation of the first assumption in Section 4.1). Thus, the method tries to find a common plane for the background *and* part of the foreground.

The circles in the disparity images of the other scenes mark places where the disparities of an untextured area are wrong. All of these errors are near image borders. At these places, the chances that SGM has propagated correct disparities into the untextured area is reduced as only a part of the untextured area and not the complete borders of it are seen. For the same reason, testing of different hypothesis is more error prone. Due to these reasons, it may be better to set untextured areas at image borders to invalid (i.e. unknown).

## 6. Conclusion

It has been shown that the proposed intensity consistent selection of disparities as well as the discontinuity preserving interpolation of disparities improve the performance of SGM especially in structured environments.

The new CSGM method performs accurate matching and produces sharp object boundaries, even in the presence untextured background areas. It can handle complex shapes and fine structures in the presence of texture and falls back to a planar model at untextured areas only. Possible radiometric differences are handled robustly by Mutual Information as in SGM. Left/right consistency checking and subpixel estimation is performed. Furthermore, invalid disparities are interpolated in a discontinuity preserving way.

A comparison has shown that CSGM can compete with the currently top-ranked stereo algorithms, but at a much lower complexity and runtime. These features make CSGM a very valuable tool for many practical situations.

## References

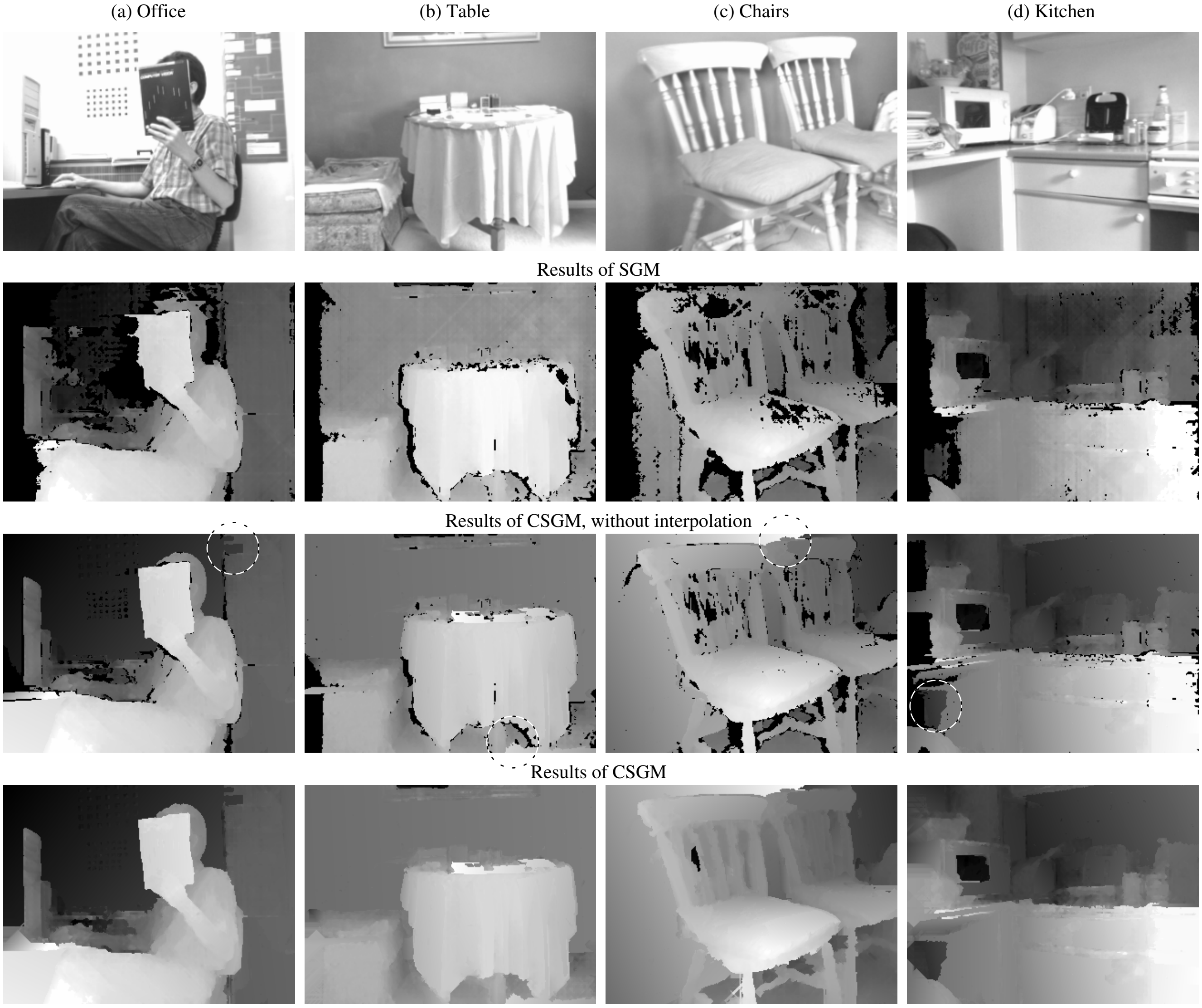
[1] S. Birchfield and C. Tomasi. Depth discontinuities by pixel-to-pixel stereo. In *Proceedings of the Sixth IEEE International Conference on Computer Vision*, pages 1073–1080, Mumbai, India, January 1998.

[2] M. Bleyer and M. Gelautz. A layered stereo matching algorithm using image segmentation and global visibility constraints. *ISPRS Journal of Photogrammetry and Remote Sensing*, 59(3):128–150, 2005.

[3] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):1–18, May 2002.

[4] E. R. Davis. *Machine Vision: Theory, Algorithms, Practicalities*. Academic Press, 2nd edition, 1997.

[5] H. Hirschmüller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 807–814, San Diego, CA, USA, June 2005.



(a) Office (b) Table (c) Chairs (d) Kitchen

Results of SGM

Results of CSGM, without interpolation

Results of CSGM

Figure 7. Examples of SGM and CSGM on images of structured environments.

[6] H. Hirschmüller, F. Scholten, and G. Hirzinger. Stereo vision based reconstruction of huge urban areas from an airborne pushbroom camera (hrsc). In *Proceedings of the 27th DAGM Symposium*, volume LNCS 3663, pages 58–66, Vienna, Austria, August/September 2005. Springer.

[7] V. Kolmogorov and R. Zabih. Computing visual correspondence with occlusions using graph cuts. In *International Conference for Computer Vision*, volume 2, pages 508–515, 2001.

[8] D. Scharstein and R. Szeliski. Middlebury online stereo evaluation. www.middlebury.edu/stereo.

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

[10] D. Scharstein and R. Szeliski. High-accuracy stereo depth maps using structured light. In *IEEE Conference for Computer Vision and Pattern Recognition*, volume 1, pages 195–202, Madison, Winsconsin, USA, June 2003.

[11] J. Sun, Y. Li, S. Kang, and H.-Y. Shum. Symmetric stereo matching for occlusion handling. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 399–406, San Diego, CA, USA, June 2005.

[12] P. Viola and W. M. Wells. Alignment by maximization of mutual information. *International Journal of Computer Vision*, 24(2):137–154, 1997.

[13] Q. Yang, L. Wang, R. Yang, H. Stewenius, and D. Nister. Stereo matching with color-weighted correlation, hirarchical belief propagation and occlusion handling. In *IEEE Conference on Computer Vision and Pattern Recognition*, New York, NY, USA, 17-22 June 2006.

[14] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski. High-quality video view interpolation using a layered representation. In *SIGGRAPH*, 2004.