Photogrammetrie • Fernerkundung • Geoinformation 2/2006, S. 139–151, 11 figs., 4 tabs.

Cooperative Disparity Estimation and its Improvement

HELMUT MAYER, Neubiberg

Keywords: mathematics, photogrammetry, stereo imagery, cooperative disparity estimation, evaluation, visualization

Summary: ZITNICK & KANADE (2000) proposed a cooperative approach for disparity estimation from stereo imagery based on support and inhibition in three-dimensional (3D) disparity space. We describe this approach and show how a significant improvement over the results of the evaluation reported for (ZITNICK & KANADE 2000) in (SCHARSTEIN & SZELISKI 2002) can be obtained by several means. The results of the evaluation of our approach are in the upper third in the online version of (SCHARSTEIN & SZELISKI 2002), while the numerical complexity compares favorably with other approaches. We have analyzed the different means for improvement including their performance gain. They comprise symmetric support, the combination of absolute differences and (normalized) cross-correlation weighted by the strength of the horizontal gradient, the use of auto-correlation to estimate the significance of a matching score, the preference for small disparities to obtain more meaningful results for occluded regions, and the enforcement of the alignment of disparity and image gradient. The parameters of these means are tuned with respect to given data sets. However, results using the same set of parameters with other images confirm, that our implementation is applicable to a wide range of imagery.

Zusammenfassung: *Kooperative Disparitätsschätzung und ihre Verbesserung*. ZITNICK & KANADE (2000) schlugen einen kooperativen Ansatz für die Disparitätsschätzung aus Stereobildern basierend auf Unterstützung und Hemmung im dreidimensionalen (3D) Disparitätsraum vor. Dieser Ansatz wird hier beschrieben und es wird gezeigt, wie durch verschiedene Maßnahmen eine signifikante Verbesserung gegenüber den Ergebnissen der Evaluierung erzielt werden kann, die für (ZITNICK & KANADE 2000) in (SCHARSTEIN & SZELISKI 2002) dargestellt wurden. Die Ergebnisse der Evaluierung des verbesserten Ansatzes befinden sich im oberen Drittel in der online-Version von (SCHARSTEIN & SZELISKI 2002), wohingegen die numerische Komplexität im Vergleich zu anderen Ansätzen vorteilhaft ist. Die verschiedenen Maßnahmen wurden mit besonderem Schwerpunkt auf der Verbesserung der Leistungsfähigkeit analysiert. Sie umfassen symmetrische Unterstützung, die Kombination von absoluten Differenzen und (normalisiertem) Kreuzkorrelationskoeffizienten gewichtet mit der Stärke des horizontalen Gradienten, die Nutzung von Autokorrelation, um die Signifikanz einer Zuordnungsstärke zu schätzen, die Bevorzugung von kleinen Disparitäten, um für verdeckte Regionen sinnvollere Ergebnisse zu erzielen, und zuletzt die Forcierung der gleichen Lage von Disparitäts- und Bildgradienten. Die Parameterwerte der Maßnahmen wurden auf gegebene Datensätze abgestimmt. Ergebnisse mit anderen Bildern, die denselben Satz von Parameterwerten verwenden, zeigen jedoch, dass die vorliegende Implementierung für eine große Bandbreite von Bildern geeignet ist.

1432-8364/06/2006/0139 $ 3.25

1 Introduction

Image pairs are often ''normalized'' before dense matching by means of epipolar resampling (McGlone et al. 2004). Epipolar lines become the rows of the resampled image, meaning that only horizontal x-parallaxes termed ''disparities'' remain. In combination with the orientation and calibration of the cameras they alone determine the distances of the corresponding points from the cameras. The elimination of y-parallaxes allows for simplified matching schemes working on corresponding image rows only.

The estimation of disparities from stereo pairs has received considerable attention over the last decades. Yet, there is still not one or even a set of 'gold standard' approaches which can deal with a broad range of imagery. An excellent recent survey (Scharstein & Szeliski 2002) has grouped existing approaches into a taxonomy. The assumption is, that matching algorithms perform all or several of the following four steps:

1. matching cost computation,
2. cost (support) aggregation,
3. disparity computation/optimization,
4. disparity refinement.

Typical matching costs are squared differences or absolute differences of gray values. Aggregation can be done as simple as summing up over a square window and disparity computation by determining the minimum matching cost at a position. For the (normalized) cross-correlation-coefficient, steps 1 and 2 are combined. Finally, global algorithms are based on explicit smoothness assumptions for the object surface which are solved by means of optimization. As the number of approaches is vast, we refer for the literature to (Scharstein & Szeliski 2002), only introducing recent approaches and comparing them to our ideas and results.

Scharstein & Szeliski (2002) have also introduced an evaluation metric as well as test data to compare different approaches. Together with a web page (http://www.middlebury.edu/stereo/) listing the results of all approaches for which results have been submitted under the constraint, that the same set of parameters have been used for all image pairs, and an ordering according to the performance, this has sparked competition and progress in the field.

Zitnick & Kanade 2000, that our work is based on, refers to work proposed at the end of the seventies (Marr & Poggio 1976, 1979). The basic idea is to employ explicitly stated global constraints on uniqueness and continuity of the disparities. While Marr & Poggio (1976), (1979) have used two-dimensional (2D) regions to enforce continuity by fusing support among disparity estimates, Zitnick & Kanade (2000) employ 3D support regions. Matching scores are calculated for a disparity range (search width) and then stored in a 3D array. This array is filtered with a 3D box-filter to obtain the local support for a match from all close-by matches. Assuming opaque, diffuse-reflec-ting surfaces, the uniqueness constraint requires that on one ray of view only one point is visible. This implies an inhibition which is realized by weighting down all scores besides the strongest. Support and inhibition are iterated. Thereby, information is propagated more globally. We have chosen (Zitnick & Kanade 2000) because it can deal with strong occlusions and large disparity ranges and have extended it by the following means:

The smoothness of the output is improved by sub-pixel estimation. By a recursive implementation of the 3D box-filter we have sped up the computation. We determine the convergence automatically and employ symmetric support, considerably improving the results. As proposed by Scharstein & Szeliski (2002), we combine for the matching scores cross-correlation with absolute differences, employing correlation particularly for horizontally textured regions. As we are looking for unambiguous matches, the matching scores are weighted down when there is repetitive texture determined by a special type of auto-correlation. It was found that, using color improves the result. As occluded regions have a smaller disparity than their occluding regions, we have introduced a small preference for smaller dispar-

ities. By combining image gradient and disparity gradient to control the amount of smoothing as proposed by ZHANG & KAMBHAMETTU (2002), we avoid blurring disparity discontinuities and the elimination of narrow linear structures. Finally, determining occlusions and reducing the probabilities for large disparities in these regions is another means to obtain more meaningful, smaller disparities in occluded regions.

The paper which is an extension of MAYER (2003) is organized as follows. First we give a short account of cooperative disparity estimation as proposed by ZITNICK & KANADE (2000). Section presents the evaluation metric of SCHARSTEIN & SZELISKI (2002) and our results for the four image pairs obtained using one set of parameters. In Section 4 we present the means for improvement in more detail and we analyze them by assessing their performance gain. In Section 5 additional results are presented. The paper ends up with conclusions.

2 Cooperative Disparity Estimation

The main idea of ZITNICK & KANADE (2000) is a cooperation between support and inhibition (cf. Fig.). The support region is a 2D-region or usually a 3D-box. All matching scores in this box, derived, e. g., by (normalized) cross-correlation, corroborate to generate a disparity map which is locally continuous. When employing a 3D-box, also sloped regions are modeled, although only implicitly.

Inhibition enforces the uniqueness of a match. Assuming opaque and diffuse-reflecting surfaces, a ray of view emanating from a camera will hit the scene only at one point. The idea is to gradually weight down all matches on a ray of view besides the strongest. For a stereo pair there are two rays (cf. Fig., right). The matching scores are stored in a 3D array. Therefore, for the left image the ray of view is a column in the 2D-slice of width and disparity. Because we work in disparity and not in depth space, the ray of view of the right image consists of the 45° left-slanted diagonal through the pixel of interest. Putting everything together, the support S_n for a pixel at row r and column c with disparity d is defined as

$$S_n(r,c,d) = \sum_{(r',c',d')\in\Phi} L_n(r+r', c+c', d+d'), \tag{1}$$

with L_n the score for the preceding iteration and Φ the support region. The new score for iteration $n+1$ is obtained as

$$L_{n+1}(r,c,d) = \left(\frac{S_n(r,c,d)}{\sum_{(r'',c'',d'')\in\Psi} S_n(r'',c'',d'')} \right)^{\alpha} * L_0(r,c,d) \tag{2}$$

with Ψ the union of the left and right inhibition region and α an exponent controling the speed of convergence. α has to be chosen greater than 1 to make the scores converge to 1. The multiplication with the original matching score L_0 avoids hallucination in

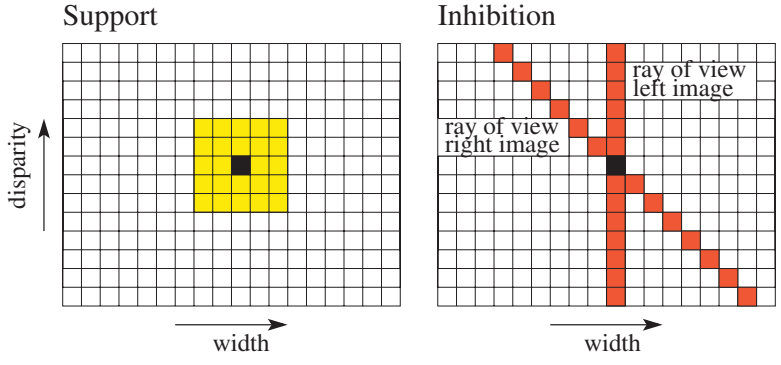

Fig. 1: Support and inhibition.

weak matching regions. Finally, for each pixel of the left image the disparity is chosen which has the maximum score. Practically, it is important to correct the inhibition value for the fact, that on the left and the right side of the image a number of pixels depending on the search width are not matched and, therefore, do not contribute to the inhibition.

3 Evaluation

For the evaluation we used the data and the code available at www.middlebury.edu/stereo (cf. Fig. 2) employing the search widths given there. The measures used in (Scharstein & Szeliski 2002) and here comprise the number of bad pixels, i. e., pixels which are further away from the given ground truth map than a tolerance δ_d. As in (Scharstein & Szeliski 2002), we also use $\delta_d = 1{,}0$ and the following measures:

- Bad pixels nonocc (all) – $B_{\overline{O}}$: % bad pixels in non-occluded regions. Used as overall performance measure.
- Bad pixels untextured (untex.) – $B_{\overline{T}}$: % bad pixels in untextured regions.
- Bad pixels discont (disc.) – B_D: % bad pixels near discontinuities.

For sub-pixel estimation a parabola involving the matching scores of the voxels having a smaller (d-1) and larger ($d + 1$) disparity than the given disparity for a pixel d (with $l_n(d) = L_n(r,c,d)$ is used ($-0{,}5 \le \Delta d \le 0.5$):

$$\Delta d = \frac{l_n(d+1) - l_n(d-1)}{2(2l_n(d) - l_n(d+1) - l_n(d-1))} \qquad (3)$$

The results presented in Figs 3 and 4, for Tsukuba and Map also compared to their ground-truth, give an indication of the quality obtained. Tab. 1 gives evaluation results which are in the upper third of the approaches presented in the online version of (Scharstein & Szeliski 2002) at www.middlebury.edu/stereo. As required there, only one set of parameters given in Tab. 3 was used. We were ranked number three in the individual result page of the online version of (Scharstein & Szeliski 2002) as of April 3, 2003 and are still number fourteen of forty as of August 21, 2005. Run time for all images is about 102 seconds on a 2.5 GHz PC. This time is better than those reported for the seventh (2706 seconds) and fifth (528 seconds) performing algorithms in (Scharstein & Szeliski 2002). The times for Tsukuba, Sawtooth, Venus, and Map are 22, 28, 36, and 16 seconds, respectively.

Tab. 1: Percentage of bad pixels with all means for improvement included (right three columns: sub-pixel precise results).

				subpixel		
	all	untex.	disc.	all	untex.	disc.
Tsukuba	1.67	0.77	9.67	2.24	1.58	11.70
Sawtooth	1.21	0.17	6.90	0.72	0.03	6.82
Venus	1.04	1.07	13.68	0.78	0.68	10.66
Map	0.29	0.00	3.65	0.24	0.00	3.36

Tab. 2: RMS error with all means for improvement included (right three columns: sub-pixel precise results).

				subpixel		
	all	untex.	disc.	all	untex.	disc.
Tsukuba	0.83	0.63	1.74	0.87	0.56	1.90
Sawtooth	0.61	0.31	1.70	0.56	0.24	1.67
Venus	0.47	0.44	1.31	0.38	0.35	1.27
Map	0.99	0.42	3.44	0.94	0.26	3.36

Tab. 3: Parameters for the results in the figures and tables of this paper.

Size matching	$5 \times 5 \times 1$
Size support	$11 \times 11 \times 3$
Truncation Value	4 gray values
Threshold for convergence	0.005 * *search_width*
Threshold for mixing scores	45 gray values
Preference for larger disparities	0.05 * *search_width*
Number iterations for occlusion	2

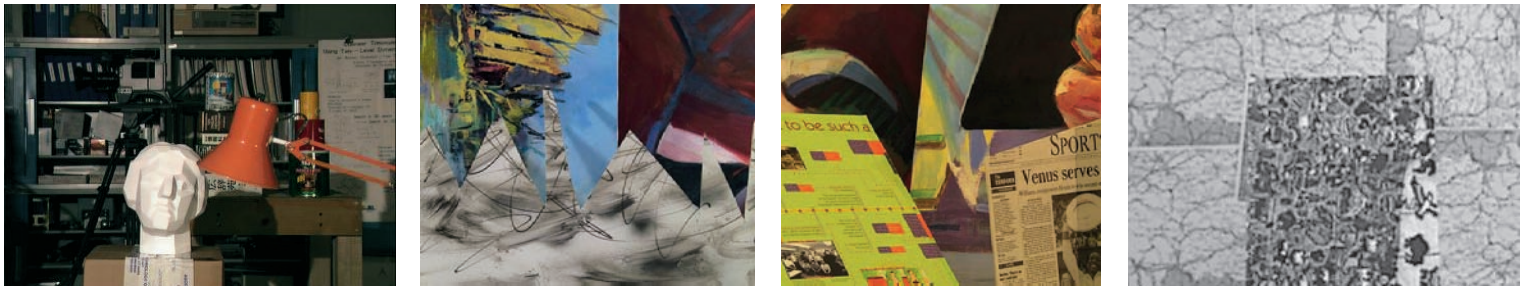

Fig. 2: Images (www.middlebury.edu/stereo) – from left to right: Tsukuba, Sawtooth, Venus, and Map.

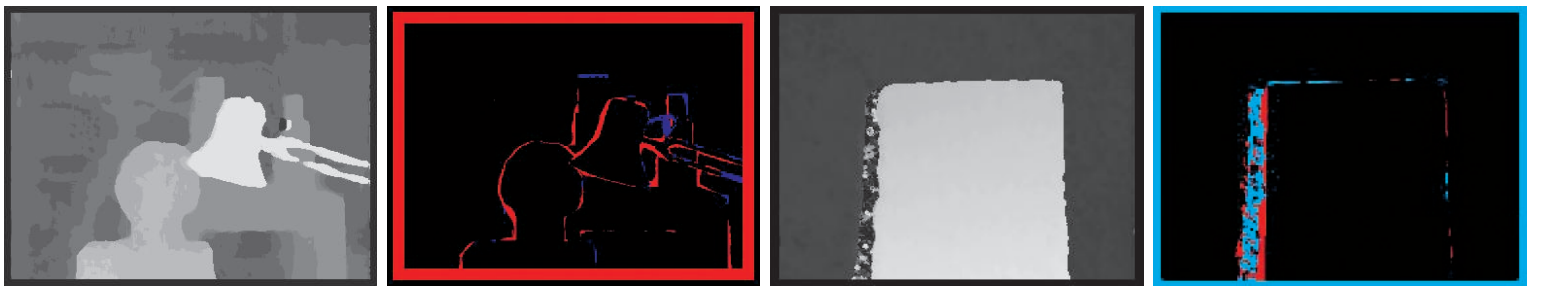

Fig. 3: Disparities and differences to ground truth for Tsukuba (left) and Map (right) – red: disparities more than 1 pixel too large; blue: disparities more than 1 pixel too small.

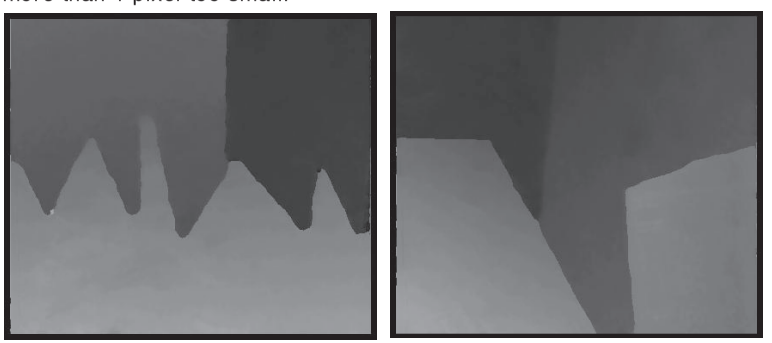

Fig. 4: Disparities for Sawtooth (left) and Venus (right) including sub-pixel estimation.

For the interpretation of the results with sub-pixel estimation, which were obtained with the same set of parameters as above, one has to consider, that while the ground-truth for Tsukuba is pixel precise, the ground-truth for the rest is sub-pixel precise. As the distance for the evaluation is fixed to one pixel and as we restrict $|\Delta d| \leq 0{,}5$, for Tsukuba sub-pixel estimation can only result into an equal or lower performance. For the other three images the performance can improve, as it does. The same is also true for the root mean square (RMS) error given in Tab. 2. Because the result for Tsukuba can only degrade for sub-pixel precise estimation, we concentrate on pixel precise disparity estimation for the rest of the paper.

4 Means for Improvement

In the remainder of the paper, we illustrate our means and show their performance gain. The two basic means presented in the first subsection only speed up the processing. The means are explained using Tsukuba as running example. Their gain is assessed in the final subsection by comparing the evaluation results when excluding the respective means from the processing to the result when all means are used.

4.1 Recursive 3D Box-filter and Convergence Determination

Filtering with a 3D box-filter based on simple summation is highly redundant. To get rid of it, we use a standard recursive filter. We separate the filter into one-dimensional (1D) staffs and 2D sheets. By adding pixels on top of each other we generate staffs (cf. Fig.). From them we build sheets and finally from the sheets the box. The update is done recursively. To filter with a translated box, instead of adding sheets we add a (new) sheet on one side and subtract the (old) sheet on the other side. The same is done for the sheets and the staffs. By this means the complexity becomes independent of the size of the box.

The performance gain depends on the size of the 3D-box, but is considerably large for meaningful box sizes. For Tsukuba of size 384x288 pixels and a disparity range, i. e., search width of 15 pixels, one iteration of the simple algorithm takes on a 2.5 GHz PC 3.2 seconds for the 11 x 11 x 3 box. The separated algorithm needs only 0.10 seconds. It is interesting to compare this with the times for the inhibition. For Tsukuba inhibition takes 0.30 seconds per iteration. If one substitutes the square, i. e., $\alpha = 2$, for the general exponential, it reduces to 0.13 seconds. Because we found that this gives also the best results in nearly all cases, we have used $\alpha = 2$ in our experiments.

The meaningful number of iterations varies for different images. It proved useful to decide about the number of iterations by convergence determination. For the latter also one parameter is needed, but empirical investigations have shown that it is relatively independent of the images at hand. To

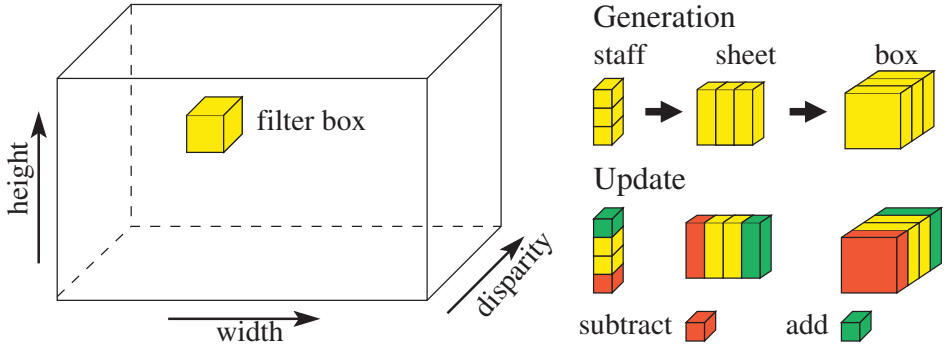

Fig. 5: Recursive filtering for 3 × 3 × 3 box: Generation and update.

determine the convergence, we compute the difference image between the disparity maps from the last two iterations and compute the standard deviation σ. Empirically we found, that a good threshold for σ is 0.005 of the search width. This results in 34 iterations for Tsukuba, 23 for Sawtooth, 30 for Venus, and 28 for Map.

4.2 Symmetric Support and Combination of Absolute Differences and Correlation

Fig. 1, right, showing the diagonal inhibition, gives a hint, that a box-shaped support as in Fig. might not be optimum. Our experiments have shown, that a symmetric support, where a box and a tilted box are added as shown in Fig., considerably improves the performance. Also the tilted box is implemented recursively by adding/subtracting staffs from the box.

As suggested by SCHARSTEIN & SZELISKI (2002), we have based the correlation scores on absolute differences. Experiments showed that the performance for squared differences was in nearly all cases worse than for absolute differences. For the absolute differences, we truncate the difference value with *trunc*. The matching score for absolute differences is $score_{abs\_diff} = 1 - abs\_diff/trunc$, with $0 \leq score \leq 1$.

When looking at results based on (normalized) cross-correlation compared to results where absolute differences have been employed, we got the idea, that the failure modes seemed to be different and that it might be useful to combine both. The combination is done by

$$score_{comb} = \frac{score_{abs\_diff} + weight * score_{corr}}{1 + weight} \quad (4)$$

with

$$weight = \frac{horiz\_grad}{threshold\_for\_mixing\_scores} \quad (5)$$

A large horizontal gradient *horiz_grad* (cf. Fig. 7 left) increases the probability for a good match for cross-correlation, because cross-correlation works best for strongly textured regions and the matching is done in horizontal direction for the normalized image pairs.

In addition to the combination, a special type of auto-correlation *auto_corr* (cf. Fig. 7 right) is used to indicate potentially false matches. It is determined as the maximum value of correlation along the horizontal line ranging from outside the matching window to the search width. If this auto-correlation

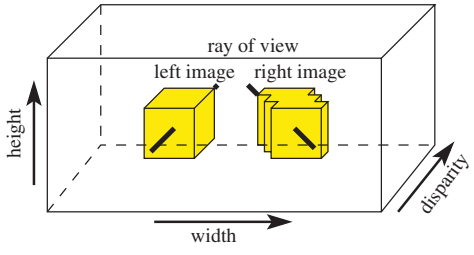

Fig. 6: Symmetric support.

Fig. 7: Horizontal gradient (left); Maximum auto-correlation along the horizontal line ranging from outside the matching window to the search width (right).

is large, it means that there are similar structures, i. e., repetitive textures, already in the reference image and, therefore, the match is highly likely to be ambiguous also in the other image. The auto-correlation is used to weight down the matching score by $score = score * (1 - 0.5 * auto\_corr)$. Both, horizontal gradient and auto-correlation are smoothed with a Gaussian filter.

4.3 Use of Color and Preference for Smaller Disparities

For color images we take the average of the individual results for the three colors for absolute differences. As we found, that color does not help too much for correlation, we correlated only the average images of the colors.

As noted in (Zhang & Kambhamettu 2002), there is a tendency of the cooperative approach to fatten regions with larger disparities. We counteract this by reducing the matching scores by (d is disparity)

$$score_{red} = score * (1 - \frac{d * preference}{search\_width} * (1 - 0.5 * auto\_corr)) \quad (6)$$

This is motivated as follows: Occluded regions must have a smaller disparity than their occluding regions. As there is no correct matching possible for occlusions, introducing a slight bias towards smaller disparities increases the probability, that occluded regions obtain correct, smaller disparities. For *preference*, a value of 0.05 was found suitable empirically. By reducing the matching score, there is a tendency for regions with a large auto-correlation (cf. above) to obtain a wrong, too small disparity value. Therefore, we reduce the preference with the same factor as above.

4.4 Enforcement of the Alignment of Image and Disparity Gradients

In many cases the materials or surface characteristics are considerably different at both sides of a disparity discontinuity or the disparity continuity casts a dark shadow. This results in a typical alignment of large disparity gradients with strong image gradients. While in (Zhang & Kambhamettu 2002) the image is segmented into several regions and the support is restricted to these regions, we take a more conservative policy. Additionally to the support size given in Tab. 3 (box_{supp_size}) we smooth the image with a 3 x 3 x 3 box filter (box_{333}) and mix the results according to the combined strength of $gradient_{comb} = gradient_{image} * gradient_{disparity}/255$. This combined gradient is then smoothed by a Gaussian filter. Both gradients are determined as absolute values by 3 x 3 Sobel-filters and scale from 0 to 255. An adequate weight and threshold was empirically found to be the combination of the combined gradient with half the search width:

$$weight = \frac{gradient_{comb}}{0.5 * search\_width} \quad (7)$$

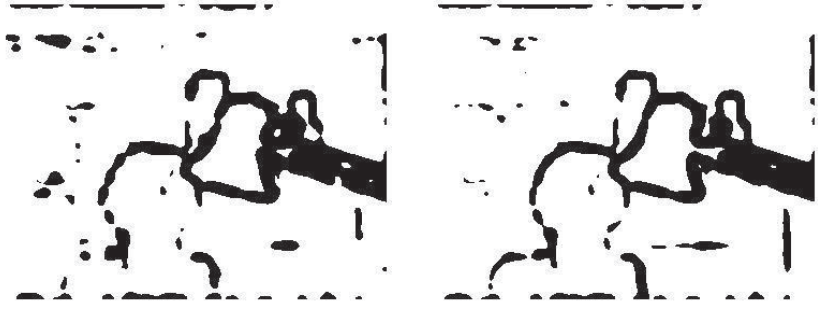

Fig. 8: Regions where the smoothing of the image function is reduced (black) due to large image and disparity gradients after 5 (left) and after 30 (right) iterations.

We truncate values below 1. The combined blurring reads

$$r(box_{comb}) = \frac{r(box_{\sup p\_size}) + weight * r(box_{333})}{1 + weight} \quad (8)$$

where $r(box_{xxx})$ stands for the result of filtering with the respective box-filter. The smaller box_{333} is only employed for values above the half search width. As can be seen, the regions with reduced smoothing fit better to the actual disparity continuities after convergence (right).

Fig. 9: Occluded regions at convergence.

4.5 *Determination of Occlusions and Sharpening of Disparity Discontinuities*

The tendency to estimate too large disparities (cf. also Section 4.3) is especially true for occluded regions. EGNAL & WILDES (2002) describe different approaches to determine occlusions. We use one of these approaches and reduce the probability of larger disparities for occluded regions, for which no matching is possible, and which must have a smaller disparity than their occluding regions. The determination of occlusions works best, when the result is already cleaned from gross errors. Empirically, the optimum procedure was found to let the basic algorithm converge first, and then to multiply $L_0(r, c, d)$ by $(search\_width - d)/search\_width$. This reduces the energy or probability of the original matching scores for larger disparities. They influence the process via equation (2). The reduction of the original matching scores is done several times. For the experiments in this papers it is done two times. After the reduction, the algorithm runs until σ falls again below the given threshold for convergence.

Because the disparities are already smooth when the algorithm has converged for the first time, it is sufficient to compute an indication for an occluded region by what EGNAL & WILDES (2002) term occlusion constraint. Here it is determined by the predicate $((d - d_{occ}) + (c - c_{occ})) < 0$. d and c are the disparity and the column coordinate of the point under investigation. d_{occ} and c_{occ} are the disparity and the column coordinate of the preceding point when starting from the left side of the image if no occluding point was found yet. If an occluding point was found, it is only updated to be the preceding point, when the above predicate is not true any more. To obtain compact regions, morphological opening and closing with circular structuring elements with a radius of 2.5 pixels are used. In Fig. the occlusions determined for the final convergence of the algorithm are shown.

4.6 *Assessment of the Gain of the Means*

Tab. 4 gives in the first row as reference the results when all means are employed. In the other rows results which are considerably worse than the reference are shown in bold, while results which are considerably better are marked in italics.

For the symmetric support in the second row the result is clear-cut. Apart from the untextured regions in Venus, there is an improvement nearly everywhere. The third row shows the results for absolute differences only, with the optimum truncation value of 4 gray values. As can be seen, the performance gain is considerable for the combination for all images besides Tsukuba. Our interpretation of this is as follows: Absolute differences make use of brightness differences even for weakly textured regions. This is useful only for constant lighting conditions, similar viewing angles, and well-behaved reflection functions. Yet, it is an advantage compared to (normalized) correla-

Tab. 4: Comparison of Different Means for Improvement: Percentage of bad pixels of results without using the respective means (worse results are marked in bold and better results in italics).

Means	Tsukuba			Sawtooth			Venus			Map		
	all	untex.	disc.	all	untex.	disc.	all	untex.	disc.	all	untex.	disc.
everything included	1.67	0.77	9.67	1.21	0.17	6.90	1.04	1.07	13.68	0.29	0.00	3.65
no symmetric support (4.2)	**2.06**	**1.19**	**11.90**	**1.49**	**0.43**	**10.29**	0.97	*0.67*	**14.67**	**0.39**	0.00	**4.63**
absolute differences only (4.2)	1.73	0.69	9.84	**1.42**	**0.20**	6.97	**1.32**	1.14	*10.38*	**2.53**	**1.19**	**16.82**
without use of auto-correlation (4.2)	**1.99**	**1.01**	**11.42**	1.22	0.16	6.82	1.17	**1.37**	**15.34**	0.29	0.00	3.68
no color used (4.3)	**2.15**	**1.14**	**12.39**	**1.40**	**0.44**	7.01	*0.82*	*0.71*	*10.16*	0.29	0.00	3.65
no preference for small disparities (4.3)	**2.05**	**1.18**	**11.78**	1.23	0.19	7.06	1.18	1.34	14.24	0.27	0.00	*3.24*
no alignment of gradients (4.4)	**2.82**	**1.53**	**16.07**	1.20	0.18	6.92	1.08	1.09	13.80	**0.51**	**0.24**	**6.63**
no occlusion modeling (4.5)	1.67	0.77	9.67	1.22	0.17	6.87	1.05	1.10	13.76	**0.39**	**0.71**	3.68

tion which is invariant to differences in brightness and contrast. Correlation can therefore produce a high score when matching a smooth bright to a smooth or even textured dark region when the weak texture happens to be similar, even though this is practically implausible. On the other hand, by restricting ourselves to relatively small truncation values, we do not make full use of heavily textured regions by absolute differences, where correlation works best.

From the fourth row it can be seen, that auto-correlation helps, though mostly for Tsukuba and Venus. Both have strong repetitive textures in the form of the books for Tsukuba and the rows of letters for Venus. The fifth row shows, that color is helpful. Yet, for Venus there is still ample room for improvement. This might stem from the fact, that Venus is partly relatively greenish and we only sum up the color information without weighting it according to contrast. From the sixth row one can see that no preference for small disparities results in a noticeable degradation of the overall results especially for Tsukuba, while for Map there is only a small improvement and for the other images there is none. The improvement by means of the enforcement of the alignment of image and disparity gradient in the sixth row is extremely large for Tsukuba and considerable for Map. Modeling occlusion in the last row only helps for Map. Yet, the importance of this means is still not absolutely clear.

5 Additional Results

To show, that our means for improvement and the set of parameters used are not only valid for the data set at www.middlebury.edu/stereo, we experimented with other image pairs with the same set of parameters given in Tab. 3. The only modification was to make the absolute differences invariant against a different average brightness of the image windows. This had to be done, because, opposed to the data set at www.middlebury.edu/stereo, many other image pairs have a significantly different gray value for homologous windows.

For the image pair Sport (cf. Fig. 10) from INRIA's Syntim image database one can see, that the approach works reasonably well for a relatively large disparity range (45 pixels search width for the epipolar resampled image Sport reduced to 267 x 271 pixels). The image pair Kitchen (cf. Fig. 11) stems from http://research.microsoft.com/virtuamsr/virtuatour.html (Antonio Criminisi & Phil Torr). The results show the high quality achievable with the improved approach. Similar results were obtained also for a large number of other images.

6 Conclusions

Ranking our results in the frame of the online version of Scharstein & Szeliski (2002) at www.middlebury.edu/stereo shows, that we have obtained a relatively

Fig. 10: Image pair Sport from INRIA's Syntim image database, result (occluded regions in red) with the same set of parameters as in Tab. 3 (left) and visualization (occluded regions in black).

Fig. 11: Image pair Kitchen from web page TORR & CRIMINISI, result, and visualization; occlusions and set of parameters cf. Fig. 10.

good performance also compared to the run time of our algorithm. On one hand, we have fine-tuned our approach for an optimum performance with the given data set. On the other hand, the last section has shown, that we obtain reasonable results also for other image pairs using the same set of parameters.

The results reported in Sun et al. (2002) are partly better than that presented in this paper. Though, it takes 288 seconds on a 500 MHz PC for Tsukuba, i. e., more than double as long as ours when scaled to 2.5 GHz. Graph cuts (Boykov et al. 2001, Kolmogorov & Zabih 2001) with and without the handling of occlusions also have a similar or better performance than our algorithm especially in combination with the fast max-flow algorithm. Yet, an interesting question would be if it might be possible to reach an improvement by some of our means for these algorithms. Especially the combination of correlation and absolute differences as well as using the auto-correlation function to characterize probably unreliable regions with repetitive texture might be fruitful in terms of performance as well as speed. Zhang & Kambhamettu (2002) has an advantage for depth discontinuities due to a more advanced modeling of the image function, but also it could possibly benefit from our more wide range of means of improvements.

Ways to proceed are for instance the use of more images as in Koch et al. (1999), Kolmogorov & Zabih (2002), where merging of pairs is done in object space based on relaxation, or to locally optimize the window size and shape (Veksler 2002). We have started to project the results into a third image by means of the trifocal tensor to obtain more evidence via cross-correlation especially for occluded regions. Finally, recent approaches such as Strecha et al. (2003, 2004) resemble the multi-image least squares matching approaches of Ebner et al. (1987) and Wrobel (1987). We are right now working in this area taking as a basis a large number of highly reliable points from automatic 3D reconstruction, e. g., Mayer (2005), as in Lhuillier & Quan (2002), to initialize the optimization of the surface.

Acknowledgement

We thank the anonymous reviewer for their helpful comments.

References

Boykov, Y., Veksler, O. & Zabih, R., 2001: Fast Approximate Energy Minimization via Graph Cuts. – IEEE Transactions on Pattern Analysis and Machine Intelligence **23** (11): 1222–1239.

Ebner, H., Fritsch, D., Gillessen, W. & Heipke, C., 1987: Integration von Bildzuordnung und Objektrekonstruktion innerhalb der digitalen Photogrammetrie. – Bildmessung und Luftbildwesen **5/87**: 194–203.

Egnal, G. & Wildes, R., 2002: Detecting Binocular Half-Occlusions: Empirical Comparison of Five Approaches. – IEEE Transactions on Pattern Analysis and Machine Intelligence **24**(8): 1127–1133.

Koch, R., Pollefeys, M. & Van Gool, L., 1999: Robust Calibration and 3D Geometric Modeling from Large Collections of Uncalibrated Images. – Mustererkennung 1999, Springer-Verlag, Berlin, Germany, 413–420.

Kolmogorov, V. & Zabih, R., 2001: Computing Visual Correspondence with Occlusions Using Graph Cuts. – Eighth International Conference on Computer Vision, 508–515.

Kolmogorov, V. & Zabih, R., 2002: Multi-Camera Scene Reconstruction via Graph Cuts. – Seventh European Conference on Computer Vision, Volume III, 82–96.

Lhuillier, M. & Quan, L., 2002: Match Propagation for Image Based Modeling and Rendering. – IEEE Transactions on Pattern Analysis and Machine Intelligence **24** (8): 1140–1146.

Marr, D. & Poggio, T., 1976: Cooperative Computation of Stereo Disparity. – Science **194:** 209–236.

Marr, D. & Poggio, T., 1979: A Computational Theory of Human Stereo Vision. – Proceedings Royal Society London B, **204:** 301–328.

Mayer, H., 2003: Analysis of Means to Improve Cooperative Disparity Estimation. – The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume (34) 3/W8, 25–31.

Mayer, H., 2005: Robust Least Squares Adjustment Based Orientation and Auto-Calibration of Wide-Baseline Image Sequences. – ICCV –

ISPRS Workshop Towards Benchmarking Automated Calibration, Orientation, and Surface Reconstruction from Images.

McGlone, J., Bethel, J. & Mikhail, E., 2004 (Ed.): Manual of Photogrammetry. – American Society of Photogrammetry and Remote Sensing, Bethesda, USA.

Scharstein, D. & Szeliski, R., 2002: A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. – International Journal of Computer Vision **47**(1): 7–42.

Strecha, C., Fransen, R. & Van Gool, L., 2004: Wide-Baseline Stereo from Multiple Views: A Probabilistic Account. – Computer Vision and Pattern Recognition, 552–559.

Strecha, C., Tuytelaars, T. & Van Gool, L., 2003: Dense Matching of Multiple Wide-Baseline Views. – Ninth International Conference on Computer Vision, Volume II: 1194–1201.

Sun, J., Shum, H.-Y. & Zheng, N.-N., 2002: Stereo Matching Using Belief Propagation. – Seventh European Conference on Computer Vision, Volume II, 510–524.

Veksler, O., 2002: Stereo Correspondence with Compact Windows via Minimum Ratio Cycles. – IEEE Transactions on Pattern Analysis and Machine Intelligence **24** (12): 1654–1660.

Wrobel, B., 1987: Digitale Bildzuordnung durch Facetten mithilfe von Objektraummodellen. – Bildmessung und Luftbildwesen **3/87**: 129–140.

Zhang, Y. & Kambhamettu, C., 2002: Stereo Matching with Segmentation-Based Cooperation. – Seventh European Conference on Computer Vision, Volume II, 556–571.

Zitnick, C. & Kanade, T., 2000: A Cooperative Algorithm for Stereo Matching and Occlusion Detection. – IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(7): 675–684.

Anschrift des Autors:

Prof. Dr.-Ing. Helmut Mayer
Universität der Bundeswehr München, Institut für Photogrammetrie und Kartographie, D-85577 Neubiberg
Tel.: +49-89-6004-3429, Fax: -6004-4090
e-mail: Helmut.Mayer@unibw.de

Manuskript eingereicht: Oktober 2005
Angenommen: Dezember 2005