Accurate and Efficient Stereo Processing by Semi-Global Matching and Mutual Information

Heiko Hirschmüller
Institute of Robotics and Mechatronics Oberpfaffenhofen
German Aerospace Center (DLR)
P.O. Box 1116, 82230 Wessling, Germany
heiko.hirschmueller@dlr.de

Abstract

This paper considers the objectives of accurate stereo matching, especially at object boundaries, robustness against recording or illumination changes and efficiency of the calculation. These objectives lead to the proposed Semi-Global Matching method that performs pixelwise matching based on Mutual Information and the approximation of a global smoothness constraint. Occlusions are detected and disparities determined with sub-pixel accuracy. Additionally, an extension for multi-baseline stereo images is presented. There are two novel contributions. Firstly, a hierarchical calculation of Mutual Information based matching is shown, which is almost as fast as intensity based matching. Secondly, an approximation of a global cost calculation is proposed that can be performed in a time that is linear to the number of pixels and disparities. The implementation requires just 1 second on typical images.

1. Introduction

Accurate, dense stereo matching is an important requirement for many applications, like 3D reconstruction. Most difficult are often the boundaries of objects and fine structures, which can appear blurred. Additional practical problems originate from recording and illumination differences or reflections, because matching is often directly based on intensities that can have quite different values for corresponding pixels. Furthermore, fast calculations are often required, either because of real-time applications or because of large images or many images that have to be processed efficiently.

An application were all of the three objectives come together is the reconstruction of urban terrain, captured by an airborne pushbroom camera. Accurate matching at object boundaries is important for reconstructing structured environments. Robustness against recording differences and illumination changes is vital, because this often cannot be controlled. Finally, efficient (off-line) processing is necessary, because the images and disparity ranges are huge (e.g. several 100MPixel with 1000 pixel disparity range).

2. Related Literature

There is a wide range of dense stereo algorithms [8] with different properties. Local methods, which are based on correlation can have very efficient implementations that are suitable for real time applications [5]. However, these methods assume constant disparities within a correlation window, which is incorrect at discontinuities and leads to blurred object boundaries. Certain techniques can reduce this effect [8, 5], but it cannot be eliminated. Pixelwise matching [1] avoids this problem, but requires other constraints for unambiguous matching (e.g. piecewise smoothness). *Dynamic Programming* techniques can enforce these constraints efficiently, but only within individual scanlines [1, 11]. This typically leads to streaking effects. Global approaches like *Graph Cuts* [7, 2] and *Belief Propagation* [10] enforce the matching constraints in two dimensions. Both approaches are quite memory intensive and Graph Cuts is rather slow. However, it has been shown [4] that Belief Propagation can be implemented very efficiently.

The matching cost is commonly based on intensity differences, which may be sampling insensitive [1]. Intensity based matching is very sensitive to recording and illumination differences, reflections, etc. *Mutual Information* has been introduced in computer vision for matching images with complex relationships of corresponding intensities, possibly even images of different sensors [12]. Mutual Information has already been used for correlation based stereo matching [3] and Graph Cuts [6]. It has been shown [6] that it is robust against many complex intensity transformations and even reflections.

3. Semi-Global Matching

3.1. Outline

The Semi-Global Matching (SGM) method is based on the idea of pixelwise matching of Mutual Information and approximating a global, 2D smoothness constraint by combining many 1D constraints. The algorithm is described in distinct processing steps, assuming a general stereo geometry of two or more images with known epipolar geometry. Firstly, the pixelwise cost calculation is discussed in Section 3.2. Secondly, the implementation of the smoothness constraint is presented in Section 3.3. Next, the disparity is determined with sub-pixel accuracy and occlusion detection in Section 3.4. An extension for multi-baseline matching is described in Section 3.5. Finally, the complexity and implementation is discussed in Section 3.6.

3.2. Pixelwise Cost Calculation

The matching cost is calculated for a base image pixel $\mathbf{p}$ from its intensity $I_{b\mathbf{p}}$ and the suspected correspondence $I_{m\mathbf{q}}$ at $\mathbf{q} = e_{bm}(\mathbf{p}, d)$ of the match image. The function $e_{bm}(\mathbf{p}, d)$ symbolizes the epipolar line in the match image for the base image pixel $\mathbf{p}$ with the line parameter d. For rectified images $e_{bm}(\mathbf{p}, d) = [\mathbf{p}_x - d, \mathbf{p}_y]^T$ with d as disparity.

An important aspect is the size and shape of the area that is considered for matching. The robustness of matching is increased with large areas. However, the implicit assumption about constant disparity inside the area is violated at discontinuities, which leads to blurred object borders and fine structures. Certain shapes and techniques can be used to reduce blurring, but it cannot be avoided [5]. Therefore, the assumption of constant disparities in the vicinity of $\mathbf{p}$ is discarded. This means that only the intensities $I_{b\mathbf{p}}$ and $I_{m\mathbf{q}}$ itself can be used for calculating the matching cost.

One choice of pixelwise cost calculation is the sampling insensitive measure of Birchfield and Tomasi [1]. The cost $C_{BT}(\mathbf{p}, d)$ is calculated as the absolute minimum difference of intensities at $\mathbf{p}$ and $\mathbf{q} = e_{bm}(\mathbf{p}, d)$ in the range of half a pixel in each direction along the epipolar line.

Alternatively, the matching cost calculation is based on Mutual Information (MI) [12], which is insensitive to recording and illumination changes. It is defined from the entropy H of two images (i.e. their information content) as well as their joined entropy.

$$MI_{I_1,I_2} = H_{I_1} + H_{I_2} - H_{I_1,I_2} \tag{1}$$

The entropies are calculated from the probability distributions P of intensities of the associated images.

$$H_I = -\int_0^1 P_I(i) \log P_I(i) di \tag{2}$$

$$H_{I_1,I_2} = -\int_0^1 \int_0^1 P_{I_1,I_2}(i_1, i_2) \log P_{I_1,I_2}(i_1, i_2) di_1 di_2 \tag{3}$$

For well registered images the joined entropy H_{I_1,I_2} is low, because one image can be predicted by the other, which corresponds to low information. This increases their Mutual Information. In the case of stereo matching, one image needs to be warped according to the disparity image D for matching the other image, such that corresponding pixels are at the same location in both images, i.e. $I_1 = I_b$ and $I_2 = f_D(I_m)$.

Equation (1) operates on full images and requires the disparity image a priori. Both prevent the use of MI as matching cost. Kim et al. [6] transformed the calculation of the joined entropy H_{I_1,I_2} into a sum of data terms using Taylor expansion. The data term depends on corresponding intensities and is calculated individually for each pixel $\mathbf{p}$.

$$H_{I_1,I_2} = \sum_{\mathbf{p}} h_{I_1,I_2}(I_{1\mathbf{p}}, I_{2\mathbf{p}}) \tag{4}$$

The data term h_{I_1,I_2} is calculated from the probability distribution P_{I_1,I_2} of corresponding intensities. The number of corresponding pixels is n. Convolution with a 2D Gaussian (indicated by $\otimes g(i,k)$) effectively performs Parzen estimation [6].

$$h_{I_1,I_2}(i,k) = -\frac{1}{n} \log(P_{I_1,I_2}(i,k) \otimes g(i,k)) \otimes g(i,k) \tag{5}$$

The probability distribution of corresponding intensities is defined with the operator T[], which is 1 if its argument is true and 0 otherwise.

$$P_{I_1,I_2}(i,k) = \frac{1}{n} \sum_{\mathbf{p}} \mathrm{T}[(i,k) = (I_{1\mathbf{p}}, I_{2\mathbf{p}})] \tag{6}$$

Kim et al. argued that the entropy H_{I_1} is constant and H_{I_2} is almost constant as the disparity image merely redistributes the intensities of I_2. Thus, $h_{I_1,I_2}(I_{1\mathbf{p}}, I_{2\mathbf{p}})$ serves as cost for matching the intensities $I_{1\mathbf{p}}$ and $I_{2\mathbf{p}}$. However, if occlusions are considered then some intensities of I_1 and I_2 do not have a correspondence. These intensities should not be included in the calculation, which results in non-constant entropies H_{I_1} and H_{I_2}. Therefore, it is suggested to calculate these entropies analog to the joined entropy.

$$H_I = \sum_{\mathbf{p}} h_I(I_{\mathbf{p}}), \quad h_I(i) = -\frac{1}{n} \log(P_I(i) \otimes g(i)) \otimes g(i) \tag{7}$$

The probability distribution P_I must not be calculated over the whole images I_1 and I_2, but only over the corresponding parts (otherwise occlusions would be ignored and H_{I_1} and H_{I_2} would be almost constant). That is easily done by just summing the corresponding rows and columns of the joined probability distribution, e.g. $P_{I_1}(i) = \sum_k P_{I_1,I_2}(i,k)$. The resulting definition of Mutual Information is,

$$MI_{I_1,I_2} = \sum_{\mathbf{p}} mi_{I_1,I_2}(I_{1\mathbf{p}}, I_{2\mathbf{p}}) \tag{8a}$$

$$mi_{I_1,I_2}(i,k) = h_{I_1}(i) + h_{I_2}(k) - h_{I_1,I_2}(i,k). \tag{8b}$$

This leads to the definition of the MI matching cost.

$$C_{MI}(\mathbf{p},d) = -mi_{I_b,f_D(I_m)}(I_{b\mathbf{p}}, I_{m\mathbf{q}})\text{with } \mathbf{q} = e_{bm}(\mathbf{p},d) \tag{9}$$

The remaining problem is that the disparity image is required for warping I_m, *before* $mi()$ can be calculated. Kim et al. suggested an iterative solution, which starts with a random disparity image for calculating the cost C_{MI}. This cost is then used for matching both images and calculating a new disparity image, which serves as the base of the next iteration. The number of iterations is rather low (e.g. 3), because even wrong disparity images (e.g. random) allow a good estimation of the probability distribution P. This solution is well suited for iterative stereo algorithms like Graph Cuts [6], but it would increase the runtime of non-iterative algorithms unnecessarily.

Therefore, a hierarchical calculation is proposed, which recursively uses the (up-scaled) disparity image, that has been calculated at half resolution, as initial disparity. If the overall complexity of the algorithm is $O(WHD)$ (i.e. width $\times$ height $\times$ disparity range), then the runtime at half resolution is reduced by factor $2^3 = 8$. Starting with a random disparity image at a resolution of $\frac{1}{16}$th and initially calculating 3 iterations increases the overall runtime by the factor,

$$1 + \frac{1}{2^3} + \frac{1}{4^3} + \frac{1}{8^3} + 3\frac{1}{16^3} \approx 1.14. \tag{10}$$

Thus, the theoretical runtime of the hierarchically calculated C_{MI} would be just 14% slower than that of C_{BT}, ignoring the overhead of MI calculation and image scaling. It is noteworthy that the disparity image of the lower resolution level is used only for estimating the probability distribution P and calculating the costs C_{MI} of the higher resolution level. Everything else is calculated from scratch to avoid passing errors from lower to higher resolution levels.

3.3. Aggregation of Costs

Pixelwise cost calculation is generally ambiguous and wrong matches can easily have a lower cost than correct ones, due to noise, etc. Therefore, an additional constraint is added that supports smoothness by penalizing changes of neighboring disparities. The pixelwise cost and the smoothness constraints are expressed by defining the energy $E(D)$ that depends on the disparity image D.

$$\begin{aligned} E(D) = \sum_{\mathbf{p}} C(\mathbf{p}, D_{\mathbf{p}}) + \sum_{\mathbf{q} \in N_{\mathbf{p}}} P_1 \mathrm{T}[|D_{\mathbf{p}} - D_{\mathbf{q}}| = 1] \\ + \sum_{\mathbf{q} \in N_{\mathbf{p}}} P_2 \mathrm{T}[|D_{\mathbf{p}} - D_{\mathbf{q}}| > 1] \end{aligned} \tag{11}$$

The first term is the sum of all pixel matching costs for the disparities of D. The second term adds a constant penalty P_1 for all pixels $\mathbf{q}$ in the neighborhood $N_{\mathbf{p}}$ of $\mathbf{p}$, for which the disparity changes a little bit (i.e. 1 pixel). The third term adds a larger constant penalty P_2, for all larger disparity changes. Using a lower penalty for small changes permits an adaptation to slanted or curved surfaces. The constant penalty for all larger changes (i.e. independent of their size) preserves discontinuities [2]. Discontinuities are often visible as intensity changes. This is exploited by adapting P_2 to the intensity gradient, i.e. $P_2 = \frac{P_2'}{|I_{b\mathbf{p}} - I_{b\mathbf{q}}|}$. However, it has always to be ensured that $P_2 \geq P_1$.

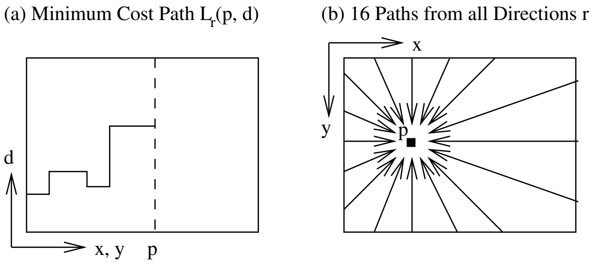
The problem of stereo matching can now be formulated as finding the disparity image D that minimizes the energy $E(D)$. Unfortunately, such a global minimization (2D) is NP-complete for many discontinuity preserving energies [2]. In contrast, the minimization along individual image rows (1D) can be performed efficiently in polynomial time using Dynamic Programming [1, 11]. However, Dynamic Programming solutions easily suffer from streaking [8], due to the difficulty of relating the 1D optimizations of individual image rows to each other in a 2D image. The problem is, that very strong constraints in one direction (i.e. along image rows) are combined with none or much weaker constraints in the other direction (i.e. along image columns).

This leads to the new idea of aggregating matching costs in 1D from *all* directions equally. The aggregated (smoothed) cost $S(\mathbf{p}, d)$ for a pixel $\mathbf{p}$ and disparity d is calculated by summing the costs of all 1D minimum cost paths that end in pixel $\mathbf{p}$ at disparity d (Figure 1). It is noteworthy that only the cost of the path is required and not the path itself.

Let $L'_{\mathbf{r}}$ be a path that is traversed in the direction $\mathbf{r}$. The cost $L'_{\mathbf{r}}(\mathbf{p}, d)$ of the pixel $\mathbf{p}$ at disparity d is defined recursively as,

$$\begin{aligned} L'_{\mathbf{r}}(\mathbf{p}, d) = C(\mathbf{p}, d) + \min(&L'_{\mathbf{r}}(\mathbf{p} - \mathbf{r}, d), \\ &L'_{\mathbf{r}}(\mathbf{p} - \mathbf{r}, d-1) + P_1, L'_{\mathbf{r}}(\mathbf{p} - \mathbf{r}, d+1) + P_1, \\ &\min_i L'_{\mathbf{r}}(\mathbf{p} - \mathbf{r}, i) + P_2). \end{aligned} \tag{12}$$

The pixelwise matching cost C can be either C_{BT} or C_{MI}. The remainder of the equation adds the lowest cost of the


(a) Minimum Cost Path $L_r(p, d)$

(b) 16 Paths from all Directions r

Figure 1. Aggregation of costs.

previous pixel $\mathbf{p}-\mathbf{r}$ of the path, including the appropriate penalty for discontinuities. This implements the behavior of equation (11) along an arbitrary 1D path. This cost does not enforce the *visibility* or *ordering* constraint, because both concepts cannot be realized for paths that are not identical to epipolar lines. Thus, the approach is more similar to *Scanline Optimization* [8] than traditional Dynamic Programming solutions.

The values of L' permanently increase along the path, which may lead to very large values. However, equation (12) can be modified by subtracting the minimum path cost of the previous pixel from the whole term.

$$\begin{aligned} L_{\mathbf{r}}(\mathbf{p},d) = C(\mathbf{p},d) + \min(&L_{\mathbf{r}}(\mathbf{p}-\mathbf{r},d),\\ &L_{\mathbf{r}}(\mathbf{p}-\mathbf{r},d-1)+P_1, L_{\mathbf{r}}(\mathbf{p}-\mathbf{r},d+1)+P_1,\\ &\min_i L_{\mathbf{r}}(\mathbf{p}-\mathbf{r},i)+P_2) - \min_k L_{\mathbf{r}}(\mathbf{p}-\mathbf{r},k) \end{aligned} \tag{13}$$

This modification does not change the actual path through disparity space, since the subtracted value is constant for all disparities of a pixel $\mathbf{p}$. Thus, the position of the minimum does not change. However, the upper limit can now be given as $L \leq C_{max} + P_2$.

The costs $L_{\mathbf{r}}$ are summed over paths in all directions $\mathbf{r}$. The number of paths must be at least 8 and should be 16 for providing a good coverage of the 2D image.

$$S(\mathbf{p},d) = \sum_{\mathbf{r}} L_{\mathbf{r}}(\mathbf{p},d) \tag{14}$$

The upper limit for S is easily determined as $S \leq 16(C_{max} + P_2)$.

3.4. Disparity Computation

The disparity image D_b that corresponds to the base image I_b is determined as in local stereo methods by selecting for each pixel $\mathbf{p}$ the disparity d that corresponds to the minimum cost, i.e. $\min_d S(\mathbf{p},d)$. For sub-pixel estimation, a quadratic curve is fitted through the neighboring costs (i.e. at the next higher or lower disparity) and the position of the minimum is calculated.

Using a quadratic curve is theoretically justified only for a simple correlation using the sum of squared differences. However, is is used as an approximation due to the simplicity of calculation.

The disparity image D_m that corresponds to the match image I_m can be determined from the same costs, by traversing the epipolar line, that corresponds to the pixel $\mathbf{q}$ of the match image. Again, the disparity d is selected, which corresponds to the minimum cost, i.e. $\min_d S(e_{mb}(\mathbf{q},d),d)$. However, the cost aggregation step does not treat the base and match images symmetrically. Therefore, better results can be expected, if D_m is calculated from scratch. Outliers are filtered from D_b and D_m, using a median filter with a small window (i.e. 3×3).

The calculation of D_b as well as D_m permits the determination of occlusions and false matches by performing a consistency check. Each disparity of D_b is compared with its corresponding disparity of D_m. The disparity is set to invalid (D_{inv}) if both differ.

$$D_{\mathbf{p}} = \begin{cases} D_{b\mathbf{p}} & \text{if } |D_{b\mathbf{p}} - D_{m\mathbf{q}}| \leq 1,\ \mathbf{q} = e_{bm}(\mathbf{p}, D_{b\mathbf{p}}), \\ D_{inv} & \text{otherwise.} \end{cases} \tag{15}$$

The consistency check enforces the *uniqueness constraint*, by permitting one to one mappings only.

3.5. Extension for Multi-Baseline Matching

The algorithm could be extended for multi-baseline matching, by calculating a combined pixelwise matching cost of correspondences between the base image and all match images. However, valid and invalid costs would be mixed near discontinuities, depending on the visibility of a pixel in a match image. The consistency check (Section 3.4) can only distinguish between valid (visible) and invalid (occluded or mismatched) pixels, but it can not separate valid and invalid costs afterwards. Thus, the consistency check would invalidate all areas that are not seen by all images, which leads to unnecessarily large invalid areas. Without the consistency check, invalid costs would introduce matching errors near discontinuities, which leads to fuzzy object borders.

Therefore, it is better to calculate several disparity images from individual image pairs, exclude all invalid pixels by the consistency check and then combine the result. Let the disparity D_k be the result of matching the base image I_b against a match image I_{mk}. The disparities of the images D_k are scaled differently, according to some factor t_k. For rectified images, this factor corresponds to the length of the baseline between I_b and I_{mk}.

The robust combination selects the median of all disparities $\frac{D_{k\mathbf{p}}}{t_k}$ for a certain pixel $\mathbf{p}$. Additionally, the accuracy is increased by calculating the weighted mean of all correct disparities (i.e. within the range of 1 pixel around the median). This is done by using t_k as weighting factor.

$$D_{\mathbf{p}} = \frac{\sum_{k \in V_{\mathbf{p}}} D_{k\mathbf{p}}}{\sum_{k \in V_{\mathbf{p}}} t_k}, \quad V_{\mathbf{p}} = \{k | \left| \frac{D_{k\mathbf{p}}}{t_k} - \operatorname*{med}_i \frac{D_{i\mathbf{p}}}{t_i} \right| \leq \frac{1}{t_k}\} \quad (16)$$

This combination is robust against matching errors in some disparity images and it also increases the accuracy.

3.6. Complexity and Implementation

The calculation of the pixelwise cost C_{MI} starts with collecting all alleged correspondences (i.e. defined by an initial disparity as described in Section 3.2) and calculating P_{I_1,I_2}. The size of P is the square of the number of intensities, which is constant (i.e. 256×256). The subsequent operations consist of Gaussian convolutions of P and calculating the logarithm. The complexity depends only on the collection of alleged correspondences due to the constant size of P. Thus, $O(WH)$ with W as image width and H as image height.

The pixelwise matching costs for all pixels $\mathbf{p}$ at all disparities d are scaled to 11 bit integer values and stored in a 16 bit array $C(\mathbf{p}, d)$. Scaling to 11 bit guarantees that all aggregated costs do not exceed the 16 bit limit. A second 16 bit integer array of the same size is used for the aggregated cost values $S(\mathbf{p}, d)$. The array is initialized with 0. The calculation starts for each direction $\mathbf{r}$ at all pixels $\mathbf{b}$ of the image border with $L_{\mathbf{r}}(\mathbf{b}, d) = C(\mathbf{b}, d)$. The path is traversed in forward direction according to equation (13). For each visited pixel $\mathbf{p}$ along the path, the costs $L_{\mathbf{r}}(\mathbf{p}, d)$ are added to $S(\mathbf{p}, d)$ for all disparities d. The calculation of equation (13) requires $O(D)$ steps at each pixel, since the minimum cost of the previous pixel (e.g. $\min_k L_{\mathbf{r}}(\mathbf{p} - \mathbf{r}, k)$) is constant for all disparities and can be pre-calculated. Each pixel is visited exactly 16 times, which results in a total complexity of $O(WHD)$. The regular structure and simple operations (i.e. additions and comparisons) permit parallel calculations using integer based SIMD[1] assembler instructions.

The disparity computation and consistency check requires visiting each pixel at each disparity a constant number of times. Thus, the complexity is $O(WHD)$ as well.

The 16 bit arrays C and S have a size of $W \times H \times D$, which can exceed the available memory for larger images and disparity ranges. The suggested remedy is to split the input image into tiles, which are processed individually. The tiles overlap each other by a few pixels, since the pixels at the image border receive support by the global cost function only from one side. The overlapping pixels are ignored for combining the tiles to the final disparity image. This solution allows processing of almost arbitrarily large images.

4. Experimental Results

4.1. Stereo Images with Ground Truth

Three stereo image pairs with ground truth [8, 9] have been selected for evaluation (first row of Figure 2). The images 2 and 4 have been used from the Teddy and Cones image sequences. All images have been processed with a disparity range of 32 pixel.

The MWMF method is a local, correlation based, real time algorithm [5], which has been shown [8] to produce better object borders (i.e. less fuzzy) than many other local methods. The second row of Figure 2 shows the resulting disparity images. The blurring of object borders is typical for local methods. The calculation of Teddy has been performed in just 0.071s on a Xeon with 2.8GHz.

Belief Propagation (BP) [10] minimizes a global cost function (e.g. equation (11)) by iteratively passing messages in a graph that is defined by the four connected image grid. The messages are used for updating the nodes of the graph. The disparity is in the end selected individually at each node. Similarly, SGM can be described as passing messages independently, from all directions along 1D paths for updating nodes. This is done sequentially as each message depends on one predecessor only. Thus, messages are passed through the whole image. In contrast, BP sends messages in a 2D graph. Thus, the schedule of messages that reaches each node is different and BP requires an iterative solution. The number of iterations determines the distance from which information is passed in the image.

The efficient BP algorithm[2] [4] uses a hierarchical approach and several optimizations for reducing the complexity. The complexity and memory requirements are very similar to SGM. The third row of Figure 2 shows good results of Tsukuba. However, the results of Teddy and especially Cones are rather blocky, despite attempts to get the best results by parameter tuning. The calculation of Teddy took 4.5s on the same computer.

The Graph Cuts method[3] [7] iteratively minimizes a global cost function (e.g. equation (11) with $P_1 = P_2$) as well. The fourth row of Figure 2 shows the results, which are much better for Teddy and Cones, especially near object borders and fine structures like the leaves. However, the complexity of the algorithm is much higher. The calculation of Teddy has been done in 55s on the same computer.

The results of SGM with C_{BT} as matching cost (i.e. the same as for BP and GC) are shown in the fifth row of Figure

[1] Single Instruction, Multiple Data

[2] http://people.cs.uchicago.edu/˜pff/bp/

[3] http://www.cs.cornell.edu/People/vnk/software.html

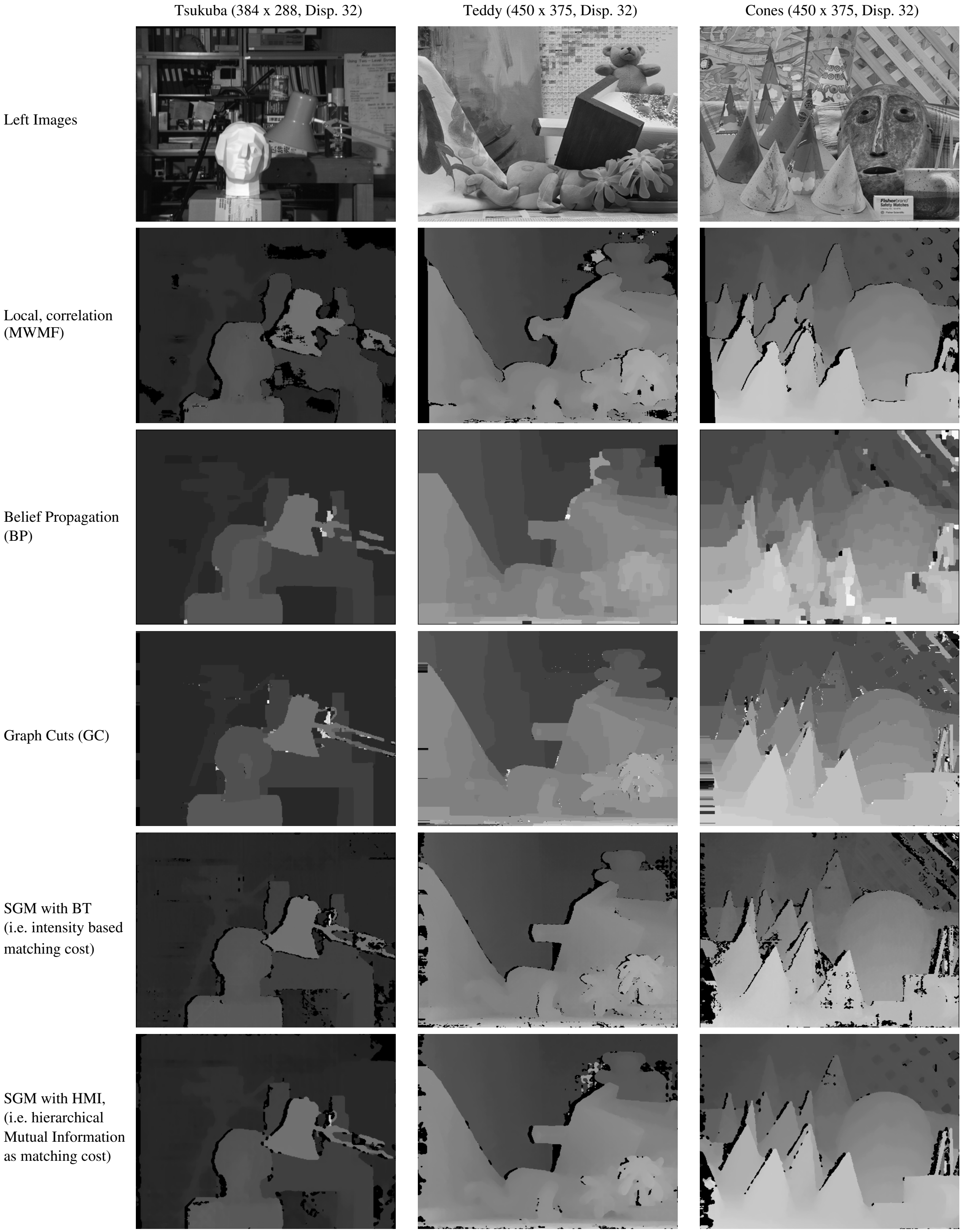
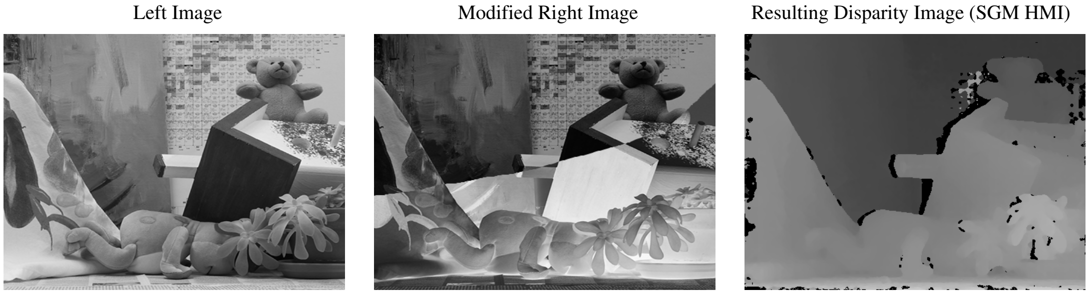

Tsukuba (384 x 288, Disp. 32)
Teddy (450 x 375, Disp. 32)
Cones (450 x 375, Disp. 32)
Left Images
Local, correlation (MWMF)
Belief Propagation (BP)
Graph Cuts (GC)
SGM with BT (i.e. intensity based matching cost)
SGM with HMI, (i.e. hierarchical Mutual Information as matching cost)

Figure 2. Comparison of different stereo methods.


Left Image

Modified Right Image

Resulting Disparity Image (SGM HMI)

Figure 3. Result of matching modified Teddy images with SGM (HMI).

2, using the best parameters for the set of all images. The quality of the result comes close to Graph Cuts. Only, the textureless area on the right of the Teddy is handled worse. Slanted surfaces appear smoother than with Graph Cuts, due to sub pixel interpolation. The calculation of Teddy has been performed in 1.0s. Cost aggregation requires almost half of the processing time.

The last row of Figure 2 shows the result of SGM with the hierarchical calculation of C_{MI} as matching cost. The disparity image of Tsukuba and Teddy appear equally well and Cones appears much better. This is an indication that the matching tolerance of MI is beneficial even for carefully captured images. The calculation of Teddy took 1.3s. This is just 30% slower than the non-hierarchical, intensity based version.

The disparity images have been compared to the ground truth. All disparities that differ by more than 1 are treated as errors. Occluded areas (i.e. identified using the ground truth) have been ignored. Missing disparities (i.e. black areas) have been interpolated by using the lowest neighboring disparities. Figure 4 presents the resulting graph. This quantitative analysis confirms that SGM performs as well as other global approaches. Furthermore, MI based matching results in even better disparity images.

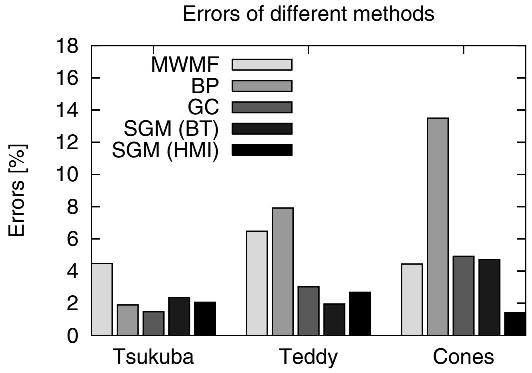


Figure 4. Errors of different stereo methods.

The power of MI based matching can be demonstrated by manually modifying the right image of Teddy by dimming the upper half and inverting the intensities of the lower half (Figure 3). Such an image pair cannot be matched by intensity based costs. However, the MI based cost handles this situation easily as shown on the right. More examples about the power of MI based stereo matching are shown by Kim et al. [6].

4.2. Stereo Images of an Airborne Pushbroom Camera

The SGM (HMI) method has been tested on huge images (i.e. several 100MPixel) of an airborne pushbroom camera, which records 5 panchromatic images in different angles. The appropriate camera model and non-linearity of the flight path has been taken into account for calculating the epipolar lines.

A difficult test object is Neuschwanstein castle (Figure 5a), because of high walls and towers, which result in high disparity changes and large occluded areas. The castle has been recorded 4 times using different flight paths. Each flight path results in a multi-baseline stereo image from which the disparity has been calculated. All disparity images have been combined for increasing robustness.

Figure 5b shows the end result, using a hierarchical, correlation based method [13]. The object borders appear fuzzy and the towers are mostly unrecognized. The result of the SGM (HMI) method is shown in Figure 5c. All object borders and towers have been properly detected. Stereo methods with intensity based pixelwise costs (e.g. Graph Cuts and SGM (BT)) failed on these images completely, because of large intensity differences of correspondences. This is caused by recording differences as well as unavoidable changes of lighting and the scene during the flight (i.e. corresponding points are recorded at different times on the flight path). Nevertheless, the MI based matching cost handles the differences easily.

The processing time is one hour on a 2.8GHz Xeon for matching 11MPixel of a base image against 4 match images with an average disparity range of 400 pixel.

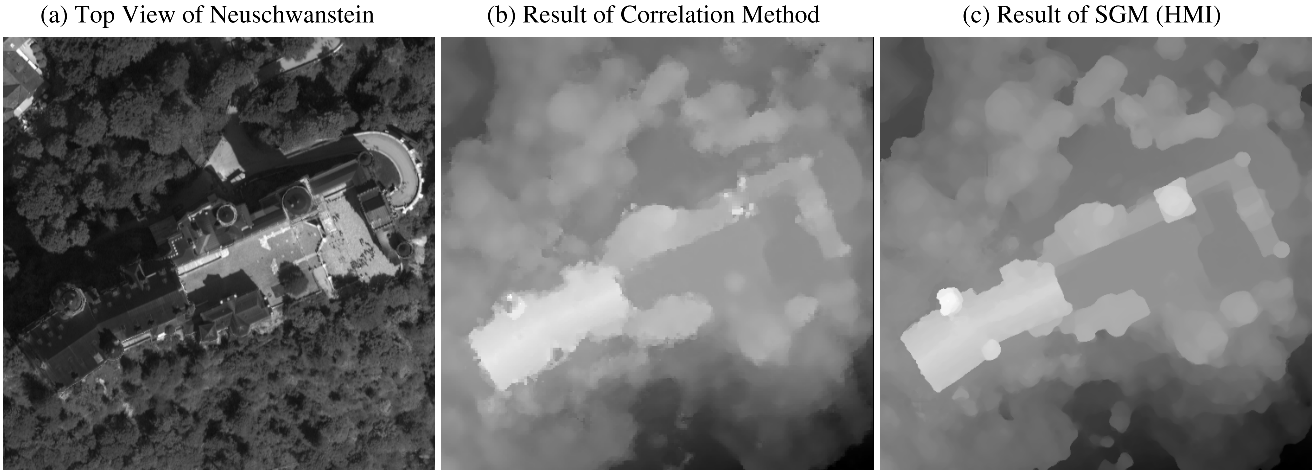

(a) Top View of Neuschwanstein
(b) Result of Correlation Method
(c) Result of SGM (HMI)

Figure 5. Neuschwanstein castle (Germany), recorded by an airborne pushbroom camera.

5. Conclusion

It has been shown that a hierarchical calculation of a Mutual Information based matching cost can be performed at almost the same speed as an intensity based matching cost. This opens the way for robust, illumination insensitive stereo matching in a broad range of applications. Furthermore, it has been shown that a global cost function can be approximated efficiently in $O(WHD)$.

The resulting Semi-Global Matching (SGM) method performs much better matching than local methods and is almost as accurate as global methods. However, SGM is much faster than global methods. A near real-time performance on small images has been demonstrated as well as an efficient calculation of huge images.

6. Acknowledgments

I would like to thank Klaus Gwinner, Johann Heindl, Frank Lehmann, Martin Oczipka, Sebastian Pless, Frank Scholten and Frank Trauthan for inspiring discussions and Daniel Scharstein and Richard Szeliski for makeing stereo images with ground truth available.

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] Y. Boykov, O. Veksler, and R. Zabih. Efficient approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, 2001.

[3] G. Egnal. Mutual information as a stereo correspondence measure. Technical Report MS-CIS-00-20, Computer and Information Science, University of Pennsylvania, Philadelphia, USA, 2000.

[4] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient belief propagation for early vision. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2004.

[5] H. Hirschmüller, P. R. Innocent, and J. M. Garibaldi. Real-time correlation-based stereo vision with reduced border errors. *International Journal of Computer Vision*, 47(1/2/3):229–246, April-June 2002.

[6] J. Kim, V. Kolmogorov, and R. Zabih. Visual correspondence using energy minimization and mutual information. In *International Conference on Computer Vision*, 2003.

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

[8] 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.

[9] 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.

[10] J. Sun, H. Y. Shum, and N. N. Zheng. Stereo matching using belief propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(7):787–800, July 2003.

[11] G. Van Meerbergen, M. Vergauwen, M. Pollefeys, and L. Van Gool. A hierarchical symmetric stereo algorithm using dynamic programming. *International Journal of Computer Vision*, 47(1/2/3):275–285, April-June 2002.

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

[13] F. Wewel, F. Scholten, and K. Gwinner. High resolution stereo camera (hrsc) - multispectral 3d-data acquisition and photogrammetric data processing. *Canadian Journal of Remote Sensing*, 26(5):466–474, 2000.