MONOCULAR-DEPTH ASSISTED SEMI-GLOBAL MATCHING

M. Hödel[1,*], T. Koch[2], L. Hoegner[1], U. Stilla[1]

[1] Photogrammetry and Remote Sensing, Technical University of Munich (TUM), Germany
[2] Chair of Remote Sensing Technology - Technical University of Munich (TUM), Germany
- (max.hoedel, tobias.koch, ludwig.hoegner, stilla)@tum.de

ICWG II/III: Pattern Analysis in Remote Sensing

KEY WORDS: mono-depth, single-image depth estimation, SGM, 3D reconstruction, image matching

ABSTRACT:

Reconstruction of dense photogrammetric point clouds is often based on depth estimation of rectified image pairs by means of pixel-wise matching. The main drawback lies in the high computational complexity compared to that of the relatively straightforward task of laser triangulation. Dense image matching needs oriented and rectified images and looks for point correspondences between them. The search for these correspondences is based on two assumptions: pixels and their local neighborhood show a similar radiometry and image scenes are mostly homogeneous, meaning that neighboring points in one image are most likely also neighbors in the second. These rules are violated, however, at depth changes in the scene. Optimization strategies tend to find the best depth estimation based on the resulting disparities in the two images. One new field in neural networks is the estimation of a depth image from a single input image through learning geometric relations in images. These networks are able to find homogeneous areas as well as depth changes, but result in a much lower geometric accuracy of the estimated depth compared to dense matching strategies. In this paper, a method is proposed extending the Semi-Global-Matching algorithm by utilizing a-priori knowledge from a monocular depth estimating neural network to improve the point correspondence search by predicting the disparity range from the single-image depth estimation (SIDE). The method also saves resources through path optimization and parallelization. The algorithm is benchmarked on Middlebury data and results are presented both quantitatively and qualitatively.

1. INTRODUCTION

Semi-Global Matching (SGM) (Hirschmueller, 2005) is a computer vision method for finding the correlation of pixel pairs in stereo images by determining the *disparity* value. This value corresponds to the sensor distance in pixels between equivalent points of an image (corrected by an offset). By doing this across the image a disparity map can be calculated, which in return can be converted to a metric depth image. On modern architectures (GPUs) SGM can be performed in near real-time, allowing for computation of such maps on a frame-by-frame basis, but at high computational cost. The applications of disparity mapping cover many 3D stereo vision tasks, but is best suited for those which require quick or instantaneous estimations rather than precise, post processed depth maps.

Recently, *convolutional neural networks (CNNs)* have provided new possibilities in dense matching. (Zbontar et al., 2016) proposes a deep learning-based matching method based on CNNs (MC-CNN) by substituting handcrafted cost functions, such as Census or Mutual Information, by training a network on pairs of small image patches with known ground-truth disparity maps. (Luo et al., 2016) learns informative image patch representations by employing a siamese network which extracts marginal distributions over all possible disparities for each pixel. Another branch in deep learning methods tackles disparity map computation from single views. The task of single-image depth estimation is ambiguous and NP-hard. However, tremendous results have been achieved with recent deep learning-based approaches by training a network with RGB and corresponding depth map pairs to regress pixel-wise depth predictions for single-view RGB images. The ability to estimate dense depth maps requires a huge amount of training data, usually obtained from stereo image sequences (Geiger et al., 2013), RGB-D video streams (Silberman et al., 2012) or synthetic datasets (McCormac et al., 2017).

Despite the vast progress in producing reasonable depth maps from single views there are still deficits in the preservation of depth discontinuities and planar surfaces, as well as a high error-proneness of textured or illuminated planar objects (Koch et al., 2018). Although the results of current single-image depth estimation methods are not comparable with the results of classic stereo approaches, they can still provide valuable scene information that can in turn help improve and accelerate stereo matching methods.

This information can be exploited in an effort to reduce the time and operations needed to perform SGM. By interpreting the input from a monocular-depth estimating neural network, boundary estimations can be derived allowing for a restriction of the disparity range (inverse depth range) that needs evaluation. This significantly reduces the number of necessary operations while still preserving the accuracy of output depth maps, allowing for faster computation across a wide variety of scenes.

2. CNN-BASED SINGLE-IMAGE DEPTH ESTIMATION

Current CNN architectures are capable of implicitly inferring geometric information solely from RGB images and approach the problem of depth estimation as a pixel-level regression task. Since the early work of (Eigen et al., 2014), proposing a deep learning approach for the task of SIDE, this field has become

*Corresponding author

(a) RGB (b) Ground truth (c) (Liu et al., 2015) (d) (Laina et al., 2016) (e) SGM (ours)

Figure 1. Samples (rows) from the Middlebury dataset comparing ground-truth depth maps (b), SIDE predictions (c+d), and SGM depth maps (e)

increasingly relevant in Computer Vision, leading to ever improving results. As one of the first related works, (Liu et al., 2015) proposes a deep convolutional neural field (DCNF) for depth prediction combining CNNs and *conditional random fields (CRFs)* in a unified framework on a superpixel level. By training a CNN on a large dataset of ground-truth depth maps and corresponding RGB images (Silberman et al., 2012), the network demonstrates the superiority of deep features over hand-crafted ones. (Laina et al., 2016) proposes a *fully convolutional network (FCN)* for depth prediction. The convolutional layers from pre-trained networks, such as AlexNet (Krizhevsky et al., 2012), VGG (Simonyan, Zisserman, 2015), or ResNet (He et al., 2016) are fine-tuned on RGB-D image pairs, while the fully connected layers are re-learned from scratch. Additionally, up-convolution blocks are introduced to address the decrease of resolution of the output maps due to repeated pooling operations. A visual comparison of depth maps generated with different SIDE methods and a standard SGM for samples of the Middlebury 2014 stereo dataset (Scharstein et al., 2014) are shown in Figure 1.

The proposed method is not limited to any particular SIDE methods and can therefore be combined with various current state-of-the-art approaches. In the following, the methods of both (Laina et al., 2016) and (Liu et al., 2015) will both be used for boundary estimations for mdaSGM.

As stated previously, the disparity value sought by SGM is inversely proportional to physical depth. This can be exploited by converting the depth estimation from the neural network into a pseudo-disparity estimation. The minimum and maximum of these pseudo-disparities can then be used to restrict the range needing evaluation by the SGM algorithm, as depicted in Figure 2. This task is not trivial, as the network does not always correctly estimate depth and outliers can still occur. The first problem is attributed to the neural-net depth estimation itself, while the second can be approached using optimized boundary estimation.

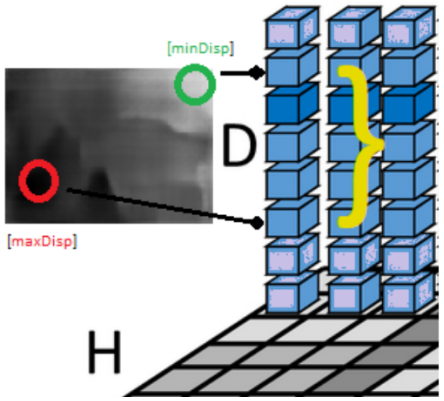


Figure 2. Disparity restriction from neural net estimation. Adapted from: (Hu et al., 2015)

3. PROBLEMS AND LIMITATIONS OF SGM

The problem of tackling disparity range estimation represents the main goal of mdaSGM. There are, however, other limitations on SGM approached by the method which warrant discussion.

3.1 Path Indexing

One such problem is the indexing of paths along which the most-likely disparities are calculated. The standard SGM method uses eight paths (in some implementations even 16). In order to evaluate the disparity pairs along all paths, a consistent indexing system must be conceived to address all pixels from all paths containing them. The naive approach calculates all eight full-length paths for all pixels in the master image. This results in both empty and partially populated paths, as shown in Figure 3. The proposed method utilizes case-by-case differentiation, only creating indices that will be populated, reducing computational effort by approximately 20%, regardless of the disparity range evaluated.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-IV-2-W7-55-2019 | © Authors 2019. CC BY 4.0 License.

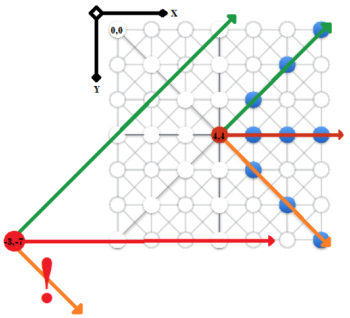


Figure 3. Problem: evaluation of nonexistent paths.
Adapted from: (Drory et al., 2014)

3.2 Dimensionality

Another problem lies in the high dimensionality and resulting complexity of the method. The tensor of SGM is ultimately a 4D matrix (*X* x *Y* x *Disp* x *Paths*), leading to a fourth degree complexity and long (CPU) runtime. The method approaches this problem through parallelization, resulting in eight 3D tensors that can be evaluated simultaneously, versus the sequential evaluation of a single 4D tensor. This poses the most significant time-efficiency gain of the method, but is seen as a more general optimization problem rather than a mono-depth expansion.

3.3 Resolution

On the neural-net side of the method, the resolution of the mono-depth maps is another limiting factor. Depending on the network used, the resolution of the SIDE image can be substantially lower than that of the input image. This distorts the mathematical relation between disparity and depth, leading to a systematic falsification of the output depth maps. Additionally, some mono-depth estimating networks crop the image as a means of border handling, leading to a further falsification in scale, as well as a loss in scene content and with it the elimination of potentially important depth constraints.

The first network used in development of the method cropped and resized the input image, leading to these errors and a vastly incorrect depth map. This could be partially compensated for by means of multiple rescaling factors, yet proved to be unstable. By switching to full-resolution networks with proper border handling, such as those from *Liu* and *Laina*, this problem has been alleviated.

4. mdaSGM

To approach the limitations mentioned before, an SGM algorithm has been implemented, initially transposed into Python from Hirschmueller's 2008 MATLAB script. Since then it has been comprehensively rewritten from a script into an I/O controller making use use of a newly written library which accomplishes SGM. The method is in its essence completely identical to classic SGM, only with new, variable parameters derived either empirically from observation or statistically from mono-depth estimations. The input to mdaSGM is either a single image pair, or an entire dataset. Additionally, the mono-depth estimation for the scene is required as an input for each image pair, forming the final input triplet to the algorithm.

From the mono-depth estimation, the minimum and maximum disparities are identified and passed as boundary parameters to the ensuing matching process. Matching is then done using the optimized parallel approach, returning a full-resolution depth map of equal quality of standard SGM, given correct boundary estimations. Accuracy also depends upon the quality of calibration info, which is required for the depth-to-disparity conversion. While the earlier mono-depth network with rescaled mono-depth maps made this calculation unstable, the more recent full-resolution depth maps eliminate this issue.

4.1 Parameters and Constraints

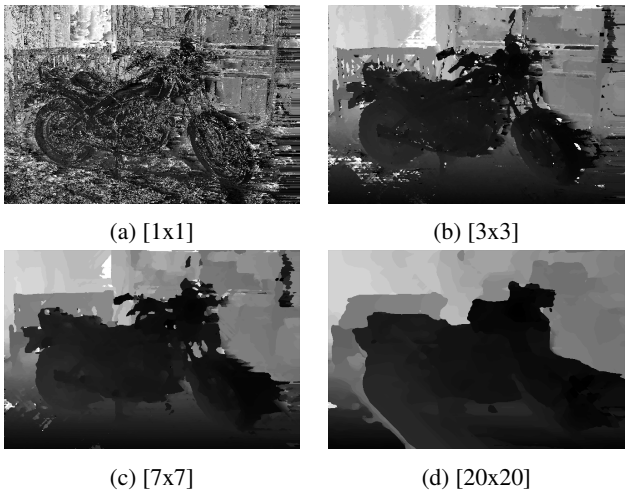
Before the algorithm can be usefully applied to an image pair or benchmarked on a dataset, sensible boundary conditions must be established. These conditions are primarily given by the hyper-parameters of the SGM algorithm. These parameters affect the outcome of matching in distinct ways, meaning they need to be investigated and constrained to plausible and useful values. By doing so, the focus of the evaluation can be shifted towards the quality and performance of the algorithm itself. The main hyper-parameters to be set are as follows:

4.1.1 Aggregation block size Rather than evaluating individual elements of the 4D cost matrix (3D with parallelization), a convolutional summing block is applied across it. The size of this block determines the homogeneity of the resulting depth map. If this block is very small, including a reduction to [1x1], the resulting map displays large amounts of noise. If it is too large, the image becomes over-blended. This effect is illustrated in Figure 4, with multiple block sized being applied to the same image. The optimum for the evaluated Middlebury data appears to lie between five and seven, and seven will be used as a constant for the following experiments.

4.1.2 Path number and direction The standard SGM approach utilizes eight paths, yet results show similar quality with as little as three paths. Given a camera setup along a baseline parallel to the the pixel coordinate system, as well as perfectly rectified images, then decent results can be obtained from one path alone. Due to the fact that all Middlebury images are precisely calibrated and utilize a baseline exactly in X direction, the gains of evaluating multiple paths diminish. They do not vanish entirely, however, as a higher number of paths also serves to increase homogeneity by reducing the chance of false matches. This is demonstrated in Figure 5. In benchmarking, the three path method has demonstrated similar results to the standard eight. Nonetheless, the use of multiple paths is central to the concept of SGM, which is why the three, six and eight path variants will be used for evaluation.

4.1.3 Disparity range This is the most central hyper-parameter to mdaSGM, since it is the one derived from additional neural-net information. The naive approach evaluates all disparities from zero or one (zero theoretically implies infinite distance), through a conservative upper bound. One of the main goals of mdaSGM is lowering this bound as far as possible without cutting out true disparity values in the scene. Figure 6a shows a reference of an ideal SGM disparity map, depicting a Motorcycle in a garage, made using an ideal disparity range extracted directly from the Middlebury information. In the following, several disparity range failure modes will be discussed using maps of identical intensity scaling.

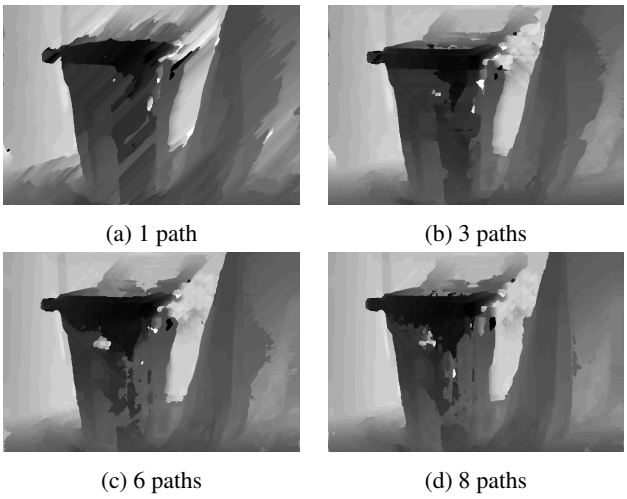
If the upper disparity bound is reduced too far, mismatches and non-matches will occur, resulting in the nearest objects being distorted or missing from the depth map. This is seen



(a) [1x1] (b) [3x3]

(c) [7x7] (d) [20x20]

Figure 4. mdaSGM: influence of sum aggregation block size



(a) 1 path (b) 3 paths

(c) 6 paths (d) 8 paths

Figure 5. mdaSGM: influence of path selection

in Figure 6b, where the parts of the motorcycle nearest to the camera fail to match (black areas). If the bound is set too high, calculation costs will rise exponentially and results will also be negatively influenced, as shown in Figure 6c. This is due to the the fact that nonexistent disparity ranges are occasionally assigned to pixel pairs as a result of the coincidental repetition of gray values. Since cost aggregation is conducted across entire paths, containing both true and false disparity values, the outcome is that the resulting arguments of aggregated disparities are forced to a lower value across the entire image. This can be interpreted as an equilibrium problem and demonstrates the importance of setting this boundary value properly.

The inverse holds true for the lower disparity bound. By beginning naively at zero or one, it is guaranteed that no lower bound disparities will be missed (meaning objects at or near infinity). Since objects never lie at infinity, and SGM is conducted primarily in the near field, it is safe to assume that a disparity of zero will not occur. It is therefore sensible to have mdaSGM estimate this lower bound as well, further reducing operations and processing time needed. An underestimation of this bound is usually not as critical as with the upper bound, since the theoretical minimum disparity of one is usually not far off the true minimum disparity. Nonetheless, the same shifting effect discussed above can become present if the underestimation is too large. An overestimation of the lower bound will in turn lead to the furthest objects in the image being cut out, as shown in Figure 6d, where the motorcycle is still visible but the more distant background can no longer be matched.

With this information it becomes clear that conservative boundary overestimation is no longer an acceptable solution, both in terms of quality as well efficiency. During the evaluation, the estimated disparity ranges will be compared with the ground-truth values from information files to determine how well this estimation is working on each image.

5. EXPERIMENTS

Now that the boundary conditions and main failure modes of the method are known, several experiments can be conducted. To prepare, the mono-depth predictions are acquired utilizing the methods of (Laina et al., 2016) and (Liu et al., 2015) for each image pair of the Middlebury 2014 benchmark dataset. Disparity maps are then calculated from our baseline SGM implementation (identical to Hirschmueller, 2008), as well as from the modified mdaSGM method using both Liu and Laina mono-depth estimations. These can then evaluated in accordance with different quality metrics and compared to the ground truth disparity maps provided by Middlebury.

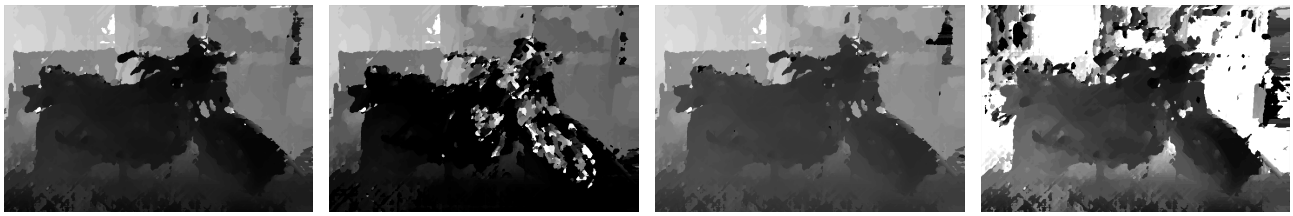
5.1 Operations vs. paths

The first adaptive component of mdaSGM is the number of paths, which is no longer fixed at 8, as with standard SGM, but rather variable at will. If image pairs are perfectly rectified along a known baseline, a reduction in the number of paths needing evaluation is conceivable. This simple extension allows for significant reduction in operations needed, at a negligible quality penalty. As the number of grayscale operations is linear with regard to the number of paths evaluated, then reducing this number will lead logically result in a linear reduction in operations needed. In testing it has been demonstrated that a reduction from eight to three paths quickly amounts to a savings of tens of millions of operations. In the older single threaded version of the method this amounts to an equally drastic increase in runtime, quickly amounting to several minutes. When using the later multithreaded version of the method instead, this penalty becomes negligible, as each processor / thread can simultaneously evaluate its own path.

Figure 7 shows the quality of the resulting disparity maps as a function of the number of paths evaluated, as denoted by the average per-pixel disparity error, as well as the ”Bad5” error metric (percentage of disparities off by 5 or more). It can be seen that as the number of paths decreases, both the average pixel error and ”Bad5” error increase, but only very slightly. Nonetheless, as the quality penalty is negligible when considering the vast savings in computational efficiency, the three-path method proves itself to be an efficient approach for faster depth estimation, or for use on less powerful hardware.

5.2 Operations vs. disparity range

As with the number of paths, the disparity range also has a profound impact on the number of operations. Unlike with the path number, however, the impact of the disparity range cannot be as easily mitigated with multithreading. Figure 8 shows the decrease in operations as the bounds move from naive overestimations toward ground-truth bounds. The impact



(a) Good bounds (b) Upper bound too low (c) upper bound too high (d) lower bound too high

Figure 6. Effect of incorrect disparity bounds

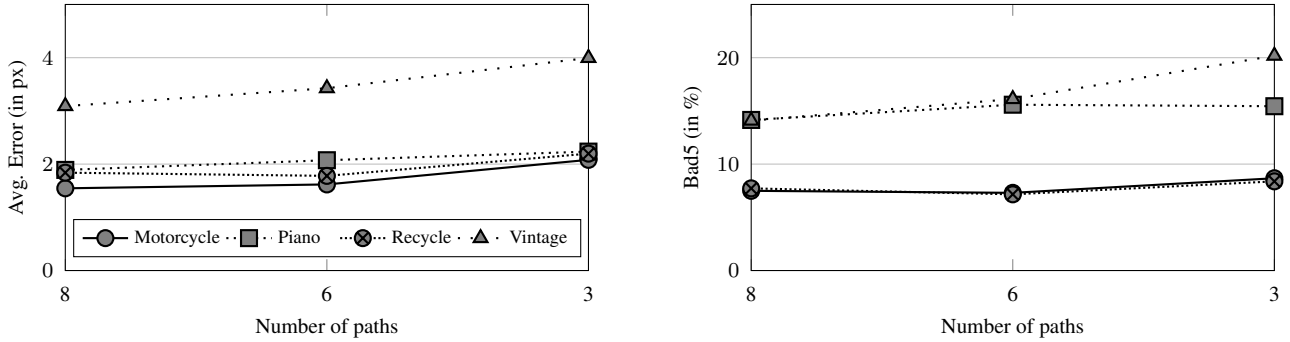


Figure 7. Matching quality as a function of paths for (left) average pixel error and (right) "Bad5" error on samples from the Middlebury 2014 benchmark dataset

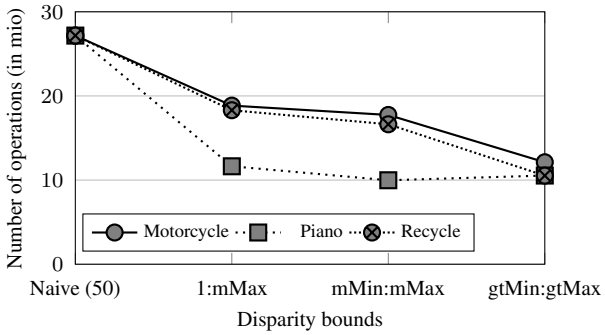


Figure 8. Number of operations as a function of disparity range

of poorly selected bounds is shown previously in Figures 6a and 6d. It can be summarily stated that the minimum number of operations is achieved if and only if the disparity bounds correspond exactly to ground-truth. This is nearly impossible, which is why excellent mono-depth estimation is absolutely key to optimizing mdaSGM.

5.3 Disparity range estimation

Since its inception mdaSGM has gone through three permutations of disparity range estimation. In the following, all three methods will be discussed and evaluated.

5.3.1 Raw Pixel The first mono-depth disparity range estimation method implemented is a raw pixel min/max identifier, simply finding the absolute maximum and minimum values in the mono-depth image and converting these to pseudo-disparities. While this method is computationally simple and generally captures the true disparity range, it demonstrates a tendency towards overestimation and is unstable with regard to outliers. In order to tighten the estimated bounds and improve stability a different method is required.

5.3.2 Median Filter The second method is largely identical to the first, but employs a convolutional median filter in order to reduce susceptibility to outliers. Experimentation shows that this method has the effect of squeezing the boundaries of the disparity range proportionally to the convolution mask size. This proves helpful in some image sets, yet detrimental in others. An additional drawback is the overhead computational time required in order to conduct median filtering. While an improvement over raw-pixel identification, yet another method is still needed.

5.3.3 Histogram Evaluation The third and final method developed within the bounds of mdaSGM is one based on the histogram of the pseudo-disparity maps calculated from the mono-depth images. By forming these disparity histograms, a better understanding of the distribution of disparities can be ascertained, leading to a better boundary estimation. Figure 9 shows mono-depth and ground-truth disparity maps, as well as their respective disparity histograms. By analyzing these histograms, outliers can be removed and better boundaries applied. The method accomplishes this by simply discarding the upper and lower one-percentiles and selecting the boundaries as the arguments of pseudo-disparity at these locations of the histogram.

Applied to a subset of the Middlebury 2014 dataset, the different boundary approaches result in the boundaries shown in image Figure 10. The ground-truth disparity boundaries are shown in black. Method one (blue) sets conservative, yet wasteful boundaries. Method two (green) shrinks these boundaries, yet in doing so occasionally cuts into the true disparity range and wastes processing time performing median-filtering. Method three (orange) is less computationally wasteful, and generally lines up decently with ground-truth disparities. The method can still fail, but usually only where the other two methods fail as well. Given its good performance, this method has been selected as standard for disparity range estimation.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-IV-2-W7-55-2019 | © Authors 2019. CC BY 4.0 License.

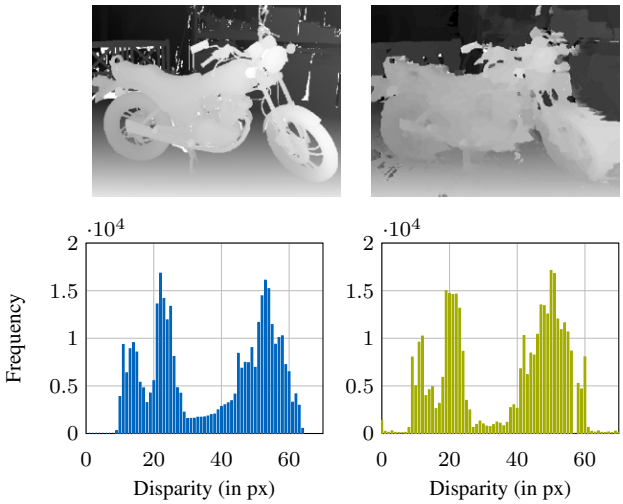


Figure 9. Disparity histograms for ground truth (left) and mono-depth prediction (right) for Motorcycle sample

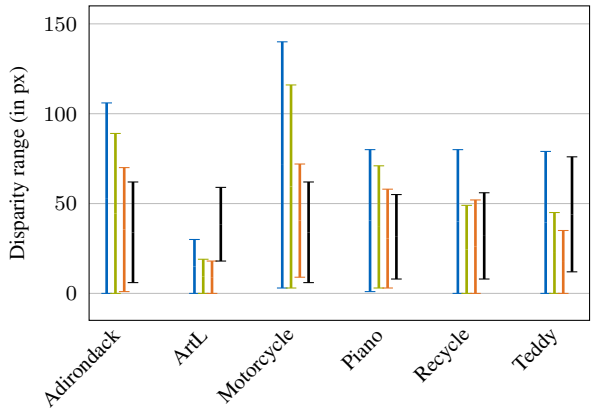


Figure 10. Comparison of disparity range estimation for raw pixels (—), 9x9 median filter (—), and histogram-based (—) w.r.t. the ground truth disparity range (—)

6. BENCHMARKING

With results from the experiments allowing for a more targeted matching approach via path number restriction, multithreading and optimized path estimation, the algorithm has been benchmarked using the MiddEval v3 Evaluation tool. The results from this tool are presented for the entire Middlebury 2014 benchmarking dataset, as well as for three individual examples.

6.1 Middlebury 2014 Full Dataset

Table 1 shows the following quality metrics: percentile of pixels with a ground-truth disparity deviation of more than 1, the equivalent percentile for deviations over 5, over 10, as well as the average per-pixel error and RMSE. Additionally, the number of operations is also introduced as a metric in order to quantify the efficiency of the method. The values shown are averages taken across the entire dataset. With these six metrics, the performance of the algorithm can be more objectively assessed. The top block shows three naive SGM approaches, whereby the disparity range is estimated and set to a fixed span, as well as a fourth "oracle" method using ground truth disparity ranges from calibration files. This latter method represents a quasi-optimum, whereby only truly occurring disparities are evaluated.

From a first glance it can be seen that the "Bad1" percentage is smaller for mdaSGM (for both SIDE methods) than for any of the three naive SGM approaches. Only the optimum "oracle" method performs better, which is to be expected since it is making use of ground-truth data otherwise not available to the algorithm. When proceeding to the "Bad5" metric, it can be seen that values become slightly higher than for the naive approach, a trend which continues over to "Bad10". This implies that, while mdaSGM calculates more disparities correctly than the standard method, once disparities are no longer correct, their deviations tend to become slightly higher. It should be noted, however, that these average values are detrimentally affected by individual bad examples, where both SIDE methods return vastly incorrect disparity estimations. In the following section, these problems will be further elaborated upon.

Another notable difference in the output quality is the dependency on the respective SIDE method used in calculation. According to our evaluation, depth predictions using the method of (Liu et al., 2015) provide more accurate disparity ranges than the method of (Laina et al., 2016). This is reflected in all six metrics, leading to the conclusion that the Liu method is better suited for mdaSGM than Laina (though it should be noted that neither method was originally designed for this task).

With regard to calculative efficiency it is also shown that mdaSGM (with Liu SIDE predictions) consistently performs better than the naive approach, with the exception of SGM(1-40). As most datasets have upper disparity bounds of at least 60, this method usually return very inaccurate disparity maps, as shown by the metrics in the top row of the table. A better comparison can me made to SGM(1-100), which will usually capture the entire disparity range of the scene. Even here, mdaSGM + Liu offers greater calculative efficiency. Figure 11 summarily compares the relation of performance and operational complexity between the baseline SGM with fixed disparity ranges and dynamic disparity ranges from mdaSGM for the entire dataset. The figure demonstrates a decent compromise between the number of operations and resulting quality for both mono-depth prediction methods utilized in mdaSGM.

6.2 Analysis of Individual Image Pairs

Beyond analyzing the average quality metrics from MiddEval it should also be discussed where the method does a good job of estimating and calculating disparities, as well as where it fails. By understanding the main failure modes of the method, further considerations for SIDE methodologies, as well as improvements for mdaSGM itself, can be made. To this end, three examples will be analyzed.

6.2.1 "Piano" An example of good disparity estimation and calculation is given by the "Piano" image pair. For this image pair, the ground-truth disparity range is given with [9-54], while the mono-depth estimation is [3-57]. This implies that the SIDE method (in this case Liu) correctly estimates the depth span and thus the disparity range of the pictured scene. As a result, the method returns a good disparity map (Figure 12, top-right) with "Bad10" error of 5.4%, as well as an average pixel error of 3.3 px. These values lie well under the full-set average of 23.8% and 9.7 px, respectively. By comparison, a naive fixed-disparity calculation per SGM(1-200) returns a "Bad10" error of 11.7% and an average pixel error of 8.3 px. This is attributed to the effects of overestimation as discussed in section 4.1.3.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-IV-2-W7-55-2019 | © Authors 2019. CC BY 4.0 License.

Table 1. Quantitative comparison of baseline SGM and mdaSGM applied on the Middlebury 2014 stereo benchmark. Disparity ranges for baseline methods either fixed or have access to ground truth disparity maps. mdaSGM utiziles dynamic disparity ranges. Results are listd w.r.t. to different error metrics

Method	Bad 1 ↓ (in %)	Bad 5 ↓ (in %)	Bad 10 ↓ (in %)	Avg. Err ↓ (in px)	RMSE ↓ (in px)	NOPS ↓ (in mio)
SGM(1-40)	75.5	40.3	28.6	10.4	16.0	36.1
SGM(1-100)	67.4	22.8	15.5	6.1	12.0	91.7
SGM(1-200)	68.5	26.5	19.9	16.0	35.7	184.4
SGM(oracle)	41.8	19.7	11.5	3.8	7.8	61.4
mdaSGM(Laina)	57.0	41.3	36.0	23.3	32.1	100.3
mdaSGM(Liu)	52.7	31.4	23.8	9.7	17.2	73.7

6.2.2 "Playtable" On the contrary, the "Playtable" image set demonstrates an example of poor disparity map calculation, owing to a failure mode caused by an underestimation of the upper disparity bound, also discussed in section 4.1.3. In this case, the ground-truth disparity range is given with [6-67], while the mono-depth estimation returns [1-38], well below the true upper bound. The resulting disparity map (Figure 12, center-right) is therefore spotty, losing information on objects closest to the camera. This is also reflected in the poor values for "Bad10" at 24.1% and average pixel error at 8.8 pix. Here the naive SGM(1-200) method returns a disparity map with values of 7.6% and 4.1 px, respectively.

6.2.3 "Vintage" Complementary to "Playtable", the "Vintage" image pair demonstrates the effect of overestimating the lower disparity boundary. For this scene, the ground-truth disparity range is given with [8-180], while the prediction is [44-162]. The disparity map (Figure 12, bottom-right), returns a "Bad10" error of 13.1% and an average pixel error of 4.0 px. By comparison, the SGM(1-200) method returns a map with values of 14.5% and 5.5 px. Even though the lower bound is significantly off (and the upper bound slightly off), the resulting map is still superior than the one from the naive method, and better than the average across the entire benchmarking set. Nonetheless it is still inferior to the "Piano" example, which can be expected due to the bounds being too tight.

These three cherry-picked examples demonstrate where mdaSGM and its underlying SIDE predictions perform excellently, adequately, as well as poorly. By understanding this behavior, as well as the main failure modes of the method, further insight can be gained into improving the absolutely critical task of disparity range estimation.

7. CONCLUSION AND OUTLOOK

Semi-Global Matching is a fundamentally simple yet computationally costly approach for disparity determination from raw grey value differences of calibrated stereo images. The resulting disparity maps implicitly allow for depth mapping and have become a common tool in machine vision. The main problem of SGM and its derivatives still lies in the determination of which disparity range to evaluate. This range can be conservatively overestimated using calibration information, but at a quality and efficiency penalty. In order to minimize this penalty, mdaSGM presents a method for using ancillary information from a neural network, allowing the inference of sensible disparity ranges. This has led to a significant reduction in the number of computations required, as well as an improvement in the quality of results (assuming disparity ranges are overestimated in the classic approach). The method performs adequately on a CPU, processing a mid sized image triplet in well under a minute, versus over three minutes for standard, overestimated SGM.

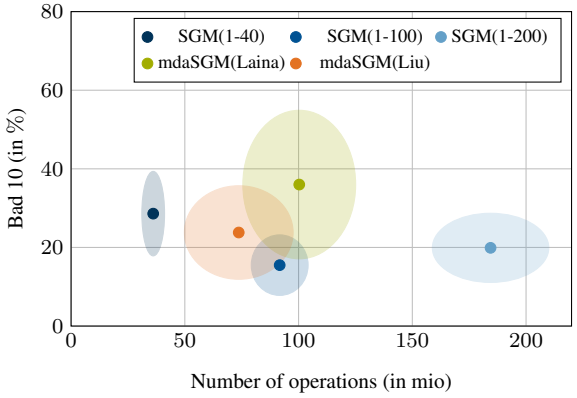


Figure 11. Relation between performance and computational complexity of mdaSGM and baseline SGMs for the Middlebury 2014 Stereo benchmark. Mean and standard deviation are represented by dots and ellipses

Looking forward, area-based disparity restriction is a topic of interest. By determining the approximate disparity on a pixel-wise basis, the length of paths needing evaluation can be made variable, depending on the area of the image currently being analyzed. This means that areas of the scene closer to the camera can be evaluated across a higher range than those in the background, where it is "known" from SIDE information that the disparity will be lower. The caveat of this is idea is the quasi geometry-free, energy-flow nature of the problem, where the goal is finding as the minimum argument of an aggregated sum across multiple dimensions. Varying the disparity range and thus the tensor across the image would prove more difficult to handle, yet will be the subject of upcoming research. Finally, adaptation to a GPU framework and object-oriented language will have an exponential impact on efficiency, with the high-dimensioned matrix calculation being well parallelizable on GPUs with the efficient use of pointers versus variable storage.

Despite being over a decade old, Semi-Global Matching still holds is weight to this day and new optimizations are continuously being found. With mdaSGM, an optimization for disparity range estimation has been presented, from which both classic SGM and related machine vision algorithms can profit going forward.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-IV-2-W7-55-2019 | © Authors 2019. CC BY 4.0 License.

(a) RGB image (b) Ground truth disparity (c) Baseline SGM (d) mdaSGM

Figure 12. Qualitative results of mdaSGM on different samples (from top to bottom: Piano, Playroom, Vintage) from the Middlebury 2014 Stereo dataset (Scharstein et al., 2014) showing ground-truth disparity maps (b) and disparity maps derived from the baseline SGM (c) and mdaSGM (d), respectively. The method of (Liu et al., 2015) was used for mono-depth predictions in mdaSGM

REFERENCES

Drory, A., Haubold, C., Avidan, S., Hamprecht, F. A., 2014. Semi-global matching: A principled derivation in terms of message passing. *German Conference on Pattern Recognition (GCPR)*, 43–53.

Eigen, D., Puhrsch, C., Fergus, R., 2014. Depth map prediction from a single image using a multi-scale deep network. *Advances in Neural Information Processing Systems (NIPS)*, 2366–2374.

Geiger, A., Lenz, P., Stiller, C., Urtasun, R., 2013. Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research*, 32(11), 1231–1237.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Hirschmueller, H., 2005. Accurate and Efficient Stereo Processing by Semi-Global Matching and Mutual Information. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hu, H., Rzhanov, Y., Hatcher, P. J., Bergeron, R. D., 2015. Binary adaptive semi-global matching based on image edges. *Seventh International Conference on Digital Image Processing (ICDIP)*, 9631, International Society for Optics and Photonics, 96311D.

Koch, T., Liebel, L., Fraundorfer, F., Körner, M., 2018. Evaluation of cnn-based single-image depth estimation methods. *European Conference on Computer Vision (ECCV)*, Springer, 331–348.

Krizhevsky, A., Sutskever, I., Hinton, G. E., 2012. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems (NIPS)*, Curran Associates, Inc., 1097–1105.

Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N., 2016. Deeper depth prediction with fully convolutional residual networks. *Fourth International Conference on 3D Vision (3DV)*, 239–248.

Liu, F., Shen, C., Lin, G., 2015. Deep convolutional neural fields for depth estimation from a single image. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Luo, W., Schwing, A. G., Urtasun, R., 2016. Efficient deep learning for stereo matching. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5695–5703.

McCormac, J., Handa, A., Leutenegger, S., Davison, A. J., 2017. Scenenet rgb-d: Can 5m synthetic images beat generic imagenet pre-training on indoor segmentation? *IEEE International Conference on Computer Vision (ICCV)*, 2678–2687.

Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nešić, N., Wang, X., Westling, P., 2014. High-resolution stereo datasets with subpixel-accurate ground truth. *German Conference on Pattern Recognition (GCPR)*, Springer, 31–42.

Silberman, N., Hoiem, D., Kohli, P., Fergus, R., 2012. Indoor segmentation and support inference from rgbd images. *European Conference on Computer Vision (ECCV)*, Springer, 746–760.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations (ICL)*.

Zbontar, J., LeCun, Y. et al., 2016. Stereo Matching by Training a Convolutional Neural Network to Compare Image Patches. *Journal of Machine Learning Research*, 17(1-32), 2.

Revised May 2019