Know Your Limits: Accuracy of Long Range Stereoscopic Object Measurements in Practice

Peter Pinggera[1,2], David Pfeiffer[1], Uwe Franke[1], and Rudolf Mester[2,3]

[1] Environment Perception, Daimler R&D, Sindelfingen, Germany
[2] VSI Lab, Computer Science Dept., Goethe University Frankfurt, Germany
[3] Computer Vision Laboratory, Dept. EE, Linköping University, Sweden

Abstract. Modern applications of stereo vision, such as advanced driver assistance systems and autonomous vehicles, require highest precision when determining the location and velocity of potential obstacles. Sub-pixel disparity accuracy in selected image regions is therefore essential. Evaluation benchmarks for stereo correspondence algorithms, such as the popular Middlebury and KITTI frameworks, provide important reference values regarding dense matching performance, but do not sufficiently treat local sub-pixel matching accuracy. In this paper, we explore this important aspect in detail. We present a comprehensive statistical evaluation of selected state-of-the-art stereo matching approaches on an extensive dataset and establish reference values for the precision limits actually achievable in practice. For a carefully calibrated camera setup under real-world imaging conditions, a consistent error limit of 1/10 pixel is determined. We present guidelines on algorithmic choices derived from theory which turn out to be relevant to achieving this limit in practice.

1 Introduction

Stereo vision has been an area of active research for several decades and applications have found their way into a wide variety of industrial and consumer products. Quite recently, stereo cameras have attracted renewed attention as the central sensor module in modern driver assistance systems, and even in first fully autonomous driving applications [7].

Part of the practicability and performance of modern stereo vision algorithms can arguably be attributed to the seminal Middlebury benchmark study [27], which first provided a comprehensive framework for evaluation and enabled algorithm analysis and comparison. Ten years later, the KITTI project [10] presented a new realistic and more challenging benchmark with stereo imagery of urban traffic scenes, triggering a new wave of improved stereo vision algorithms. These major benchmark studies focus on dense stereo correspondence and are naturally required to provide *both* dense *and* accurate ground truth data. Algorithm performance is mainly judged by the percentage of pixels whose disparity estimates fall within a given accuracy threshold. The threshold is commonly set to several pixels (KITTI), or half pixels at best (Middlebury).

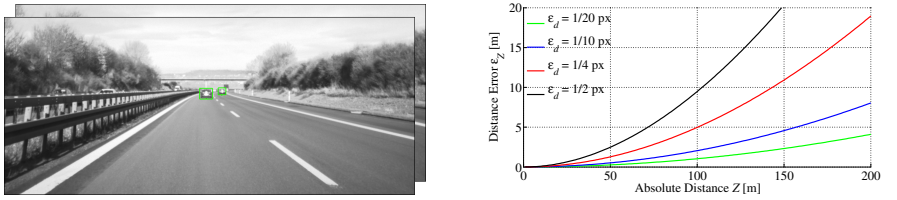


Fig. 1. Highway driving scene with relevant objects at distances of 80 and 140 m (left). Metric distance errors ϵ_Z increase non-linearly for given stereo disparity errors ϵ_d(right)

However, for safety-critical applications such as environment perception in autonomous driving, sub-pixel disparity accuracy is essential. Furthermore, not all parts of the considered images may require the same level of attention. Obstacles in the path of motion are most relevant to the driving task, and their location and velocity have to be determined with maximum precision. Fig. 1 illustrates such critical object locations and the significant impact of sub-pixel disparity errors on the respective distance estimates. Note that for a subsequent estimation of relative object *velocities*, these errors can have an even more serious influence. Unfortunately, this important aspect lies outside the scope of existing major stereo benchmarks, leaving open the question of the actually achievable disparity estimation accuracy where it matters most.

The present paper intends to fill this gap by providing an extensive statistical evaluation of object stereo matching algorithms and establishing a reference for the achievable sub-pixel accuracy limits in practice. We employ a large real-world dataset in an automotive scenario and consider various state-of-the-art stereo matching algorithms, including local differential matching and segmentation-based approaches as well as global optimization in both discrete and continuous settings. Moreover, we investigate possibilities for performance improvement rooted in signal- and estimation theory which are partly used in other areas of computer vision such as medical or super-resolution imaging. Finally, we provide practical guidelines on which algorithmic aspects are essential to achieving the accuracy limits and which are not, also taking into account the trade-off between precision and computational complexity.

2 Related Work

In major dense stereo correspondence benchmarks (Middlebury [27], KITTI [10]) the number of images is kept relatively small for practical reasons, and algorithm performance is derived from pixel-wise match evaluation, weighting each pixel equally. To determine the percentage of erroneous matches, the KITTI benchmark employs a minimum threshold of two pixels. Alternatively, the *average* disparity error on the dataset can be considered, where the top-ranking algorithms at the time of writing achieve a value of 0.9 pixels [33]. This value however provides no information on the matching accuracy for isolated salient objects.

Notably, all top-performing dense methods make use of generic smoothness constraints on the disparity solution, either by global optimization in discrete or continuous disparity space or by integrated image segmentation and parametric model refinement. Taking a closer look at sub-pixel matching precision, it becomes clear that techniques in a discrete setting entail inherent difficulties. Sub-pixel results are obtained by fractional sampling of the disparity space and/or a curve fit to the computed matching cost volume [30]. Depending on the used matching cost measure, these methods usually suffer from the so-called pixel-locking effect, i.e. an uneven sub-pixel disparity distribution. Various approaches have been proposed to alleviate this effect, including two-stage shifted matching [28], symmetic refinement [17], design of optimal cost interpolation functions [11] and disparity smoothing filters [9]. In contrast, methods set in a continuous framework [21] or based on segment model fitting [33] do not suffer from pixel-locking and have been shown to outperform discrete techniques in accuracy.

When shifting the focus from dense disparity maps to isolated objects, the properties of local area-based matching techniques have to be investigated. Within the context of image registration, Robinson and Milanfar [22] presented a comprehensive analysis of the fundamental accuracy limits under simple translatory motion. In low noise conditions, iterative differential matching methods [15] were shown to reach errors of below 1/100 pixels. The corresponding Cramer-Rao Lower Bound (CRLB) for registration errors turns out to be a combination of noise and bias terms, with bias being caused by suboptimal methods for image derivative estimation and image interpolation as well as mathematical approximations. Similar results were reported in [29] for stereoscopic high-precision strain analysis applications. The optimal design of derivative filters and interpolation kernels was also identified as an essential issue in optical flow [26], super-resolution [3], and medical imaging [5, 31] literature.

Perhaps most relevant to the present work is a recent study on local stereo block matching accuracy by Sabater et al. [24]. In contrast to the work mentioned above, realistic noise conditions were investigated and a theoretical formulation for the expected disparity error was derived. Results from a phase-correlation local matching algorithm were shown to agree with the presented theory, demonstrating an accuracy of down to 1/20 pixel on pre-selected pixel locations. However, experiments were performed only on a set of three synthetic stereo pairs and the four classic Middlebury images. Finally, aiming at a more practical automotive setting, in [19] we proposed a joint differential matching and object segmentation approach, yielding errors of 1/10 pixel on actual real-world data. However, our evaluation was also restricted to a very limited amount of sequences.

An important aspect, but outside the scope of the present object-based statistical evaluation, is the data-driven pre-selection of reliable matching points. For local differential methods, matching accuracy can be predicted based on the local image structure [6]. Point selection methods based on various confidence measures have been explored for local [25] as well as global methods [18].

3 Long Range Object Stereo: Algorithm Overview

All algorithms in the present evaluation assume a calibrated stereo camera setup and rectified image pairs. For each relevant object in the scene, a single representative disparity value is determined. This makes sense in the considered scenario, where it is sufficient to model the visible relevant objects as fronto-parallel planes. Note that at large distances, where accurate disparity estimation is actually most important, this model is also valid for more general scenarios.

For the purpose of this study, approximate image locations and sizes of objects are given in advance. Corresponding rectangular patches in the left stereo images are provided as input to the matching algorithms (cf. Fig. 1). Details on the generation of these object patches can be found in Sect. 4.1.

We first define a general stereo matching model by considering the discrete left and right image patch values $I_l(x,y)$ and $I_r(x,y)$ as noisy samples of the observed continuous image signal f at positions (x,y). In this simplified model, $\eta_l(x,y)$ and $\eta_r(x,y)$ represent additive Gaussian noise with variance σ^2, while the shift d denotes the object stereo disparity.

$$I_l(x,y) = f(x,y) + \eta_l(x,y) \tag{1}$$

$$I_r(x,y) = f(x+d,y) + \eta_r(x,y) \tag{2}$$

3.1 Local Differential Matching (LDM)

Iterative local differential matching methods, originally proposed by Lucas and Kanade [15], have proven to perform exceptionally well at high-accuracy displacement estimation [22, 29, 19]. The image difference $I_S(x,y) = I_r(x,y) - I_l(x,y)$ is approximated by linearization and Taylor expansion of f around $d = 0$, with ${f'}_x$ denoting the signal derivative in direction x. Following (1), η now represents Gaussian noise with variance ${\sigma_s}^2 = 2\sigma^2$:

$$I_S(x,y) = f(x+d,y) - f(x,y) + \eta(x,y) \tag{3}$$

$$= d \cdot {f'}_x(x,y) + R_{res}(x,y,d) + \eta(x,y). \tag{4}$$

The disparity d is estimated as the least squares solution to $(d \cdot {f'}_x(x,y) - I_S(x,y))^2 \stackrel{!}{=} 0$, using all pixels of the input image patch. Applying this concept iteratively, the image patch I_r is successively warped by the current estimate of d, and additive parameter updates Δd are computed as described above. This effectively minimizes the influence of the residual R_{res} of the Taylor expansion, and the solution in fact converges to the Maximum Likelihood (ML) estimate. A good initial value for d is required and is commonly provided by a pyramidal implementation. In our sequences, we use a robust global stereo method (Sect. 3.4) for initialization or, if available, the estimation result from the previous frame. In most cases the algorithm converges in less than five iterations. To minimize errors due to global intensity offsets, image patches are mean-corrected before computation.

Table 1. Separable pre-smoothing (left) and derivative filter kernels (right). Complement symmetric and antisymmetric values respectively

Scharr 3×3	$[\ldots, 0.5450, 0.2275]$	$[\ldots, 0, 0.5]$
Scharr 5×5	$[\ldots, 0.4260, 0.2493, 0.0377]$	$[\ldots, 0, 0.2767, 0.1117]$
Central Diff. 5×5	$[\ldots, 0.4026\ 0.2442, 0.0545]$	$\frac{1}{12}[1, -8, 0, 8, -1]$

Image Derivative Estimation. In practice, the signal derivatives ${f'}_x$ in (4) are not known and have to be approximated from I_r using discrete derivative filters. However, inexact derivatives lead to matching bias [22, 3], requiring the use of optimal filter kernels. Jähne [13] derived an optimized second order central differences kernel which requires a separate smoothing step for signal bandwidth limitation. Simoncelli and Scharr [5, 26] on the other hand proposed the joint optimization of pairs of signal pre-smoothing and derivative filters. We investigate both methods, using 3×3 and 5×5 Scharr kernels as well as a 5×5 central difference kernel with a $5 \times 5/\sigma = 1$ Gaussian pre-smoother, cf. Table 1.

Inverse Compositional Algorithm (IC). In [2], Baker et al. presented the so-called inverse compositional algorithm, reversing the roles of the input images and introducing compositional parameter updates to the differential matching framework. This not only lowers the amount of required computations per iteration, but according to [29] also reduces matching bias. The estimated signal derivatives do not have to be warped in each iteration but are computed just once and only at integer pixel positions in I_l, which avoids errors from interpolating derivative kernel responses. In our evaluations, the IC variant is therefore used as the default LDM implementation.

Image Interpolation. Even when using the IC matching algorithm, the iterative nature of the approach still requires warping the image patch I_r in each iteration. Naturally, this step involves the evaluation of intensity values at sub-pixel positions and therefore makes a suitable image interpolation method necessary. In previous studies on image interpolation [31], approaches based on B-Spline representations clearly outperformed simpler methods such as cubic convolution [14] and bilinear interpolation. We investigate the impact of interpolation on disparity accuracy, with cubic B-Splines as the reference method [32].

Estimation-Theoretic Approach (LDM+). Looking at the derivation of the common LDM approach (cf. [15], Sect. 3.1), it can be seen that the algorithm actually computes the ML estimate for the model defined in (3), and not for the original measurement model from (1) and (2). In fact, only (2) depends on the parameter d, leading to the corresponding ML estimate

$$d_{ML} \leftarrow arg\,min_d(I_r(x, y) - f(x + d, y))^2, \tag{5}$$

which involves the unknown signal f. However, a recent theoretic study [16] argued that for the optimal solution of the correspondence problem *both* displacement *and* the unknown signal should be estimated at the same time. We follow this idea and implement a practical algorithm that performs a joint optimization (LDM+). The disparity is computed according to (5), while the signal f is re-estimated in each iteration as the mean of the respectively aligned input image patches. The signal derivatives are computed by applying derivative filters to the current estimate of f. Note that in this case the modifications used in the IC algorithm do not apply.

3.2 Joint Matching and Segmentation (SEG)

Common local matching techniques, such as the LDM algorithm, inherently make the assumption that all pixels in the input image patches conform to a single simple displacement model. Outliers corresponding to a different model can significantly distort estimation results. The approach presented in [19] handles this problem by jointly optimizing the patch shape and the corresponding parametric displacement model. A probabilistic multi-cue formulation integrating disparity, optical flow and pixel intensity is proposed to reliably segment the relevant object from its surroundings. At the same time the iterative approach refines disparity and optical flow parameters in an LDM manner.

We apply the method of [19] but do not make use of the optical flow cue for segmentation. As in the basic LDM algorithm, the results of our SEG algorithm thus depend only on the most recent image pair. After two segmentation iterations, the LDM+ approach as described above is applied for final disparity refinement.

Scene Flow Matching and Segmentation (SEG+). In order to investigate the impact of exploiting the full data from two consecutive stereo pairs, we again follow the approach of [19], but extend it by introducing an additional scene flow segmentation constraint and using all four images for disparity refinement.

In the original formulation, a Gaussian noise model is applied directly to the disparity d and optical flow vectors $\boldsymbol{v}$, which allows for the formulation of probabilistic segmentation criteria by regarding the degraded versions $\tilde{d}$ and $\tilde{\boldsymbol{v}}$ as conditionally independent random variables given $\ell, \mathcal{I}$. Here $\mathcal{I}$ denotes the stereo image data and ℓ the pixel labeling due to the segmentation result.

The scene flow constraint now couples the disparity displacements d between left and right stereo images at time t with the optical flow vectors $\boldsymbol{v}$ between consecutive left images, while the respective degradations due to noise are still considered to be conditionally independent. The constraint is expressed as

$$I_{l,t-1}(x + \tilde{v}_x, y + \tilde{v}_y,) = I_{r,t}(x - \tilde{d}, y). \tag{6}$$

Linearization and Taylor expansion as in (4) and [19] yields

$$I_{l,t-1}(x + \tilde{v}_x, y + \tilde{v}_y,) - I_{r,t}(x - \tilde{d}, y) \tag{7}$$

$$= df'_x(x, y) - \boldsymbol{v}f'_v(x, y) + \eta(x, y), \tag{8}$$

where the noise term $\eta(x, y)$ with variance ${f'}_x^2 \cdot \sigma_d^2 + {f'}_v^2 \cdot \sigma_v^2$ stems from the assumed degradation models of $\tilde{d}$ and $\tilde{\boldsymbol{v}}$. Following [19], the additional random variable $\boldsymbol{w}$ representing the scene flow constraint can be derived from (8). The optimized patch shape is then computed by assigning optimal segment models for pixel intensity i, disparity d and optical flow $\boldsymbol{v}$ under the scene flow constraint $\boldsymbol{w}$, thus maximizing the segmentation likelihood $p(\ell|\mathcal{I}, \boldsymbol{v}, d, \boldsymbol{w}, i)$.

Having obtained an optimized patch shape, for the final disparity refinement step we again resort to the LDM+ algorithm, but now aligning all four input images to estimate the unknown signal f.

3.3 Total Variation Stereo (TV)

As a representative for global stereo matching approaches in a continuous setting, we investigate a differential matching algorithm with variational optimization. Total Variation (TV) based algorithms, originally designed for optical flow estimation [35, 34], have been shown to perform very well in stereo applications [21]. Specifically, we use a total variation Huber-L1 stereo implementation [20] adapted from [35]. The algorithm uses an iterative pyramidal approach to globally optimize an energy of the form

$$E = \int \int \lambda |I_r(x - d, y) - I_l(x, y)| + \sum_{k=1}^{2} |\nabla d_k|_\epsilon \, dy \, dx, \tag{9}$$

where the regularization term $|\nabla d_k|_\epsilon$ penalizes the spatial variation of disparity values, using the robust Huber norm with threshold ϵ. For algorithm details we refer to [35]. We set $\epsilon = 0.01$, $\lambda = 25$ and use five image pyramid levels. For robustness with regard to changes in illumination, the structure-texture decomposition of [34] is applied.

Estimation-Theoretic Approach (TV+). Since the variational approach makes use of the same differential matching principle as the local LDM method on a pixel-wise basis, the estimation-theoretic considerations of the LDM+ algorithm can also be applied. We include a TV+ variant which performs the joint estimation of both displacement and unknown image signal at each iteration. To estimate the required image derivatives, a 3×3 Scharr kernel is used.

Object Measurement. While the resulting dense disparity map provided by the global algorithm is useful for many applications, an additional processing step is needed to arrive at representative disparity values for isolated objects. We compute the interquartile mean of the pixel disparities within the input image patch to obtain a robust object disparity estimate for evaluation.

3.4 Semi-Global Matching (SGM)

Finally, we evaluate the discrete Semi-Global Matching (SGM) algorithm of [12]. The method approximates a two-dimensional optimization with truly global con-

straints by first computing pixel-wise matching costs and then applying one-dimensional regularization along paths from eight directions at each pixel. The nature of the approach allows for efficient computation, and a fast implementation on specialized hardware has been presented in [8].

While all algorithms described above perform matching using image intensities directly, here we employ the census transform and corresponding Hamming distances as a matching cost. This provides a very robust algorithm suitable for challenging real-world scenarios [8]. Sub-pixel results are computed by a symmetric V-fit to three adjacent values in the regularized matching cost volume [11]. Again, we compute the interquartile mean to obtain object disparities.

Pixel Locking Compensation (SGM+PLC). As mentioned previously, matching methods in a discrete setting suffer from the so-called pixel-locking effect, i.e. a biased distribution of sub-pixel disparity values (cf. Fig. 7e). The severity of this effect depends on the used cost metric. While the census transform provides robust matching results, the associated pixel-locking effect is particularly prominent. For general stereo applications, different methods to alleviate the effect have been presented [28, 11]. However, for the scenario at hand we propose a straightforward and efficient post-processing step, which largely neutralizes object-based pixel-locking errors. With ground truth data for the desired object disparities available, the systematic sub-pixel bias can be estimated from a set of raw measurements directly. To this end we project both expected and measured disparity values into the sub-pixel interval $[0, 1]$ and fit a low-order polynomial to the resulting two-dimensional point cloud. This curve is stored and directly provides the necessary offsets for an efficient online correction of the object disparities.

4 Evaluation

4.1 Dataset

A central aspect of the present evaluation is the use of an extensive dataset to allow for a meaningful statistical analysis. Furthermore, we exclusively use real-world data to be able to draw conclusions most relevant for practical applications.

The dataset consists of 70,000 grayscale image pairs recorded from a vehicle-mounted stereo camera system in highway scenarios at mostly sunny weather conditions. It includes approximately 250 unique vehicles representing relevant objects, which gives a total of more than 36,000 disparity measurements. The setup exhibits a baseline of 38 centimeters and a focal length of 1240 pixels, with spatial and radiometric resolutions of 1024×440 pixels and 12 bits, respectively.

We consider disparities between 9 and 3 pixels, corresponding to a distance range of approximately 50 to 160 meters. To also analyze matching accuracy as a function of absolute distance, we divide the overall range into intervals of two meters and evaluate each interval separately. The respective distribution of object observations in the dataset is visualized in Fig. 2.

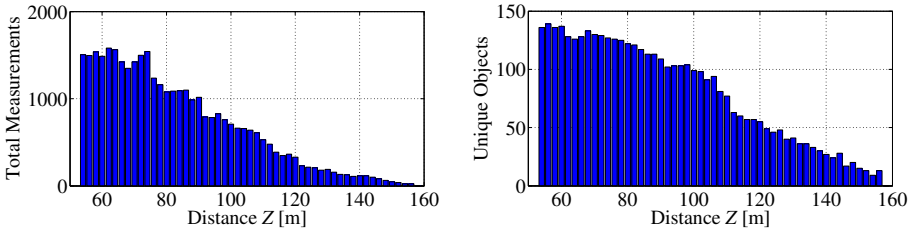


Fig. 2. Distribution of total measurements (left) and unique observed objects (right) in the dataset

Ground truth is provided by a long range radar sensor. Owing to its underlying measurement principle, radar is able to determine longitudinal distances of isolated moving objects with high precision. The used reference sensor yields a measurement uncertainty of $3\sigma \cong 0.5$ m over the full considered distance range.

Generation of Object Patches. To detect relevant objects in the images and provide them as input to the stereo algorithm evaluation, we apply a combined detection and tracking method. A texture-based pattern classifier using a multilayer neural network with local receptive field features (NN/LRF) as described in [4] is used to first locate potential vehicles. These are then tracked over time, accumulating confidence in the process. For evaluation we consider objects which have been tracked for more than 15 frames. The objects are represented by a rectangular patch in the left stereo image, two examples can be seen in Fig. 1.

Note that, before passing the patches to the stereo algorithms, we optimize the patch fit around objects in order to minimize the amount of outlier pixels. We exploit a precomputed dense disparity result to estimate the mean disparity for each patch and decrease the patch size until the number of outliers falls below a given threshold. Subsequently, we shrink the patches by another 25%, except for the segmentation-based approaches, where we actually increase the size again by 25%. To determine the benefit of this adapted patch fit, we also apply the LDM+ algorithm to the *unmodified* patches, denoting this variant as **LDM-**.

4.2 Performance Measures

Disparity Error. The disparity error ϵ_d represents the deviation of the estimated stereo disparity from the ground truth radar value at each frame:

$$\epsilon_d = d - d_{radar}. \tag{10}$$

Temporal Disparity Error Variation. The disparity error ϵ_d as described above provides an *absolute* accuracy measure for all object observations, combining the measurements of multiple unique objects. However, it alone does not provide sufficient information on the *relative* accuracy for a single tracked object

over time. This is essential if the velocities of single objects are to be determined. In this case, the relative accuracy between consecutive measurements of the object of interest is just as important as e.g. a possible constant disparity bias. To describe the object-based relative measurement accuracy over time, we define $\nabla\epsilon_d$ as the disparity error variation using finite differences:

$$\nabla\epsilon_d = \epsilon_{d,t} - \epsilon_{d,t-1}. \tag{11}$$

We examine the distributions of ϵ_d and $\nabla\epsilon_d$ both over the complete dataset and as a function of absolute distance. In addition to robust estimates of the mean, we compute robust estimates of the standard deviation, using the location-invariant and statistically efficient scale estimator S_n of [23].

Runtime. Finally, as an important aspect for practical and possibly time critical applications, we also take the runtime requirements of the various algorithms into consideration. Timings are performed on a subset of the test data, with average unmodified object patch dimensions of approximately 30×30 pixels.

5 Results and Analysis

Table 2 gives an overview of the main quantitative results across the complete dataset. Fig. 3 shows the corresponding distributions of disparity error ϵ_d and error variation $\nabla\epsilon_d$. Examining the mean of the disparity error, it can be seen that the value consistently lies close to -0.1 pixel, varying by less than 1/30 pixel across all algorithms. Fig. 4 illustrates the consistency of this offset across the full distance range. These observations suggest a constant deviation in the stereo camera calibration, most likely caused by a minor squint angle offset.

While this fact illustrates the importance of accurate estimation *and* maintenance of calibration parameters in practice, the location-invariant scale estimates $S_n(\epsilon_d)$ and $S_n(\nabla\epsilon_d)$ provide more meaningful information regarding algorithmic matching accuracy. Note that the mean of $\nabla\epsilon_d$ is exactly zero for all algorithms, other values would in fact imply a temporal drift of the matching results.

Overall, we observe that after optimization of the selected algorithms, the differences in the results for $S_n(\epsilon_d)$ become very small. The best result of approximately 1/10 pixel is achieved by the TV approach, the combination of spatial regularization and mean object disparity estimation performing well across all object observations. However, TV performs worst with regard to temporal error variation, where it does not directly benefit from regularization. The local methods $S_n(\nabla\epsilon_d)$ do best in this category, yielding values as low as 1/20 pixel.

The order of the algorithms in terms of the specified performance measures is largely consistent over the distance range, as shown in Figs. 5 and 6. Given the properties of the used image data, the observed errors roughly agree with the results presented in [24] on synthetic data.

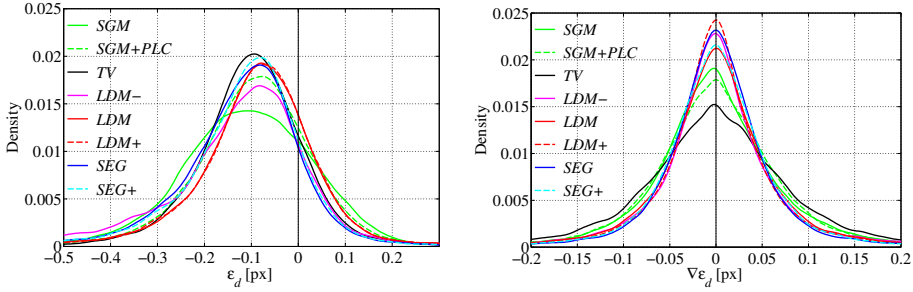


Fig. 3. Overall distributions of disparity error (left) and disparity error variation (right)

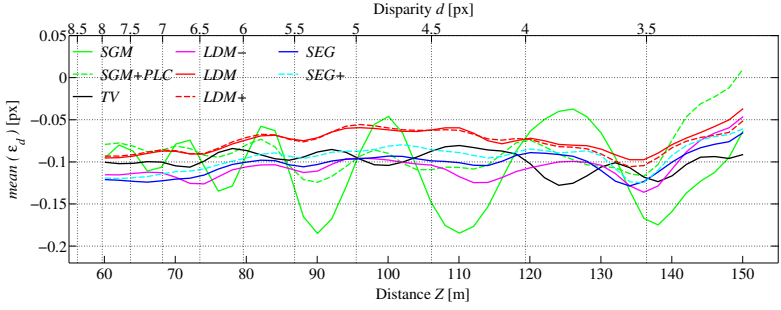


Fig. 4. Mean of disparity error over distance range

Estimation-Theoretic Approach. Now we examine the impact of the estimation-theoretic modifications used in LDM+ and TV+. As can be seen from Table 2 and Figs. 3 and 6, LDM+ yields the same results as LDM for $S_n(\epsilon_d)$, but performs slightly better in terms of error variation. TV and TV+ achieve virtually identical results, the global regularization effectively neutralizing the small differences in data terms.

Optimized Patch Fit. Optimizing the object patch fit has a relatively large impact on the disparity error, as LDM- performs notably worse than all other

Table 2. Overview of quantitative results. See text for details

Method	SGM	SGM+PLC	TV	TV+	LDM-	LDM	LDM+	SEG	SEG+
$mean(\epsilon_d)$[px]	-0.104	-0.090	-0.090	-0.097	-0.110	-0.080	-0.079	-0.109	-0.102
$S_n(\epsilon_d)$ [px]	0.139	0.117	**0.104**	**0.104**	0.135	0.113	0.112	0.114	0.111
$S_n(\nabla\epsilon_d)$ [px]	0.061	0.063	0.077	0.076	0.054	0.056	**0.049**	**0.049**	0.053
t_{avg} [ms]	25	25	65	65	**∼1**	**∼1**	**∼1**	40	80

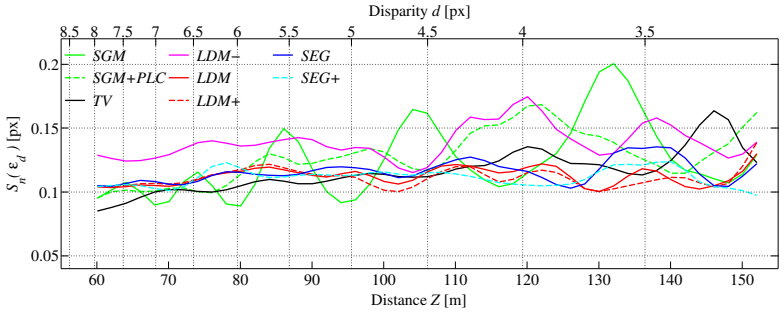


Fig. 5. Standard deviation estimate of disparity error over distance range

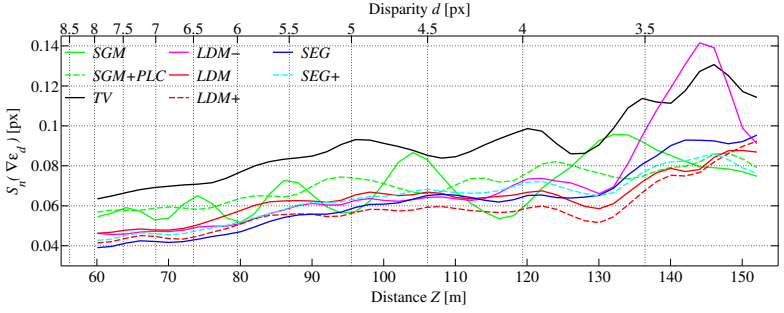


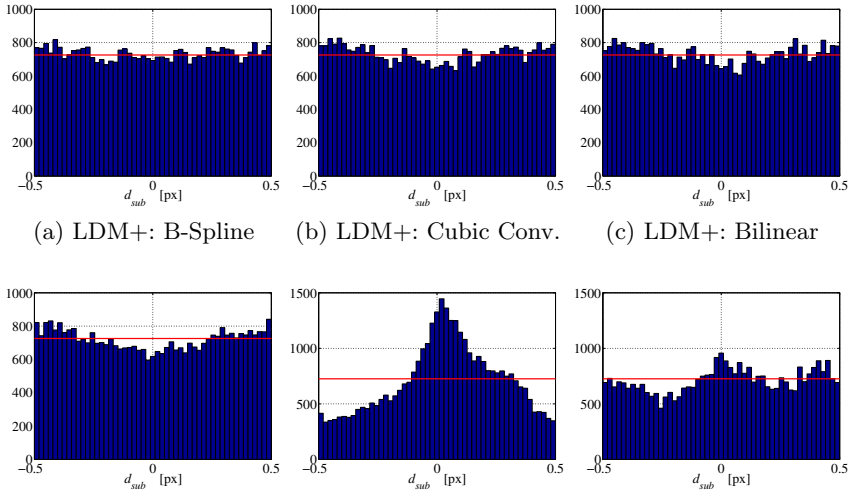
Fig. 6. Standard deviation estimate of disparity error variation over distance range

local algorithms at $S_n(\epsilon_d)$. The error variation scale $S_n(\nabla\epsilon_d)$ without optimized patch fit is also slightly higher than in the otherwise equivalent LDM+ implementation. The efficient adaptation of the rectangular patch fit leads to a similar level of accuracy as the more complex pixel-wise segmentation approaches SEG and SEG+.

Image Derivative Estimation and Interpolation. Interestingly, when comparing different derivative kernels and interpolation methods, we see only insignificant variations in the accuracy results of the differential matching algorithms. The LDM+ column of Table 2 represents our default variant, a 3×3 Scharr kernel and cubic B-Spline interpolation, whereas Table 3 displays the additional configurations. Only when looking at the actual sub-pixel disparity distributions of the different algorithms in Fig. 7, the differences between the interpolation methods become visible. Cubic B-Spline interpolation produces a nearly uniform distribution, while cubic convolution and bilinear interpolation result in a very slight bias towards half pixels. These small variations are in agreement with theoretical predictions presented in [29], but are not distinguishable by our practical disparity accuracy measures at this scale.

Table 3. Impact of derivative kernels and interpolation methods on LDM+ results

Method	Scharr 5×5	Centr. Diff. 5×5	Bilinear	Cubic Conv.
$S_n(\epsilon_d)$ [px]	0.119	0.118	0.114	0.113
$S_n(\nabla\epsilon_d)$[px]	0.054	0.050	0.050	0.050



(a) LDM+: B-Spline (b) LDM+: Cubic Conv. (c) LDM+: Bilinear

(d) TV: Bilinear (e) SGM: V-Fit (f) SGM+PLC: V-Fit

Fig. 7. Sub-pixel disparity distributions resulting from different matching and interpolation methods. Plots show the interval $[-0.5, 0.5]$ centered on full pixel disparities

Pixel-Locking Compensation. In contrast, the systematic pixel-locking effect of the census-based SGM algorithm is clearly visible in both the disparity sub-pixel distribution (Fig. 7e) and in the error measures (Figs. 4, 5, 6). However, applying the proposed compensation method considerably reduces the effect, and SGM+PLC approaches the performance of the differential matching algorithms.

Scene Flow. Finally, our evaluation shows that utilizing the data of two consecutive stereo pairs for scene flow segmentation and disparity refinement as in SEG+ does not necessarily yield a measurable improvement. This might be due to the fact that, in order to align all images, two additional sets of two-dimensional displacements have to be estimated, introducing errors not present in the standard two-image computation. Also, applying a more sophisticated imaging model for estimating the unknown signal could further improve results.

Runtime. Table 2 illustrates average algorithm runtimes, where values for local methods represent the time taken per object, while global methods are timed

per full image. Here, the time to compute object disparities from the dense disparity maps is negligible. Due to the different nature of the algorithms, the provided values are intended to serve as guidance values only. We execute the local methods on a modern four-core CPU while we make use of custom hardware implementations for the SGM (FPGA [8]) and TV (GPU [20]) algorithms.

For the considered patch sizes, the LDM versions vary only insignificantly in runtime and clearly outperform the other methods. The more complex local approaches SEG and SEG+ include an additional outer iteration loop for segmentation and require a significant amount of time for graph-cut based pixel labeling, even when using the speed-up methods of [1].

6 Conclusions

In this paper we depart from the common setting of major dense stereo benchmarks and examine the sub-pixel matching accuracy for isolated salient objects. This is motivated by modern safety-relevant applications of stereo vision, where highest sub-pixel accuracy is required in selected image areas. The presented analysis of various state-of-the-art matching approaches is based on an extensive real-world dataset, enabling meaningful statistical evaluation and providing valuable insights regarding the matching accuracy achievable in practice.

We note that the sole use of the mean absolute disparity error for evaluation proves to be problematic in practice, as even smallest deviations in the camera setup can distort results. We propose the use of robust statistical measures of scale, and additionally introduce an object-based temporal disparity error variation measure which is invariant to systematic disparity offsets. These observations also highlight the need for reliable online self-calibration algorithms.

Appropriate optimization of each selected stereo algorithm minimizes the observable differences in matching accuracy and yields consistent disparity error scale estimates of close to 1/10 pixel. While global variational approaches achieve lowest values here, they perform worst in terms of temporal error variation.

Local differential matching methods perform very well in all performance measures, achieving a temporal error variation scale of 1/20 pixel. Notably, the choice of derivative filter and interpolation method does not have a significant impact on the disparity accuracy here, while optimized patch shapes are essential. Furthermore, optimizations derived from estimation-theoretic considerations can slightly reduce the temporal error variation. Utilizing the full data of two consecutive stereo pairs does not necessarily yield the expected benefits, but shows potential for use with more sophisticated imaging and estimation models.

Pixel-locking effects of discrete matching methods such as SGM, which cause significant errors in sub-pixel disparity, can efficiently be alleviated by an object-based correction approach, moving discrete methods close to differential matching algorithms in terms of accuracy.

References

1. Alahari, K., Kohli, P., Torr, P.H.S.: Dynamic Hybrid Algorithms for MAP Inference in Discrete MRFs. TPAMI 32(10), 1846–57 (Oct 2010)
2. Baker, S., Matthews, I.: Lucas-Kanade 20 Years On: A Unifying Framework: Part 1. IJCV 56(3), 221–255 (2004)
3. Elad, M., Teo, P., Hel-Or, Y.: On the Design of Filters for Gradient-Based Motion Estimation. Journal of Mathematical Imaging and Vision 23(3), 345–365 (2005)
4. Enzweiler, M., Hummel, M., Pfeiffer, D., Franke, U.: Efficient Stixel-Based Object Recognition. In: IV (2012)
5. Farid, H., Simoncelli, E.P.: Differentiation of Discrete Multidimensional Signals. TIP 13(4), 496–508 (Apr 2004)
6. Förstner, W.: Image Matching. In: Haralick, R.M., Shapiro, L.G. (eds.) Computer and Robot Vision, chap. 16, pp. 289–372. Addison-Wesley, 2. edn. (1993)
7. Franke, U., Pfeiffer, D., Rabe, C., Knoeppel, C., Enzweiler, M., Stein, F., Herrtwich, R.G.: Making Bertha See. In: ICCV Workshops (2013)
8. Gehrig, S.K., Eberli, F., Meyer, T.: A Real-Time Low-Power Stereo Vision Engine Using Semi-Global Matching. In: Proc. ICVS (2009)
9. Gehrig, S.K., Franke, U.: Improving Stereo Sub-Pixel Accuracy for Long Range Stereo. In: ICCV Workshops (2007)
10. Geiger, A., Lenz, P., Urtasun, R.: Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In: CVPR. pp. 3354–3361 (2012)
11. Haller, I., Nedevschi, S.: Design of Interpolation Functions for Subpixel-Accuracy Stereo-Vision Systems. TIP 21(2), 889–98 (Feb 2012)
12. Hirschmüller, H.: Stereo Processing by Semiglobal Matching and Mutual Information. TPAMI 30(2), 328–41 (Feb 2008)
13. Jähne, B.: Digital Image Processing - Concepts, Algorithms, and Scientific Applications. Springer, 3rd edn. (1995)
14. Keys, R.G.: Cubic Convolution Interpolation for Digital Image Processing. ASSP 29(6), 1153–1160 (1981)
15. Lucas, B.D., Kanade, T.: An Iterative Image Registration Technique with an Application to Stereo Vision. In: Proc. Int. Joint Conf. on Artificial Intel. (1981)
16. Mester, R.: Motion Estimation Revisited: An Estimation-Theoretic Approach. In: SSIAI (2014)
17. Nehab, D., Rusinkiewiez, S., Davis, J.: Improved Sub-Pixel Stereo Correspondences Through Symmetric Refinement. ICCV pp. 557–563 (2005)
18. Pfeiffer, D., Gehrig, S., Schneider, N.: Exploiting the Power of Stereo Confidences. In: CVPR. pp. 297–304 (2013)
19. Pinggera, P., Franke, U., Mester, R.: Highly Accurate Depth Estimation for Objects at Large Distances. In: GCPR. vol. LNCS 8142, pp. 21–30. Springer (2013)
20. Rabe, C.: Detection of Moving Objects by Spatio-Temporal Motion Analysis. Phd thesis, Christian-Albrechts-Universität zu Kiel (2011)
21. Ranftl, R., Gehrig, S., Pock, T., Bischof, H.: Pushing the Limits of Stereo Using Variational Stereo Estimation. In: IV (2012)
22. Robinson, D., Milanfar, P.: Fundamental Performance Limits in Image Registration. TIP 13(9), 1185–1199 (2004)
23. Rousseeuw, P.J., Croux, C.: Alternatives to the Median Absolute Deviation. Journal of the American Statistical Association 88(424) (1993)
24. Sabater, N., Morel, J.M., Almansa, A.: How Accurate Can Block Matches Be in Stereo Vision? SIAM Journal on Imaging Sciences 4(1), 472 (2011)

25. Sabater, N., Almansa, A., Morel, J.M.: Meaningful Matches in Stereovision. TPAMI 34(5), 930–42 (May 2012)
26. Scharr, H.: Optimal Filters for Extended Optical Flow. In: LNCS 3417 - Complex Motion, vol. 1114, pp. 14–29. Springer (2007)
27. Scharstein, D., Szeliski, R.: A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. IJCV 47(1-3), 7–42 (2002)
28. Shimizu, M., Okutomi, M.: Precise Sub-pixel Estimation on Area-Based Matching. In: ICCV. pp. 90–97 (2001)
29. Sutton, M.A., Orteu, J.J., Schreier, H.W.: Image Correlation for Shape, Motion and Deformation Measurements. Springer (2009)
30. Szeliski, R., Scharstein, D.: Sampling the Disparity Space Image. TPAMI 26(3), 419–25 (2004)
31. Thévenaz, P., Blu, T., Unser, M.: Interpolation Revisited. TMI 19(7) (2000)
32. Unser, M., Aldroubi, A., Eden, M.: B-Spline Signal Processing. TSP 41(2) (1993)
33. Vogel, C., Schindler, K., Roth, S.: Piecewise Rigid Scene Flow. In: ICCV (2013)
34. Wedel, A., Pock, T., Zach, C., Bischof, H., Cremers, D.: An Improved Algorithm for TV-L1 Optical Flow. In: LNCS 5604, pp. 23–45. Springer (2009)
35. Werlberger, M., Trobin, W., Pock, T., Wedel, A., Cremers, D., Bischof, H.: Anisotropic Huber-L1 Optical Flow. In: BMVC. pp. 108.1–108.11 (2009)