Analyzing Computer Vision Data - The Good, the Bad and the Ugly

Oliver Zendel Katrin Honauer Markus Murschitz Martin Humenberger
Gustavo Fernández Domínguez
AIT, Austrian Institute of Technology, Donau-City-Strasse 1, 1220, Vienna, Austria
HCI, IWR at Heidelberg University, Berliner Strasse 43 D-69120 Heidelberg, Germany
{oliver.zendel;markus.murschitz;martin.humenberger;gustavojavier.fernandez}@ait.ac.at,
katrin.honauer@iwr.uni-heidelberg.de

Abstract

In recent years, a great number of datasets were published to train and evaluate computer vision (CV) algorithms. These valuable contributions helped to push CV solutions to a level where they can be used for safety-relevant applications, such as autonomous driving.

However, major questions concerning quality and usefulness of test data for CV evaluation are still unanswered. Researchers and engineers try to cover all test cases by using as much test data as possible.

In this paper, we propose a different solution for this challenge. We introduce a method for dataset analysis which builds upon an improved version of the CV-HAZOP checklist, a list of potential hazards within the CV domain. Picking stereo vision as an example, we provide an extensive survey of 28 datasets covering the last two decades. We create a tailored checklist and apply it to the datasets Middlebury, KITTI, Sintel, Freiburg, and HCI to present a thorough characterization and quantitative comparison. We confirm the usability of our checklist for identification of challenging stereo situations by applying nine state-of-the-art stereo matching algorithms on the analyzed datasets, showing that hazard frames correlate with difficult frames. We show that challenging datasets still allow a meaningful algorithm evaluation even for small subsets. Finally, we provide a list of missing test cases that are still not covered by current datasets as inspiration for researchers who want to participate in future dataset creation.

1. Introduction

Vision solutions are used in safety critical applications such as self-driving cars and guided surgical procedures. Rigorous quality assurance measures are thus needed to ensure safe operations. Software quality assurance provides two main techniques that can be applied in CV: verification and validation (V&V). Verification is the process of checking whether a given implementation fulfills the specifications used to define the program's behavior. In essence these are semi-automatic or automatic checks to detect software bugs and glitches. Validation on the other hand evaluates if the system fulfills a given task even under difficult circumstances. This is done by using test datasets and comparing the results obtained from the system to a defined ground truth (GT). Major questions about the quality and usefulness of test data for CV evaluation are still unanswered: What are the characteristics of a good dataset? How can shortcomings be identified and supplemented to create test datasets which are truly effective at uncovering algorithmic shortcomings? In this work we tackle the question: What constitutes good test data for robustness testing, *i.e*. the detection of possible shortcomings and weaknesses. We show that special care should be taken to cover a wide variety of difficult situations because whether for validation of CV algorithms or for training applications: Datasets need a mixture of positive cases (*the Good*), border cases (*the Bad*), and negative test cases (*the Ugly*). This paper focuses on test data for validating stereo vision algorithms but the presented methodology is applicable to basically all CV algorithms as well as the composition of machine learning training data.

To give an idea about the impact of selected datasets, Figure 1 shows the number of papers which cite stereo vision datasets published annually at three major computer vision conferences (CVPR, ICCV, and ECCV). It is interesting to note that the popular Middlebury dataset (indoor scenes) was recently overtaken by KITTI (driving scenes) which shows the importance of stereo vision in the field of autonomous driving and driver assistance systems.

Section 2 gives a thorough overview and listing of 28 stereo vision datasets and summarizes how content has changed historically. Section 3.1 reviews CV-HAZOP, a tool for systematic analysis of test datasets. It presents our improvements on the method: specialization of generic

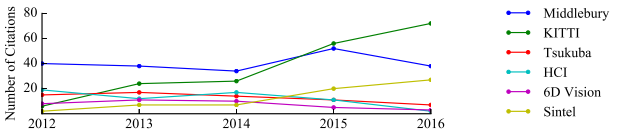

Figure 1. Number of stereo dataset citations published at CVPR+ICCV+ECCV for the years 2012-2016.

entries and instructions for easier analysis using the checklists. We apply the proposed concepts and create a specific checklist of dangerous/risky situations for stereo vision in Section 4.1. We evaluate five representative stereo vision datasets by using the proposed methodology in Section 4.2. In addition a range of stereo vision algorithms is evaluated in Section 4.3 using both traditional metrics and new metrics based on the results obtained by our checklist. Section 4.4 shows that the usage of challenging frames results in a comparable overall outcome even for a small number of test cases. Our checklist contains many critical situations that have not been found in any of the datasets. Section 4.5 presents this useful information for designing future datasets while the lessons-learned are shown in Section 4.6. Finally, Section 5 summarizes all findings and contributions of this paper.

2. State-of-the-Art

Reference data is the basis for performance analysis in computer vision. High-quality data is always well received in the community because it is essential to evaluate algorithm performance allowing the development of more accurate algorithms. Moreover, an objective comparison between algorithms using standardized data is important for a practical understanding of the current state-of-the-art in the respective area. Progress in stereo algorithm performance and the emerging applications of stereo technology motivate the need for more challenging datasets with accurate GT which emerges as a field of research. Among many others, examples of application domains are: autonomous driving (AD) [42, 66, 25, 23, 41, 60], space [24], agriculture [46], and medicine [6, 37, 36]. Early research introduced first datasets and performance metrics to show comparable results on the proposed algorithms. Initially, no common sequences/datasets were adopted. A clear domain or standard performance metrics definition were missing as well. Through the years, the CV community realized that thorough performance evaluation opens many research possibilities such as introduction of new datasets covering different scenarios and situations, analysis of performance metrics or online benchmarks comparing different algorithms. We now present the evolution of stereo vision datasets by comparing 28 datasets of the last two decades[1]. Table 1 gives an overview and presents quantitative characteristics of each dataset while Figure 2 shows representative images. We are focusing on the stereo vision test data. Many datasets contain additional GT (*e.g.* flow, segmentation, instances).

We will not compare datasets that have only RGBD data (no second camera image, *e.g.* NYU RGB-D [63, 44], TUM RGB-D [67] or the Berkeley dataset [22]). Please refer to the recent work of Firman [13] instead. There have been previous surveys on stereo vision and the interested reader is referred to [33, 57, 4, 32, 62, 30, 19].

2.1. Dataset Survey

In 2002 the Middlebury group proposed a taxonomy and a comparison framework of two-frame stereo correspondence algorithms [57]. The Middlebury website [68] evaluates stereo algorithms online, reports the performance of submitted algorithms, and offers stereo correspondence software for download. Over the years, the datasets were regularly updated: 6 datasets of piecewise planar scenes (2001), 32 datasets using structured light (between 2003 and 2006) and 43 high-resolution datasets with subpixel accurate ground truth (2014). EISATS [52] provides different video sequences for the purpose of performance evaluation. Traffic scenario scenes for evaluation of motion analysis, stereo vision, and optical flow algorithms are available to the community. Stereo sequences cover: Night vision (S1), synthesized (S2), color (S3), gray-level (S4&6), trinocular (S5&9), and consecutive stereo image pairs (S7). Neilson and Yang [45] introduced synthetic stereo pairs which were used to show their new evaluation method named cluster ranking. The dataset consists of 30 different stereo pairs containing three different baseline separations and three different noise levels and includes disparity maps and evaluation masks [48]. New College [65] is a large dataset ($\sim$ 30 GB) collected through the parks and campus of Oxford New College. The dataset focuses on outdoor SLAM (Simultaneous Localization and Mapping) applications and includes trajectories, stereo/omnidirectional imagery, as well as laser range/reflectance data. Pittsburgh Fast-Food [8] is a dataset containing 61 categories of food items. It aims to provide standard baselines for evaluating the accuracy of CV algorithms. EVD [9] dataset was developed for evaluating MODS (Matching On Demand with view Synthesis), an algorithm for wide-baseline matching of outdoor scenes but only includes homography data as GT. Ford Campus [50] dataset ($\sim$100 GB) is recorded using a 3D scanner laser and an omnidirectional camera intended for testing SLAM algorithms for AD. In 2012 Geiger et al. [15] introduced the KITTI Vision Benchmark Suite

[1] We tried to include every stereo vision dataset that also publishes GT; some datasets without GT were added due to their popularity.

Table 1. Summary of stereo datasets. 'w. GT' = number of images available with GT data, 'wo' = number without GT data, 'GT-Acc.' = GT accuracy in pixels, † =GT reported but dense GT is not available or the GT is very sparse/semantically oriented) * = algorithm results offered as GT, <1/N = granularity better than 1/N, S = synthetic, R = real, 1 = single shots, 2 = sequences of length 2, N = longer sequences

NAME	YEAR	IMAGES				DESCRIPTION
		Resolution	w. GT / wo	GT-Acc.	Type	
Middlebury [57]	2002	410 x 370	6 / —	1/8	R1	Piecewise planar cardboards
Middlebury [58]	2003	410 x 370	2 / —	1/4	R1	Cluttered still life
Middlebury [21]	2007	1390 x 1110	27 / 3	1	R1	Cluttered still life
EISATS S1 [70]	2008	640 x 481	— / 1900	—	RN	Traffic scenes
EISATS S2 [71]	2008	640 x 480	498 / —	<1/256	SN	Traffic scenes
Neilson [45]	2008	400 x 400	270 / —	1/16	S1	Still scene with var. textures/noise
EISATS S6 [53]	2009	640 x 480	— / 177	—	RN	Traffic scenes
New College [65]	2009	512 x 384	— / >100000	—	RN	Outdoor scenes for SLAM
Pittsburgh [8]	2009	1024 x 768	— / 130	*	R1	Fast food items (61 categories)
EVD [9]	2011	1000 x 750	— / 15	—	R1	Wide baseline still lifes
Ford Campus [50]	2011	1024 x 768	— / >100000	—	RN	SLAM, dynamic environments
HCI-Robust [27]	2012	656 x 541	— / 462	—	RN	Difficult road scenes
KITTI 2012 [15]	2012	1226 x 224	194 / 195	1/256	R2	Suburbs w. little traffic day time
Leuven [31]	2012	316 x 25	20 / 50	†	RN	Traffic day time
Tsukuba [38]	2012	640 x 480	1800 / —	<1/256	SN	Office cubicle still life
HCI-Synth [17]	2013	960 x 540	12 / —	1/256	S1	Texture challenges
Stixel [51]	2013	1024 x 333	2988 / —	†	RN	Highway w. good/bad weather
Daimler Urban [59]	2014	1024 x 440	— / 70000	—	RN	Urban city scenes
Malaga Urban [2]	2014	1024 x 768	— / >100000	*	RN	Dynamic environments real traffic
Middlebury [56]	2014	1328 x 1108	28 / 15	<1/256	R1	Cluttered indoor still life
Cityscapes [10]	2015	2048 x 1024	— / 20000	*	R1	Urban scenes daytime
KITTI 2015 [40]	2015	1242 x 375	200 / 200	1/256	R2	Road scenes with traffic
MPI Sintel [5]	2015	1024 x 436	1064 / —	<1/256	SN	Adventure movie scenes
Freiburg CNN [47]	2016	960 x 540	35454 / —	<1/256	SN	Road scene, animation movie
HCI Training [26]	2016	2560 x 1080	1023 / —	<1/256	RN	Difficult road scenes
SYNTHIA [55]	2016	960 x 720	>100000 / —	<1/256	SN	Diverse driving scenes
Virtual KITTI [14]	2016	1242 x 375	2126 / —	<1/256	SN	Suburban roads, currently RGBD
Oxford RobotCar [35]	To appear	1280 x 960	>100000 / —	<1/256	RN	Driving under varying weather and seasons

Figure 2. Excerpts from the discussed datasets. Images taken from the sources described in Table 1.

which includes a number of benchmarks. Stereo and optical flow data for close to 200 frames are provided. In addition, annotations include semantic and instance labels and longer image sequences of 20 frames per scene and there are about 200 frames where GT is withheld to ensure a fair evaluation on their website. In 2015 an updated

version of the dataset was released containing 400 image pairs of dynamic city scenes (200 for training and 200 for testing) and GT which was semi-automatically generated. Pixels are correctly estimated if the disparity or flow endpoint error is below a certain threshold, either 3 pixels or 5%, and it is required that the methods use the same parameter set for all test pairs. Their focus is on AD with the aim to reduce bias between real data and data generated under controlled conditions, *i.e*. laboratory environments. Objects such as cars and people are visible on each image. The Leuven [31] dataset presents image pairs from two cameras separated 1.5 meter apart from each other. The data was acquired in a public urban environment and contains both object class segmentation and dense stereo reconstruction GT for real world data. Tsukuba [38] dataset is a synthetic photo-realistic video dataset created as an reenactment of their well-known *head and lamp* stereo scene [43]. They include computer generated GT data for parameters, measurements, 3D position and distances. The 6D Vision group [11] makes two different datasets available to the community. The Daimler Urban Dataset [59] consists of video sequences recorded in urban traffic. Five semantic classes are defined (*building, ground, pedestrian, sky, and vehicle*) and 10% of the dataset is pixel-annotated using these classes. The Stixel Dataset [51] consists of 12 annotated stereo sequences acquired on a highway. Vehicle data, camera calibration, and GT generated by a fusing informations from manual annotations with ego-motion estimations are provided. HCI-Synth [17] contains four datasets, each covering a specific issue in stereo vision: visual artifacts, foreground fattening, decalibration, and textureless areas. Malaga Urban dataset [2] was recorded in urban scenarios using 9 cameras and 5 laser scanners containing real-life traffic scenes. The dataset is oriented toward object detection, SLAM, and visual odometry algorithms. The Cityscapes Dataset [10] was gathered entirely in urban street scenes focusing on semantic urban scene understanding. The dataset was recorded across several cities and different seasons. A benchmark suite, an evaluation server, and annotations (detailed for 5000 images and coarse for 20000) are also provided. The MPI Sintel Dataset [5] is derived from the animated short film Sintel containing diverse effects such as scene structure, blur, different illumination, and atmospheric effects. It is designed for the evaluation of optical flow, segmentation and stereo vision. Virtual KITTI [14] is a synthetic video dataset generated using virtual worlds. The scenarios comprise urban settings and the dataset is focused on multi-object tracking. No stereo setup has been released at the time of writing this paper (only RGBD). SYNTHIA (SYNTHetic collection of Imagery and Annotations) [55] is a synthetic dataset collected using 8 RGB cameras and 8 depth sensors. The data was acquired in different scenarios (cities, highways and green areas) under different illumination and weather conditions. The Oxford RobotCar Dataset [35] was collected by driving over the same route in Oxford throughout the year and thus represents good variations in seasons and weather.

2.2. Toward Optimal Test Data

The core problem of test data design is choosing the right number and kind of test cases. Some works in the CV community increased the number of sequences to the hundreds [12, 64, 34], but using more sequences does not necessarily increase diversity or coverage. Besides that, more data requires more GT, and GT acquisition is well known for being an error-prone and tedious task. Many recent works generate synthetic test data, where GT generation is more feasible and accuracy is higher (see [55, 18, 17, 49, 1, 5]). Another problem is dataset bias: test datasets without enough variation cannot reflect real world performance. Thus, researchers have begun to assess the role of diversity, coverage, and dataset bias. Torralba *et al*. [69] analyzed dataset bias, by training image classifiers to learn the dataset they belong to. The VOT challenge [29] performs clustering of a huge pool of sequences to reduce the size of the dataset to be evaluated while keeping in mind the diversity of the selected data. Zendel et al. [74] use a risk analysis procedure called Hazard and Operability Study (HAZOP) to evaluate and improve test datasets. HAZOP identifies difficult situations and aspects present in the dataset showing the hazard coverage of the dataset.

There are three main categories of test cases in traditional software quality assurance: positive test cases, border cases, and negative test cases. Positive test cases [61] represent normality and shall pose no problem to the algorithm. Border cases [7] are on the brink between specified and unspecified behavior but should still create meaningful outputs. Negative test cases [61] are expected to fail, but the error behavior should be well-defined (*e.g*. marking areas without meaningful values as invalid).

In this paper we concentrate on selecting challenging (*i.e*. border and negative) test cases in datasets to improve testing for robustness.

3. Methodology

Now we want to analyze some of the datasets presented in the previous section in depth and evaluate which hazards are tested by these datasets. We propose a new methodology based on an existing idea: Applying risk analysis to CV. First, this quality assurance approach is presented. Then, we extend the methodology. Finally, we apply this method to selected stereo vision datasets in Section 4.

3.1. CV-HAZOP

The systematic analysis of aspects that can influence the output performance and safety of a system is called a risk

analysis. Zendel *et al*. [74] apply a standard risk analysis called HAZOP to generic computer vision algorithms. First, they define an abstract CV model. Its components and their parameters create the basis for the HAZOP study. Then, modifier words called guide words are used to create entries representing deviations from the expected. These deviations are applied to each parameter and lead to a multitude of initial entries for the analysis. CV experts assign meanings, consequences and eventually hazards to each of these initial entries. The resulting list of identified vulnerabilities can be used to evaluate existing datasets and plan new ones. Each list entry can be referenced using its unique hazard identifier (HID). This approach allows qualitative and quantitative evaluation of datasets by identifying individual test cases that satisfy a stated checklist entry. However, there is a shortcoming with the proposed method: In order to have a unified generic checklist, each entry needs to be interpreted by the dataset analysts to their individual opinion. This results in a lot of ambiguity as different analysts might read and interpret the same entry in considerably different ways when applying it to the actual task at hand. Therefore we improve their work in the following aspects:

- Creation of specialized checklists specific to individual use cases instead of having each analyst start with the generic risk analysis lists (see Section 3.2).
- Methodology for analyzing datasets using the specialized checklist in Section 3.3.
- Application of the presented methods by creating a specialized checklist for stereo vision (Section 4.1).
- Analysis of popular stereo vision datasets using the specialized checklist presented in Section 4.3.

3.2. Checklist Specialization

The process starts with the publicly available generic CV-HAZOP checklist and transforms it into a specific one suitable for a particular domain and task:

- Decide for each entry in the list whether the hazards are relevant in the context of the actual task at hand.
- Create a single consensus summary for the entry. Write down as precisely as possible what is expected to be in a test image to fulfill the entry.
- Avoid duplicates and generate a concise list with a minimum of redundancy.

Experience has shown that the resulting list has to be revised after being used by the analysts for the first time. This resolves misunderstandings as well as annotation bias and allows to further remove redundancies.

3.3. How to Analyze a Dataset

The main goal of dataset analysis is usually to find at least one example test image for each checklist entry. This creates a rough estimate of the covered risks. First the analyst has to acquire a general overview of the dataset by noting regularities and reoccurring themes as well as special visually difficult situations such as: light sources (l.s.) visible within the image, visible specular reflections of l.s., large glare spots, large reflections showing near-perfect mirroring, transparencies, overexposure, underexposure, and large occlusions.

Now the specialist tries to find a fitting test image for each entry in the list. The restrictions found at the description are mandatory and reflect the transition from a generic hazard to the specific one. The relevant image part reducing the output quality for the target application should be large enough to have a meaningful impact (*e.g*. 1/64 of the image) and there should be valid GT available at this location. Test cases fulfilling only a single hazard with no overlap are preferred if there are multiple candidates for one entry. Otherwise images having the strongest manifestation of the hazard with largest affected areas are chosen.

4. Results

The presented methodology is applied to the stereo vision use case. A specific checklist is created and used to analyze popular existing stereo vision datasets. A thorough evaluation over a wide range of stereo vision algorithms generates an appropriate background for the following test data analysis. We show correlations between difficulty of test cases and predefined hazards from the checklist, indicate remarks about dataset size, and close with an extensive list of open issues missed in current datasets.

4.1. Stereo Vision Checklist

For our stereo vision checklist we define this use case: Calculate disparity maps from two epipolar constrained images without the use of prior or subsequent frames. The domain for which the algorithms should work is selected with the test datasets in mind: indoor scenes and outdoor driving scenes. We exclude most temporal hazards but otherwise regard all generic entries as potential candidates for our stereo vision checklist. Thus, we start with about 750 generic entries. Many hazards can quickly be disregarded as being out-of-scope for stereo vision. The remaining 350 entries are discussed and specialized. During this process some entries are deemed to be too extreme for our domain and many entries result in duplicates which are already part of the new checklist. At the end we derive 117 specialized entries from the generic list. Table 2 shows an excerpt of representative entries from the full list[2]. Each example is later identified in at least one dataset during the analysis. See Figure 3 for examples to each entry.

[2]See supplemental material or vitro-testing.com for the full list.

Table 2. Excerpts from full list of hazards for stereo vision (simplified, l.s. = light source)

hid	Loc. / GW / Param.	meaning	entry
0	L.s. / No / Number	No l.s.	Highly underexposed image; only black-level noise
26	L. s. / Part of / Position	Part of l.s. is visible	L.s. in image is cut apart by image border
142	L. s. / Less / Beam prop.	Focused beam	Scene with half lit object leaving a large portion severely underexposed
183	Medium / Less / Transparency	Medium is optically thicker than expected	Fog or haze in image reduces visibility depending on distance from observer
376	Object / Less / Complexity	Object is less complex than expected	Scene contains simple object without texture or self-shading (*e.g.* grey opaque sphere)
476	Object / No / Reflectance	Obj. has no reflectance	Well-lit scene contains a very dark object without texture nor shading
482	Object / As well as / Reflectance	Obj. has both shiny and dull surface	Object has a large glare spot on its surface that obscures same areas in the left/right image
701	Objects / Spatial aper. / Reflectance	Refl. creates a chaotic pattern	Large parts of the image show an irregular distorted mirror-like reflection
904	Obs. / Faster / Position	Observer moves too fast	Image has parts with clearly visible motion blur
1090	Obs. / No / PSF	No optical blurring	Image contains strong aliasing artifacts

Figure 3. Identified hazards in datasets corresponding to Table 2

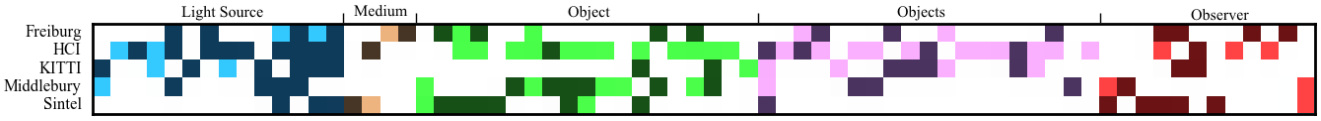

Figure 4. Distribution of hazards per dataset: Dark cells show identified hazards while light cells represent entries with no GT, too small area or disputed ones; color represents CV-HAZOP category.

4.2. Analyzing Test Data

Of all identified test datasets from Section 2 we concentrate on a specific subgroup: All datasets that are public, provide GT data, and have at least ten test images. This results in the following subsets: all Middlebury datasets, both KITTI datasets, Sintel, HCI Training 1K, and Freiburg[3]. The Oxford RobotCar and SYNTHIA datasets are certainly interesting for this evaluation but have been published too recently given their huge size for us to process.

The dataset analysis commences as described in Section 4.3. Two additional analysts as well as all authors participate, ensuring that each dataset is analyzed by at least two different people to reduce bias. In total, 76 hazards are found across all the datasets. They result in 48 unique hazards out of 117. Most hazards are found in the HCI Training Dataset, Freiburg, and Sintel (16 each) followed by the KITTI and Middlebury datasets (14 each). Figure 3 gives some examples of identified hazards. The entries correspond to the rows of Table 2. Some hazard entries are deemed to be unreliable for the upcoming evaluation due to missing GT, insufficient size, or disagreement between experts. These disputed entries were removed from the evaluation. Figure 4 visualizes the hazard distribution over all datasets. This still leaves 50 entries uncovered by any of the datasets. Section 4.5 will discuss these open issues.

[3]Freiburg is annotated without *flying things*. These scenes are too chaotic for analysts to evaluate in a reasonable time.

4.3. Dataset Evaluation

The following stereo vision algorithms are now evaluated on the analysed datasets: SAD + Texture Thresholding (TX) & Connected Component Filtering [28], SGM [20] with rank filtering (RSGM), Elas [16] + TX & Weighted Median Post Processing Filtering (WM), Cost-Volume Filtering (CVF) & WM[54], PatchMatch (PM) & WM [3], Cross-Scale Cost Aggregation using Census and Segment-Trees (ST) & WM [75, 39], SPSS [72], and MC-CNN [73] using their KITTI2012 pre-trained fast network. Average RMS and bad pixel scores for each test image in the datasets are calculated as evaluation metrics.

Figure 5 shows a summary of the difficulty for each dataset based on the performance of each algorithm. Unfilled bars visualize the relative amount of frames in the whole dataset with the specified difficulty, while filled bars denote the amount of hazard frames within this range of difficulty. All bars are normed to their respective maximum number.

As expected, the algorithms behave quite differently on the same test data due to their different implementations and performance varies depending on the dataset[4]. It is evident that hazard frames strongly group at the bins of higher difficulty. Filled bars are generally higher than unfilled bars for difficult frames (bins D/E) and lower for easier frames (bins A/B). This trend can be observed in each of the datasets for all algorithms.

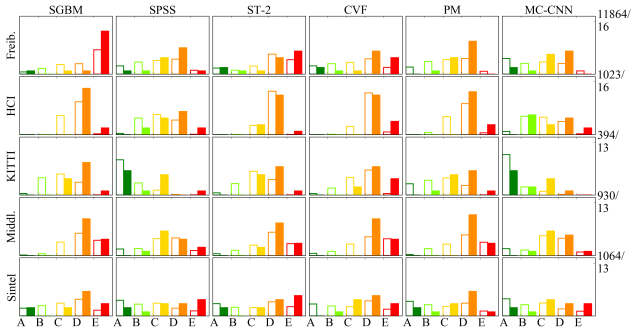

Figure 5. Difficulty distribution of frames in each dataset. Relative number of pixels having an error > 4 disparities sorted into 5 bins: A:[0-5%), B:[5-10%), C:[10-20%), D:[20-50%), E:[50-100%]. Right side: number of frames in full dataset (no-fill bars) / with hazards (solid bars). All bars (no-fill/solid) of a single plot add up to these respective numbers (first/second).

4.4. Data Size

One important aspect of test dataset design is using the right data size. Too much redundancy increases processing time and might drown relevant individual test cases in a flood of meaningless repetitions. Too few test cases, on the other hand, will prevent the detection of important shortcomings due to missing scenarios.

For our experiment we sort all frames by their difficulty according to performance per algorithms. We choose a subset of all frames and iteratively calculate the average performance over the subset adding easier frames with each step. In the first experiment we randomly pick frames from the dataset, achieving a good representation for the entire dataset. In our second experiment we only add the easiest frames of the dataset. In the third experiment we only use frames identified by the HAZOP analysis and add them in hardest-first manner. To make the results comparable we plot the accumulation of all frames up to the number of annotated hazard frames.

Figure 6 shows a comparison of the results (random, best first, HAZOP) for the Sintel dataset. Using only hazard frames allows the same level of distinction between algorithms with comparable numbers of images. Selecting hard frames is a valid way to evaluate algorithms. The advantage of using hazard frames in comparison to random picking is that they also give insights into why a specific test case failed.

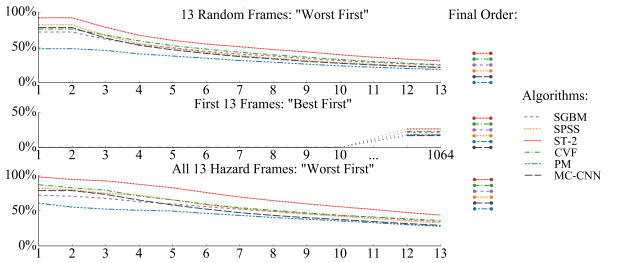

Figure 6. Comparison of cumulative average performance of 13 frames from Sintel: Random picking, easiest frames, hazard frames (all sorted by difficulty) using the bad pixel metric with a threshold of 4.

4.5. Missing Tests

There were numerous hazard entries which were not found in any of the test datasets examined by the analysts (Table 3). These entries were categorized into two groups: border cases and negative test cases. The distinction between the two is sometimes dependent on the domain (*e.g.* not every implementation has to work with a large field-of-view (FOV) or when there is rain/snow in the scene). For this checklist we tried to cover a very broad domain and require a lot of robustness from the algorithm, *i.e.* indoor scenes and outdoor street environments under difficult weather conditions. Using these guidelines we also decided on the clustering into *the Bad* and *the Ugly* groups. Positive test cases are usually easy to define. Therefore, we focus on difficult test cases.

4.6. Future Work

Testing algorithms with single test cases for each hazard allows for valuable insights, but more than a single data point is needed for representative statistics. Systematic test data, gradually increasing in difficulty, should be used to evaluate the breaking point of the algorithm (in regard to a specific hazard). Frame-based annotation should be augmented using labels within the images. This allows evaluations of hazards affecting smaller areas which otherwise get outweighed by the surrounding area's influences.

Focusing on the most difficult frames of a dataset can also give good indications about hazards without the need to inspect each frame. However, this can introduce a huge bias toward the evaluation metric used and propagate existing redundancy.

5. Conclusion

This paper focuses on analyzing datasets for their ability to test the robustness of CV applications. A thorough survey of 28 existing stereo vision test datasets demonstrates their

[4]See supplemental material for addition algorithm performance graphs.

Table 3. Selection of hazards missing from current test datasets, see supplemental material for the full list

hid	entry
Border cases (*the Bad*)	
6	L.s. and its reflection are visible on the same epipolar line
12	Multiple l.s. are periodically placed and aligned on the same epipolar line
63	L.s. visible in image with a long elongated thin shape (*e.g*. neon tube) creating an unusual overexposed area
107	L.s. projects structured pattern onto a surface that produces two distinctly different Moire patterns in both images
259	Scene is split into two equal parts: one without particles and another with considerable amount of particles
310	Two different sized objects are positioned on the same epipolar line but their projected views are identical
341	Scene contains an expanding/shrinking object resulting in noticeable radial motion blur
479	Object has strongly reflecting material that mirrors larger parts found on the same epipolar line
523	Two partially transparent objects are entangled in such a way that both allow the view on each other
694	Scene contains a clear reflection of observer together with potential matching parts on the same epipolar
754	Scene contains a prominent rainbow effect (*i.e*. mist/haze with a view-depended colour band)
758	Scene contains pronounced refraction rings (*e.g*. oil slick)
803	Cameras have both a wide FOV ($>$135deg)
918	Lens body/lens hood is prolonged and its corners are thus blocking the view
926	Two cameras both have considerable comparable amount of dirt/pollution but with different distributions
1091	Very different textures in left and right image due to large scale Moire effects
Negative test cases (*the Ugly*)	
245	Cloud of visible particles (*e.g*. pollen, small leaves) in the air are obscuring the whole scene
504	Highly transparent object encompassing a second opaque object that gets distorted due to the other object's shape
695	Scene contains a large concave mirror that shows an clean upside-down copy of parts of the scenery
719	Observer is placed between two parallel mirrors facing each other so that "infinite" number of reflections occur
790	Left and right image are the same while showing a diverse scene
916	One camera lens contains dust/dried mud that creates a partially defocused area in the image
921	Lens is broken cleanly leaving a visible crack in the image's center
933	Images contain rolling shutter artifacts
955	Images contain considerable chromatic aberration and many visible edges
983	Images have considerable amounts of vignetting and scene contains many objects close to the observer
1094	One of the two sensors is somewhat out of focus
1105	Inter-lens reflections create visible copy of objects in the image
1162	Image before rectification originates from considerably rectangular pixels (instead of square, near to e.g. 2:1 ratio)
1166	Images contain strong static image noise for well-lit scenes
1261	One camera delivers negative image (or color channels swapped)
1265	Images use logarithmic quantization instead of linear or wrong gamma mapping

progression over time. We present an improved methodology based on the CV-HAZOP checklist analysis method that identifies challenging elements in datasets. We apply this methodology to selected popular stereo datasets to identify challenging test cases. Then, we evaluate a broad range of algorithms on those selected datasets. The correlation between frames identified as challenging and test case difficulty allows these conclusions: (i) cases marked as challenging are evidently difficult independent of dataset or algorithm choice, and (ii) challenging cases of a dataset are a representative subset of the entire dataset. Testing with challenging cases only yields similar results compared to the entire dataset but contains all listed challenges.

Most importantly, we present a list of challenges that are missing from all the selected datasets. This results in a roadmap of 32 practical inputs for researchers designing new datasets.

In our opinion, new datasets should increase difficulty and variability but not necessarily size: In addition to the easy cases (*the Good*), more border cases (*the Bad*) and negative test cases (*the Ugly*) should be added. Ultimately, this will increase applicability, usefulness, and the safety of CV solutions as well as systems that rely on them.

6. Acknowledgement

This project has received funding from the Electronic Component Systems for European Leadership Joint Undertaking under grant agreement No 692480. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and Germany, Saxony, Spain, Austria, Belgium, Slovakia. See www.iosense.eu; Thanks for proofreading and good suggestions go to Daniel Steininger (AIT) and Emma Alexander (Harvard).

References

[1] D. Biedermann, M. Ochs, and R. Mester. Evaluating visual ADAS components on the COnGRATS dataset. In *2016 IEEE Intelligent Vehicles Symposium (IV)*, 2016. 4

[2] J.-L. Blanco, F.-A. Moreno, and J. González-Jiménez. The málaga urban dataset: High-rate stereo and lidars in a realistic urban scenario. *International Journal of Robotics Research*, 33(2):207–214, 2014. 3, 4

[3] M. Bleyer, C. Rhemann, and C. Rother. Patchmatch stereo-stereo matching with slanted support windows. In *British Machine Vision Conference*, 2011. 6

[4] M. Brown, D. Burschka, and G. Hager. Advances in computational stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(8):993–1008, 2003. 2

[5] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In A. Fitzgibbon et al. (Eds.), editor, *European Conf. on Computer Vision (ECCV)*, Part IV, LNCS 7577, pages 611–625. Springer-Verlag, Oct. 2012. 3, 4

[6] F. Campo, F. Ruiz, and A. Sappa. Multimodal stereo vision systems: 3d data extraction and algorithm evaluation. *IEEE Journal of Selected Topics in Signal Processing*, 6(5):437–446, 2012. 2

[7] J. Cem Kaner. What is a good test case? *STAR East*, 2003. 4

[8] M. Chen, K. Dhingra, W. Wu, L. Yang, R. Sukthankar, and J. Yang. PFID: Pittsburgh fast-food image dataset. In *Proceedings of International Conference on Image Processing*, 2009. 2, 3

[9] K. Cordes, B. Rosenhahn, and J. Ostermann. Increasing the accuracy of feature evaluation benchmarks using differential evolution. In *IEEE Symposium on Differential Evolution (SDE)*, 2011. 2, 3

[10] M. Cordts, M. Omran, S. Ramos, T. Scharwächter, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset. In *CVPR Workshop on The Future of Datasets in Vision*, 2015. 3, 4

[11] Daimler Böblingen, 6D-Vision. http://www.6d-vision.com. Accessed: 2016-11-15. 4

[12] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *CVPR*, 2009. 4

[13] M. Firman. RGBD Datasets: Past, Present and Future. In *CVPR Workshop on Large Scale 3D Data: Acquisition, Modelling and Analysis*, 2016. 2

[14] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig. Virtual worlds as proxy for multi-object tracking analysis. In *CVPR*, 2016. 3, 4

[15] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *CVPR*, 2012. 2, 3

[16] A. Geiger, M. Roser, and R. Urtasun. Efficient large-scale stereo matching. In *Asian conference on computer vision*, pages 25–38. Springer, 2010. 6

[17] R. Haeusler and D. Kondermann. Synthesizing real world stereo challenges. In *German Conference on Pattern Recognition*, pages 164–173. Springer, 2013. 3, 4

[18] V. Haltakov, C. Unger, and S. Ilic. Framework for Generation of Synthetic Ground Truth Data for Driver Assistance Applications. In J. Weickert, M. Hein, and B. Schiele, editors, *Pattern Recognition*, Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2013. 4

[19] R. Hamzah and H. Ibrahim. Literature survey on stereo vision disparity map algorithms. *Journal of Sensors*, 2016. 2

[20] H. Hirschmüller. Stereo processing by semiglobal matching and mutual information. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(2):328–341, 2008. 6

[21] H. Hirschmuller and D. Scharstein. Evaluation of cost functions for stereo matching. In *CVPR*, pages 1–8. IEEE, 2007. 3

[22] A. Janoch. The berkeley 3d object dataset. Master's thesis, EECS Department, University of California, Berkeley, May 2012. 2

[23] C. G. Keller, M. Enzweiler, and D. M. Gavrila. A new benchmark for stereo-based pedestrian detection. In *IEEE Intelligent Vehicles Symposium*, pages 691–696, 2011. 2

[24] W. Kim, A. Ansar, R. Steele, and R. Steinke. Performance analysis and validation of a stereo vision system. In *IEEE International Conference on Systems, Man and Cybernetics*, 2005. 2

[25] R. Klette, N. Kugrer, T. Vaudrey, K. Pauwels, M. van Hulle, S. Morales, F. I. Kandil, R. Haeusler, N. Pugeault, C. Rabe, and M. Lappe. Performance of correspondence algorithms in vision-based driver assistance using an online image sequence database. In *IEEE Intelligent Vehicles Symposium*, pages 2012–2026, 2011. 2

[26] D. Kondermann, R. Nair, K. Honauer, K. Krispin, J. Andrulis, A. Brock, B. Gussefeld, M. Rahimimoghaddam, S. Hofmann, C. Brenner, et al. The hci benchmark suite: Stereo and flow ground truth with uncertainties for urban autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 19–28, 2016. 3

[27] D. Kondermann, A. Sellent, B. Jähne, and J. Wingbermühle. Robust Vision Challenge, 2012. 3

[28] K. Konolige. Small vision systems: Hardware and implementation. In *Robotics Research*. Springer, 1998. 6

[29] M. Kristan, J. Matas, A. Leonardis, T. Vojir, R. Pflugfelder, G. Fernandez, G. Nebehay, F. Porikli, and L. Cehovin. A novel performance evaluation methodology for single-target trackers. *TPAMI*, Accepted, 2016. 4

[30] D. Kumari and K. Kaur. A survey on stereo matching techniques for 3D vision in image processing. *International Journal on Engineering and Manufacturing*, 4:40–49, 2016. 2

[31] L. Ladický, P. Sturgess, C. Russell, S. Sengupta, Y. Bastanlar, W. Clocksin, and P. H. Torr. Joint optimisation for object class segmentation and dense stereo reconstruction. *International Journal of Computer Vision*, 2012. 3, 4

[32] N. Lazaros, G. C. Sirakoulis, and A. Gasteratos. Review of stereo vision algorithms: From software to hardware. *International Journal of Optomechatronics*, 2:435–462, 2008. 2

[33] M. Lemmens. A survey on stereo matching techniques. In *ISPRS Congress, commission V*, pages 11–23, 1998. 2

[34] A. Li, M. Lin, Y. Wu, M.-H. Yang, and S. Yan. NUS-PRO: A new visual tracking challenge. *TPAMI*, 38(2):335–349, 2016. 4

[35] W. Maddern, G. Pascoe, C. Linegar, and P. Newman. 1 Year, 1000km: The Oxford RobotCar Dataset. *The International Journal of Robotics Research (IJRR)*, to appear. 3, 4

[36] L. Maier-Hein, A. Groch, A. Bartoli, S. Bodenstedt, G. Boissonnat, P.-L. Chang, N. Clancy, D. Elson, S. Haase, E. Heim, J. Hornegger, P. Jannin, H. Kenngott, T. Kilgus, B. Müller-Stich, D. Oladokun, S. Röhl, T. R. dos Santos, H.-P. Schlemmer, A. Seitel, S. Speidel, M. Wagner, , and D. Stoyanov. Comparative validation of single-shot optical techniques for laparoscopic 3d surface reconstruction. *Transactions on Medical Imaging*, 33(10):1913–1930, 2014. 2

[37] L. Maier-Hein, P. Mountney, A. Bartoli, H. Elhawary, D. Elson, A. Groch, A. Kolb, M. Rodrigues, J. Sorger, S. Speidel, and D. Stoyanov. Optical techniques for 3d surface reconstruction in computer-assisted laparoscopic surgery. *Medical Image Analysis*, 17(8):974–996, 2013. 2

[38] M. Martorell, A. Maki, S. Martull, Y. Ohkawa, and K. Fukui. Towards a simulation driven stereo vision system. In *ICPR*, pages 1038–1042, 2012. 3, 4

[39] X. Mei, X. Sun, W. Dong, H. Wang, and X. Zhang. Segment-tree based cost aggregation for stereo matching. In *Computer Vision and Pattern Recognition*, pages 313–320, 2013. 6

[40] M. Menze and A. Geiger. Object scene flow for autonomous vehicles. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 3

[41] N. Morales, G. Camellini, M. Felisa, P. Grisleri, and P. Zani. Performance analysis of stereo reconstruction algorithms. In *ITSC*, pages 1298–1303, 2013. 2

[42] S. Morales, T. Vaudrey, and R. Klette. Robustness evaluation of stereo algorithms on long stereo sequences. In *Intelligent Vehicles Symposium*, pages 347–352, 2009. 2

[43] Y. Nakamura, T. Matsuura, K. Satoh, and Y. Ohta. Occlusion detectable stereo-occlusion patterns in camera matrix. In *Computer Vision and Pattern Recognition, 1996. Proceedings CVPR'96, 1996 IEEE Computer Society Conference on*, pages 371–378. IEEE, 1996. 4

[44] P. K. Nathan Silberman, Derek Hoiem and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 2

[45] D. Neilson and Y.-H. Yang. Evaluation of constructable match cost measures for stereo correspondence using cluster ranking. In *Computer Vision and Pattern Recognition, 2008. Proceedings CVPR'08, 2008 IEEE Computer Society Conference on*. IEEE, 2008. 2, 3

[46] M. Nielsen, H. Andersen, D. Slaughter, and E. Granum. High-accuracy stereo depth maps using structured light. *Precision Agriculture*, 8(49):49–62, 2007. 2

[47] N.Mayer, E.Ilg, P.Häusser, P.Fischer, D.Cremers, A.Dosovitskiy, and T.Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. arXiv:1512.02134. 3

[48] U. of Alberta Stereo Vision Research, 2010. 2

[49] N. Onkarappa and D. Sappa, A. Synthetic sequences and ground-truth flow field generation for algorithm validation. *Multimedia Tools and Applications*, 2013. 4

[50] G. Pandey, J. R. McBride, and R. M. Eustice. Ford campus vision and lidar data set. *International Journal of Robotics Research*, 30(13):1543–1552, 2011. 2, 3

[51] D. Pfeiffer, S. K. Gehrig, and N. Schneider. Exploiting the power of stereo confidences. In *CVPR*, 2013. 3, 4

[52] Reinhard Klette: EISATS. http://ccv.wordpress.fos.auckland.ac.nz/eisats/. Accessed: 2016-11-15. 2

[53] R. Reulke, A. Luber, M. Haberjahn, and B. Piltz. Validierung von mobilen stereokamerasystemen in einem 3d-testfeld. (EPFL-CONF-155479), 2009. 3

[54] C. Rhemann, A. Hosni, M. Bleyer, C. Rother, and M. Gelautz. Fast cost-volume filtering for visual correspondence and beyond. In *Computer Vision and Pattern Recognition*, pages 3017–3024, 2011. 6

[55] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. Lopez. The SYNTHIA Dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*, 2016. 3, 4

[56] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nesic, X. Wang, and P. Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *Pattern Recognition*, pages 31–42. Springer, 2014. 3

[57] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 47(1/2/3):7–42, 2002. 2, 3

[58] D. Scharstein and R. Szeliski. High-accuracy stereo depth maps using structured light. In *CVPR*, pages 195–202. IEEE Computer Society, 2003. 3

[59] T. Scharwchter, M. Enzweiler, S. Roth, and U. Franke. Stixmantics: A medium-level model for real-time semantic scene understanding. In *ECCV*, 2014. 3, 4

[60] K. Schauwecker, S. Morales, S. Hermann, and R. Klette. A new benchmark for stereo-based pedestrian detection. In *IEEE Intelligent Vehicles Symposium*, 2011. 2

[61] G. S. Semwezi. Automation of negative testing. 2012. 4

[62] P. Sharma and N. Chitaliya. Obstacle avoidance using stereo vision: A survey. In *International Journal of Innovative Research in Computer and Communication Engineering*, pages 24–29, 2015. 2

[63] N. Silberman and R. Fergus. Indoor scene segmentation using a structured light sensor. In *Proceedings of the International Conference on Computer Vision - Workshop on 3D Representation and Recognition*, 2011. 2

[64] A. Smeulders, D. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah. Visual tracking: An experimental survey. *TPAMI*, 36(7):1442–1468, 2014. 4

[65] M. Smith, I. Baldwin, W. Churchill, R. Paul, and P. Newman. The new college vision and laser data set. *The International Journal of Robotics Research*, 28(5):595–599, May 2009. 2, 3

[66] P. Steingrube, S. Gehrig, and U. Franke. *Performance evaluation of stereo algorithms for automotive applications*. Computer Vision Systems, 2009. 2

[67] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of rgb-d slam systems. In *Proc. of the International Conference on Intelligent Robot Systems (IROS)*, Oct. 2012. 2

[68] The Middlebury Computer Vision Pages. http://vision.middlebury.edu/. Accessed: 2016-11-15. 2

[69] A. Torralba and A. Efros. Unbiased look at dataset bias. In *CVPR*, pages 1521–1528, 2011. 4

[70] T. Vaudrey, C. Rabe, R. Klette, and J. Milburn. Differences between stereo and motion behavior on synthetic and real-world stereo sequences. In *IVCNZ*, pages 1–6, 2008. 3

[71] A. Wedel, C. Rabe, T. Vaudrey, T. Brox, U. Franke, and D. Cremers. Efficient dense scene flow from sparse or dense stereo data. In *10th European Conference on Computer Vision (ECCV '08)*, pages 739–751, Berlin, Heidelberg, 2008. Springer-Verlag. 3

[72] K. Yamaguchi, D. McAllester, and R. Urtasun. Efficient joint segmentation, occlusion labeling, stereo and flow estimation. In *European Conference on Computer Vision*, pages 756–771. Springer, 2014. 6

[73] J. Zbontar and Y. LeCun. Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*, 17:1–32, 2016. 6

[74] O. Zendel, M. Murschitz, M. Humenberger, and W. Herzner. CV-HAZOP: Introducing test data validation for computer vision. In *ICCV*, 2015. 4, 5

[75] K. Zhang, Y. Fang, D. Min, L. Sun, S. Yang, S. Yan, and Q. Tian. Cross-scale cost aggregation for stereo matching. In *Computer Vision and Pattern Recognition*, 2014. 6